Highlights

- We investigate the sample size and covariates selection issues for using the PS method in road safety evaluation studies.
- Including the covariates that significantly affect the road accidents in the PS model is recommended, regardless of whether they affect the implementation of road safety measures.
- A proper sample size is the one that ensures relevant covariates achieve acceptable balance.
- Practical procedures for using the PS method to evaluate the effects of road safety measures are proposed.

Propensity score methods for road safety evaluation: Practical suggestions from a simulation study

Yingheng Zhang[a,b,c], Haojie Li[a,b,c] (corresponding author: h.li@seu.edu.cn), N.N. Sze[d], Gang Ren[a,b,c]

a. School of Transportation, Southeast University, China
b. Jiangsu Key Laboratory of Urban ITS, China
c. Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, China
d. Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong

Abstract: The propensity score (PS) based method has been increasingly used in road safety evaluation studies. However, several major considerations regarding its implementation arise when using the PS method. First, as is well known, the PS method is 'data hungry' in terms of the number of treated and control units, however, it is sometimes difficult and time-consuming to construct a large sample especially in road safety studies. It would be helpful to better understand how to choose a proper sample size, as well as the ratio of the number of treated units to the control ones. Second, the criteria used for covariates selection of the PS model were not fully consistent across the existing road safety evaluation studies. Due to the complicated mechanisms behind the implementation of road safety measures and policies, including all relevant covariates that affect both the selection into treatment (i.e., implementation of road safety measures) and the outcomes (i.e., road accidents) is impossible. In this paper, we conduct a simulation study to investigate such issues and provide some practical suggestions for road safety evaluations. The estimator considered in this study is the inverse probability weighting (IPW) estimator based on the PS. Our results suggest that the bias and variance of the estimated treatment effect will remain stable when the sample size reaches a certain level. A proper sample size is the one that ensures relevant covariates achieve acceptable balance. Regarding the issue of covariates selection, including the covariates that significantly affect the road accidents is recommended, regardless of whether they affect the implementation of road safety measures. This study also proposes practical procedures for using the IPW estimator to evaluate the effects of road safety treatments.

Keywords: Causal inference, Propensity score, Inverse probability weighting, Road safety evaluation

1 Introduction

Road safety has been a major issue in contemporary societies. A variety of different road safety treatments (or measures) have been implemented, including policies, legislation and enforcement, physical changes to the network, and other general-purpose measures which directly or indirectly affect traffic conditions, driver behaviors, and travel environment (Li et al., 2019). Hence, there is a growing need to evaluate the performance and effectiveness of such road safety treatments.

In road safety evaluation studies, the safety effect of a treatment is defined as an expected reduction in target accident following the implementation of the treatment (Elvik, 1997). In most cases, treatments are not randomly assigned to units (i.e., locations or road segments) as they are targeted at specific road safety concerns. Therefore, the observed relationship between the road safety treatment and road accidents may be subject to confounding. Confounding factors are the variables that influence both the treatment assignment and the outcomes. For instance, regression to the mean (RTM) is a well-known manifestation of confounding that arises in the presence of 'selection bias' (Graham et al., 2019).

Over several decades, numerous statistical approaches have been applied in road safety evaluation studies. The empirical Bayes (EB) method is one of the most popular approaches, which is viewed as a statistically defensible means of increasing the precision of estimation and accounting for the confounding effects that arise via RTM (Hauer, 1997, 2002; Persaud and Lyon, 2007). However, the EB method needs a control group that is similar to the treatment group in baseline characteristics (Hauer, 1992). When this assumption is violated, the performance of EB can be adversely affected (Lord and Kuo, 2012; Wood and Donnell, 2017). In the recent literature, the propensity score (PS) method developed by Rosenbaum and Rubin (1983) has been increasingly used for road safety evaluation as a response to the potential limitations of the EB method (Karwa et al., 2011; Sasidharan and Donnell, 2013; Wood et al., 2015a, b; Wood and Donnell, 2016; Li and Graham, 2016; Li et al., 2013, 2019, 2020, 2021; Lu et al., 2020; Li and Donnell, 2020). The PS, defined as the conditional probability of receiving a treatment given a set of observed covariates, is used to systematically address the issue of similarity between treated and control groups. As suggested by Rosenbaum and Rubin (1983), adjusting for the PS is enough to eliminate the bias due to all observed confounding factors. Several previous works compared the results obtained from the PS and EB methods, suggesting that the PS method is a viable alternative to the EB method for road safety evaluation studies (Wood et al. 2015a; Li et al. 2019).

Although the PS method has become another popular approach in the road safety research, several considerations regarding its implementation arise when using the PS method. The first issue concerned in this paper relates to the sample size and the ratio of the number of treated units to the control ones. As is well known, the PS method is 'data hungry' in terms of the sample size. However, constructing a large sample for road safety analysis is sometimes difficult and time-consuming due to data restriction. According to the existing road safety studies using the PS methods, the sample size ranged from tens to thousands, and the

treated-control ratio also varied. For example, Wood and Donnell (2016) evaluated the safety effects of continuous green T intersections based on a Florida data set containing 30 treated intersections and 38 control intersections, and a South Carolina data set containing 16 treated intersections and 21 control intersections. Another research on the safety effects of transit signal priority by Song and Noyce (2019) used a data set containing 13 treated street sections, and 10 control sections similar to the treated ones were matched using the PS matching method. Some other studies used much larger samples. In a previous work by Li et al. (2013), for instance, the safety effects of speed cameras were evaluated based on a data set containing a total number of 771 camera sites and 4787 potential control sites manually selected from eight English administrative districts. Similarly, a 6464-intersection sample consisting of 888 treated ones and 5576 control ones was used for the evaluation of roadway lighting in Sasidharan and Donnell (2013).

The second issue concerns the selection of covariates. In theory, only the covariates that influence both the treatment assignment and the outcomes should be included in the PS model. However, due to the complicated mechanisms behind the implementation of road safety measures and policies, including all relevant covariates is almost impossible. Currently, the criteria for covariates selection are not fully consistent across the existing road safety evaluation studies. For example, Wood et al. (2015a) and Wood and Donnell (2016) both noted the covariates included in the PS model should be selected based on the relevance to the treatment. Sasidharan and Donnell (2013) also focused on the treatment assignment related covariates, and the covariates that influence the outcome were included to form a rich PS model. In the studies by Li et al. (2013, 2016), the primary criterion for covariates selection was to include the covariates that influence the treatment assignment and the outcomes.

In this paper we conduct a simulation study to investigate the above issues in settings with different data conditions, and provide some suggestions for road safety evaluation studies. The remainder of the paper is organized as follows. Section 2 introduces the potential outcome framework for causal inference and the PS method in detail. Section 3 describes our simulation setup, followed by the simulation results with discussion in Section 4. A case study on the evaluation of UK's speed enforcement camera programme is presented in Section 5, and the conclusions are given Section 6.

2 Propensity score methods

2.1 Potential outcome framework

The potential outcome framework for causal inference was extended to observational studies by Rubin (1974). Therefore, it was usually referred to as Rubin Causal Model (RCM) in the recent literature. Consider a binary treatment indicator T_i for a random sample with N units from a large population, indexed by $i = 1, 2, \ldots, N$. The treatment indicator T_i equals one if unit i receives treatment and zero otherwise. Then let $Y_i(T_i)$ be the

potential outcome of unit i under treatment T_i. In the case of a binary treatment, there are two potential outcomes $Y_i(1)$ and $Y_i(0)$ for unit i.

The fundamental problem of causal inference is that it is impossible to observe the outcomes of any unit i under both treatment status (Holland, 1986). That is, only one of the potential outcomes is observed for each unit i, and the unobserved one is called 'counterfactual outcome'. Therefore, the unit treatment effects $\delta_i = Y_i(1) - Y_i(0)$ are unobservable, and we are usually interested in the average treatment effect (ATE), which can be either multiplicative or additive. The multiplicative ATE, simply defined as the expected outcome for the treated divided by the expected outcome for the untreated (Rosenbaum, 2010), can be written as:

$$\text{ATE}_{\text{multiplicative}} = \frac{\mathbb{E}[Y(1)]}{\mathbb{E}[Y(0)]}$$

Another type of treatment effect – the additive ATE – can be written as:

$$\text{ATE}_{\text{additive}} = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

In the road safety research, the *crash modification factor* (CMF) is a well-known example of multiplicative treatment effect. Therefore, we use the $\text{ATE}_{\text{multiplicative}}$ as default in the following sections. We can estimate the ATE without observing all potential outcomes under the following three key assumptions.

Assumption 1. Stable Unit Treatment Value Assumption (SUTVA)

SUTVA (Rubin, 1980, Rubin, 1986) requires that each unit has a unique potential outcome under each treatment allocation, ensuring that the observed outcome under a given treatment allocation is equivalent to the potential outcome under that same treatment allocation:

$$Y_i^{obs} = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

where Y_i^{obs} is the observed outcome. See Rubin (1990) for more discussions on some possible violations of this assumption.

Assumption 2. Unconfoundedness

The second key assumption is unconfoundedness (Rubin, 1990), which is also known as conditional independence assumption (CIA) (Lechner, 1999). This assumption states that the potential outcomes for each unit are conditionally independent of the treatment assignment given a set of observed covariates X, implying that the difference in the outcomes between treated and control units with the same X can be solely attributed to the treatment:

$$(Y_i(0), Y_i(1)) \perp T_i | X_i$$

In practice, when the dimension of vector *X* is large, it can be difficult to condition on all relevant covariates. Rosenbaum and Rubin (1983) introduced the PS, defined as the conditional probability of treatment given a set of observed covariates, to deal with this problem. They suggested that if potential outcomes are independent of treatment assignment conditional on covariates *X*, they are also independent of treatment assignment conditional on the PS:

$$(Y_i(0), Y_i(1)) \perp T_i | P(T_i | X_i)$$

In practice, the PS can be estimated by any discrete choice model (e.g., logit and probit models). In binary treatment cases, logit and probit models usually yield similar results (Smith, 1997). Hence, the logit model is used in this study:

$$P(T = 1|X) = \frac{\exp(\alpha + \beta' X)}{1 + \exp(\alpha + \beta' X)}$$

where α is the intercept term and β' is the vector of regression coefficients.

Assumption 3. Overlap

This assumption is also known as 'common support condition' (CSC), which requires that all units have positive probability of receiving the treatment. More specifically as follows:

$$0 < P(T_i = 1 | X_i = x) < 1, \forall x$$

Aforementioned three assumptions allow us to estimate the excepted outcomes $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$ from observational data using a variety of estimators. In this study, we use the inverse probability weighting (IPW) estimator based on the PS in our simulations, similar to the estimator proposed by Horvitz and Thompson (1952). In the next subsection, the IPW estimator is introduced.

2.2 Inverse probability weighting estimator

The IPW estimator exploits the PS as weight to create a balanced sample of treated and control observations. The idea behind the IPW estimator is that a pseudo population is created in which the distributions of confounding factors among the treatment and control groups are the same as the overall distribution of those in the original population (Stürmer et al., 2006). By unconfoundedness assumption, we have:

$$\mathbb{E}\left[\frac{YT}{p(x)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{YT}{p(x)}\middle|x\right]\right] = \mathbb{E}\left[\frac{\mathbb{E}[Y(1)|x]\mathbb{E}[T|x]}{p(x)}\right] = \mathbb{E}\left[\mathbb{E}[Y(1)|x]\right] = \mathbb{E}[Y(1)]$$

where $p(x) = P(T = 1|X = x)$. Similarly, we have:

$$\mathbb{E}[Y(0)] = \mathbb{E}\left[\frac{Y(1-T)}{1-p(x)}\right]$$

Then the ATE can be estimated as:

$$\text{ATE}_{IPW} = [N^{-1}\sum_i^N \frac{T_i Y_i}{p_i}] \Big/ [N^{-1}\sum_i^N \frac{(1-T_i)Y_i}{1-p_i}]$$

Compared to the matching methods, which rely on the selected (or matched) units, the PS methods based on weighting approaches are more sensitive to misspecification of the PS model. If the PS model is misspecified, the IPW estimator can be substantially biased (Zhao, 2004).

3 Simulation

3.1 Objectives

In road safety evaluation studies, the first step for using the IPW approach to estimate the effect of a particular road safety measure is to manually select a sample of units (i.e., locations or road segments) from a large population to construct a treatment group and a control group. In very few cases, we have the total population and complete relevant data or information for both groups. Sometimes, constructing a large sample for road safety analysis is difficult and time-consuming due to data restriction. The first decision has to be made concerns the sample size and a proper ratio of the number of treated units to the control ones.

The second step is to estimate the PS based on discrete choice model. With the estimated PS, safety effects can be simply calculated by the following equation: $\text{ATE}_{IPW} = [N^{-1}\sum_i^N \frac{T_i Y_i}{p_i}]/[N^{-1}\sum_i^N \frac{(1-T_i)Y_i}{1-p_i}]$.

The model specification issue is important for PS estimation. As discussed earlier in subsection 2.2, in binary treatment cases, logit and probit models usually yield similar results. Therefore, the model choice is not covered in this study. Another key issue in this step is the inclusion of covariates, and numerous discussions on this issue are available (e.g., Rubin and Thomas, 1996; Augurzky and Schmidt, 2001; Bryson et al., 2002; Brookhart et al., 2006). In theory, only the covariates that influence both the treatment assignment (i.e., implementation of road safety measures) and outcomes (i.e., road accidents) should be included in the PS

model. In most cases of road safety evaluation, however, we are uncertain about the proper model specification due to the limited knowledge on the mechanisms behind the implementation of road safety measures or the unavailability of some important data. Which covariates should be included may become a question in such cases. Currently, the criteria for covariates selection are not fully consistent across the existing road safety evaluation studies. In the following simulations, we will set up a series of relevant scenarios to investigate the sample size and covariates selection issues, and provide some practical suggestions for road safety applications.

3.2 Sample size

First, the sample size issue is investigated via a simulation of sample selection process for the treatment and control groups from a finite population of 5000 units.

Scenario 1. The *data generating process* (*DGP*$_1$) for a population of 5000 is:

$$X_{1,pre} \sim Normal(1,1)$$
$$X_{1,post} \sim Uniform(0,1) + X_{1,pre}$$
$$X_2 \sim Normal(1,1)$$

The pre-treatment *safety performance function* (*SPF*$_{pre}$), which describes the statistical relationship between road accidents and relevant covariates (Hauer, 1995), can be written as:

$$Y_{pre} \sim Poisson(\exp(\alpha_0 + \alpha_1 X_{1,pre} + \alpha_2 X_2) * \epsilon)$$
$$\epsilon \sim Gamma(2, 0.5)$$

A binary road safety measure *T* is implemented as a function of the relevant covariates and the pre-treatment road accident number Y_{pre}:

$$T \sim Bernoulli(\text{expit}(\beta_0 + \beta_1 X_{1,pre} + \beta_2 X_2 + \beta_3 Y_{pre}))$$

The post-treatment *SPF*$_{post}$ can be described as:

$$Y_{post} \sim Poisson(\exp(\alpha_0 + \alpha_1 X_{1,post} + \alpha_2 X_2 + \delta T) * \epsilon)$$

where $\alpha_0 = 1$, $\alpha_1 = 0.1$, $\alpha_2 = 0.1$, $\beta_0 = -2$, $\beta_1 = 0.1$, $\beta_2 = 0.1$, $\beta_3 = 0.1$, and $\delta = \ln(0.8)$. The true value of *T*'s treatment effect is $\tau = \exp(\delta) = 0.8$, which can be considered as a road safety measure with

$CMF = 0.8$.

Scenario 2. The DGP_2 of *scenario 2* is the same as that of *scenario 1* except the road safety measure T is implemented based on a different function:

$$T \sim Bernoulli(\text{expit}(\beta_0' + \beta_1' X_{1,pre} + \beta_2' X_2 + \beta_3' Y_{pre}))$$

where $\beta_0' = -3.2,\ \beta_1' = 1,\ \beta_2' = 0.1,\ \beta_3' = 0.1$.

We note that DGP_1 and DGP_2 will both produce around 1500 treated units and 3500 control units from the total population of 5000 units. The main difference between the two scenarios is that $X_{1,pre}$ has greater impact on the implementation of road safety measures in *scenario 2*, which can result in considerable difference in $X_{1,pre}$ between the treatment and control groups. Figure 1 displays the PS and $X_{1,pre}$ distributions of *scenario 1* and *2*. To numerically quantify the differences in covariates across treatment groups, the Cohen's *standardized mean difference* (SMD), also termed as 'normalized difference' or 'standardized difference', is also reported, which is defined as:

$$SMD = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{(s_{x,t}^2 + s_{x,c}^2)/2}}$$

where $\bar{x}_t = \frac{\sum_{i:T=1} x_i}{N_t}$, $\bar{x}_c = \frac{\sum_{i:T=0} x_i}{N_c}$, $s_{x,t}^2 = \frac{\sum_{i:T=1}(x_i - \bar{x}_t)^2}{N_t - 1}$, and $s_{x,c}^2 = \frac{\sum_{i:T=0}(x_i - \bar{x}_c)^2}{N_c - 1}$.

It can be seen that DGP_2 produces a more unbalanced population (i.e., in *scenario 2*, treatment group has higher mean value of $X_{1,pre}$ than control group, and the SMD for $X_{1,pre}$ is larger than that of *scenario 1*, see Figure 1). In the previous work by Augurzky and Schmidt (2001), they used 'strong' to describe such treatment assignment mechanisms. They also suggested that when the selection into treatment is remarkably strong, it will be difficult to achieve an acceptable balance using PS matching method. In our simulations for investigating the sample size issue, the number of selected treatment units ranges from 50 to 500 in increments of 50, and two values of treatment-control sample ratios (treatment: control = 1:1, treatment: control = 1:3) are tested for both 'strong' and 'weak' treatment assignment scenarios.

3.3 Covariates

On the basis of *scenario 1* and *2* in subsection 3.1, we further include several additional covariates to investigate the issue of covariates selection in *scenario 3*.

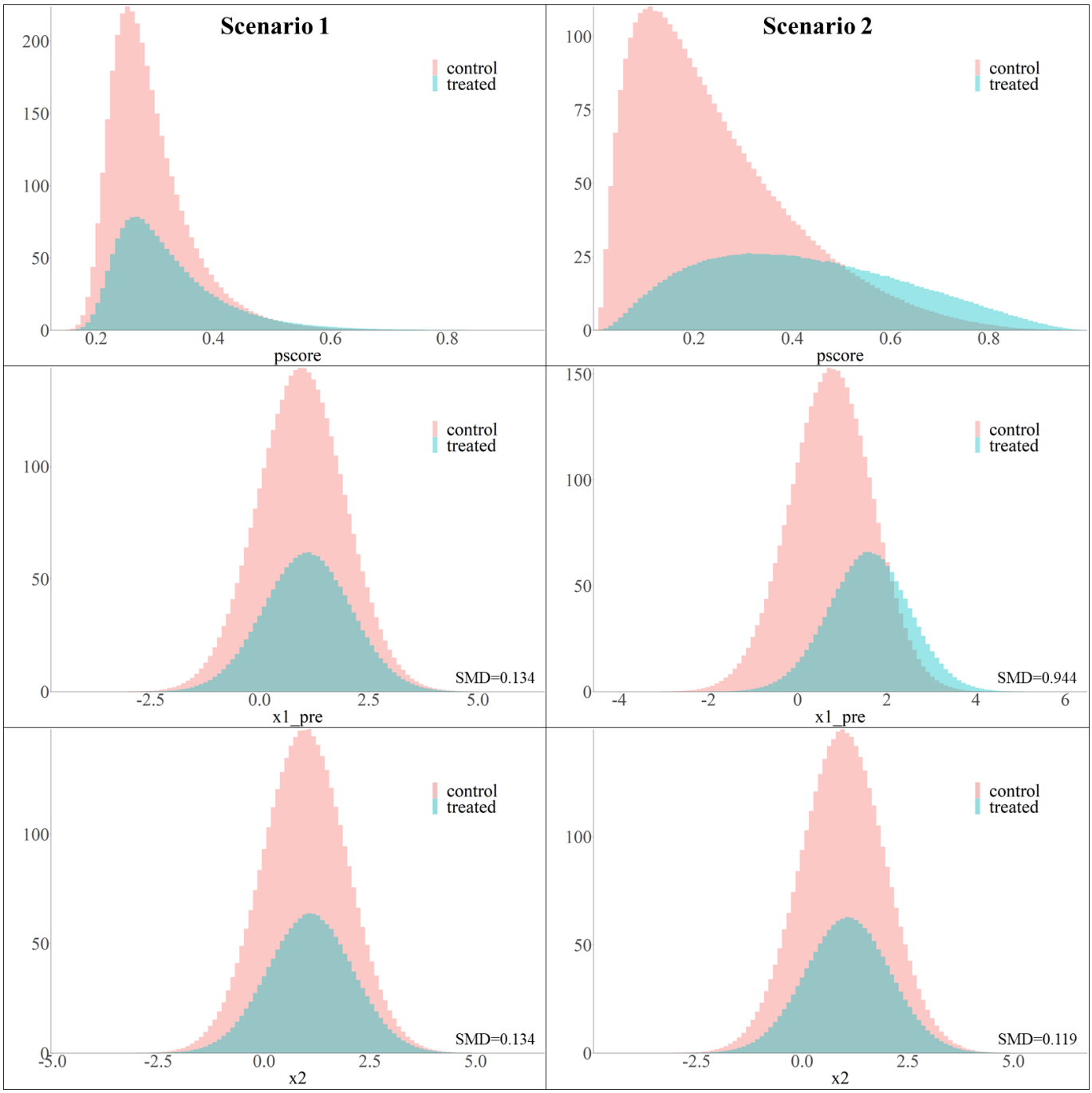


Fig. 1. PS and $X_{1,pre}$ distributions of *scenarios 1* and *2*.

Scenario 3. The DGP_3 for a population of 5000 is:

$$X_{1,pre} \sim Normal(0,1)$$
$$X_{1,post} \sim Uniform(0,1) + X_{1,pre}$$
$$X_2, X_3, X_4, X_5, X_6 \sim Normal(0,1)$$

The pre-treatment SPF_{pre} can be written as:

$$Y_{pre} \sim Poisson(\exp(\alpha_0 + \alpha_1 X_{1,pre} + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6) * \epsilon)$$

$$\epsilon \sim Gamma(2, 0.5)$$

A binary road safety measure *T* is implemented as the following function:

$$T \sim Bernoulli(\text{expit}(\beta_0 + \beta_1 X_{1,pre} + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6))$$

The post-treatment SPF_{post} can be described as:

$$Y_{post} \sim Poisson(\exp(\alpha_0 + \alpha_1 X_{1,post} + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6 + \delta T) * \epsilon)$$

where $\alpha_0 = 1$, $\alpha_1 = 0.1$, $\alpha_2 = 0.1$, $\alpha_3 = 0.1$, $\alpha_4 = 0.1$, $\alpha_5 = 0.01$, $\alpha_6 = 0.01$, $\beta_0 = -2$, $\beta_1 = 2$, $\beta_2 = 2$, $\beta_3 = 0.1$, $\beta_4 = 0.1$, $\beta_5 = 2$, $\beta_6 = 2$ (i.e., X_1 and X_2 have great impacts on both the implementation of road safety measures and the road accidents, X_3 and X_4 have little impacts on the implementation of road safety measures but great impacts on the road accidents, and X_5 and X_6 have great impacts on the implementation of road safety measures but little impacts on the road accidents). Also, $\delta = \ln(0.8)$ (i.e., the true value of *T*'s treatment effect is $\tau = \exp(\delta) = 0.8$). The assignment mechanism in *scenario 3* is a strong one in terms of $X_{1,pre}$, X_2, X_5, and X_6, as they have great impacts on the implementation of road safety measures. Four models are tested and compared in this scenario: (1) model including only X_1 and X_2, (2) model including X_1, X_2, X_3 and X_4, (3) model including X_1, X_2, X_5 and X_6, and (4) model including all the covariates. Models 1, 2, and 3 are partial models that produce inconsistent estimates of the PS, while model 4 is a full model. In this scenario, various sample sizes are also tested (100, 200, 300, 400, and 500 treated units), but the treated-control ratio is fixed at 1:3.

Scenario 4. Additional four *DGPs* ($DGP_{4.1}$, $DGP_{4.2}$, $DGP_{4.3}$, and $DGP_{4.4}$) are designed for supplementary analyses in *scenario 4*. Larger coefficient values of X_3, X_4, X_5 and X_6 are tested. That is, the relationships are not so weak as those in *scenario 3*:

(1) *Scenarios 4.1* and *4.2*: $DGP_{4.1}$ and $DGP_{4.2}$ are the same as DGP_3 except the road safety measure *T* is implemented based on a different function where β_3, $\beta_4 = 0.2$ ($DGP_{4.1}$), and β_3, $\beta_4 = 0.5$ ($DGP_{4.2}$).
(2) *Scenarios 4.3* and *4.4*: $DGP_{4.3}$ and $DGP_{4.4}$ are the same as DGP_3 except the *SPFs* are different where α_5, $\alpha_6 = 0.02$ ($DGP_{4.3}$), and α_5, $\alpha_6 = 0.05$ ($DGP_{4.4}$).

In this scenario, only the samples with 500 treated units and 1:3 treated-control ratio are tested.

All the models are simulated for 1000 iterations in each scenario. Mean values, relative bias in percentage (RB(%)), variance (Var), and the mean squared error (MSE) of the estimated treatment effects are reported.

4 Results and discussions

4.1 Results for scenarios 1 and 2 – sample size issue

In this section, the simulation results for each scenario are reported. Table 1 presents the simulation results of *scenarios 1* and *2*, and Figure 2 displays the relationships between the relative bias, MSE and sample size.

In *scenario 1*, moderate imbalance exists in the original population. When the selected sample size is small (i.e., less than 100 or 150 treated units), the estimated treatment effect is significantly biased. Relative bias of the estimated effect is less than 1% and remains stable when the number of treated units increases above 150 and 100 for the samples with 1:1 and 1:3 treated-control ratio respectively ('1:1 sample' and '1:3 sample'). Moreover, 1:3 samples always have better performance than 1:1 samples in terms of relative bias, that is, once a treated group is constructed, a larger control pool is preferred. Also, smaller MSE is observed for larger samples. More specifically, MSE is decreased by approximately 3-fold when the treated-control ratio is increased from 1:1 to 1:3. It is also worth noting that when the number of treated units is increased above 300, 1:1 and 1:3 samples performed similarly in terms of relative bias and MSE.

The level of imbalance in the original population is increased in *scenario 2* (see Figure 1) due to a stronger treatment assignment mechanism, which makes it more difficult to satisfy the overlap assumption based on the selected samples. Compared to *scenario 1*, increased relative bias and MSE of the estimated treatment effect are observed in *scenario 2*. To achieve the goal of 'RB(%) <1', 300 and 200 treated units are required for 1:1 and 1:3 samples respectively, which are larger than the corresponding figures in *scenario 1*. Moreover, greater differences in the performance between 1:1 and 1:3 samples are observed, indicating that a sample with higher level of imbalance needs a larger control group than a relatively balanced sample to give precise estimation results of the treatment effect.

4.2 Results for scenarios 3 and 4 – covariates selection issue

Table 2 reports the results obtained from *scenario 3*. Generally, partial models 1 and 2 result in similar treatment effect estimates with small degree of MSE (MSE = 0.0026 and 0.0027 respectively when the treated sample size is 500). However, the relative bias of the partial model 1 is larger than 1%. For partial model 3 and the full model, which further include X_5 and X_6, larger MSE is observed (MSE = 0.0144 and 0.0211 respectively when the treated sample size is 500). Moreover, the relative bias of the treatment effect estimated by model 3 is much more unstable, ranging from 0.18% to 2.60%. The partial model 2 can be considered as a trade-off between estimation bias and variance for varied ranges of sample size, which is more practically preferred in such scenarios.

To further investigate the reason behind such varied performances of the four models, balancing tests were conducted. Figure 3 shows the SMDs of each covariate in the unbalanced samples and weighted samples. SMD has been introduced in section 3.1, and for the weighted samples, SMD is calculated by the weighted equivalents of $\bar{x}$ and s^2. It can be seen that X_1, X_2, X_5 and X_6 have larger SMD values (approximately 0.675) in the original samples due to their larger impacts on the implementation of safety measure, while the original SMDs of X_3 and X_4 is smaller (approximately 0.032). After weighting the original samples by the PSs estimated by partial models 1 and 2, SMDs of X_1 and X_2 are significantly reduced. However, the SMDs of X_5 and X_6 are larger than 0.800, remaining at a high level, as they are not included in models 1 and 2. Smaller SMDs of X_5 and X_6 are observed (SMD < 0.100) in the weighted samples of partial model 3 and the full model, which include X_5 and X_6 as covariates. On the other hand, when the sample size is small (i.e., 100 or 200 treated units), SMDs of X_1 and X_2 are larger in the weighted samples created by model 3 and the full model than those in the weighted samples created by models 1 and 2, which is due to the fact that the additional balance of X_5 and X_6 partly sacrifices the balance of X_1 and X_2. Moreover, boxplot of the estimated PS from the four models is shown in Figure 4. Obviously, models 3 and 4, which include X_5 and X_6 result in larger differences in the estimated PSs between the treated and control groups, and therefore, they produce treatment effect estimates with larger variance and MSE.

Table 3 reports the results obtained from *scenario 4*, which is an alternative one to *scenario 3*. That is, X_3 and X_4 have increasing impacts on the implementation of road safety measure (*Scenarios 4.1* and *4.2*), and X_5 and X_6 have increasing impacts on the road accidents (*Scenarios 4.3* and *4.4*). In *scenarios 4.1* and *4.2*, when the PS models are specified by ignoring X_3 and X_4 (models 1 and 3), non-negligible bias (relative bias > 3%) in estimated effect is observed, as X_3 and X_4 play more important roles in the safety measure assignment. Also, ignoring X_5 and X_6 (models 1 and 2) can produce treatment effect estimates with lower MSE in *scenarios 4.3* and *4.4*, but the bias is non-negligible (relative bias > 2%), as X_5 and X_6 have greater impacts on the road accidents. To summarize, model 2 produces better effect estimates in *scenarios 4.1* and *4.2*, while the full model produces better effect estimates in *scenarios 4.3* and 4.4, suggesting that the importance of each covariate to the road accidents is a major consideration for the PS model specification. Based on the results obtained from *scenarios 3* and *4*, it is recommended that the covariates significantly affecting the road accidents should be included in the PS model, regardless of whether they affect the implementation of road safety measure.

Table 1

Results for *scenarios 1* and *2* (true treatment effect $\tau = 0.800$).

Sample		*Scenario 1*				*Scenario 2*			
		Mean	RB(%)	Var	MSE	Mean	RB(%)	Var	MSE
1:1	50	0.8313	3.9154	0.0157	0.0167	0.8756	9.4449	0.0459	0.0515
	100	0.8179	2.2380	0.0074	0.0077	0.8380	4.7498	0.0231	0.0245
	150	0.8077	0.9577	0.0050	0.0050	0.8282	3.5291	0.0150	0.0158
	200	0.8063	0.7924	0.0034	0.0034	0.8173	2.1565	0.0111	0.0114
	250	0.8059	0.7418	0.0029	0.0029	0.8087	1.0893	0.0081	0.0081
	300	0.8042	0.5239	0.0024	0.0024	0.8072	0.9060	0.0073	0.0073
	350	0.8025	0.3122	0.0019	0.0019	0.8075	0.9396	0.0053	0.0054
	400	0.8012	0.1550	0.0019	0.0019	0.8073	0.9084	0.0043	0.0043
	450	0.8016	0.1971	0.0016	0.0016	0.8080	1.0021	0.0041	0.0041
	500	0.8031	0.3828	0.0014	0.0015	0.8062	0.7749	0.0037	0.0037
1:3	50	0.8117	1.4574	0.0053	0.0054	0.8259	3.2413	0.0140	0.0147
	100	0.8059	0.7328	0.0024	0.0024	0.8115	1.4393	0.0066	0.0067
	150	0.8005	0.0606	0.0015	0.0015	0.8087	1.0911	0.0041	0.0042
	200	0.7998	0.0194	0.0012	0.0012	0.8041	0.5065	0.0030	0.0030
	250	0.8012	0.1544	0.0010	0.0010	0.8049	0.6147	0.0023	0.0024
	300	0.8009	0.1099	0.0008	0.0008	0.8016	0.2046	0.0018	0.0018
	350	0.8015	0.1900	0.0007	0.0007	0.8015	0.1849	0.0014	0.0014
	400	0.8012	0.1456	0.0006	0.0006	0.8010	0.1244	0.0014	0.0014
	450	0.8004	0.0521	0.0006	0.0006	0.7994	0.0702	0.0013	0.0013
	500	0.8012	0.1490	0.0006	0.0006	0.7987	0.1563	0.0011	0.0011

Notes: (1) 1:1 and 1:3 is the ratio of treated and control units, and (2) the 'Sample' column shows the number of the selected treated units.

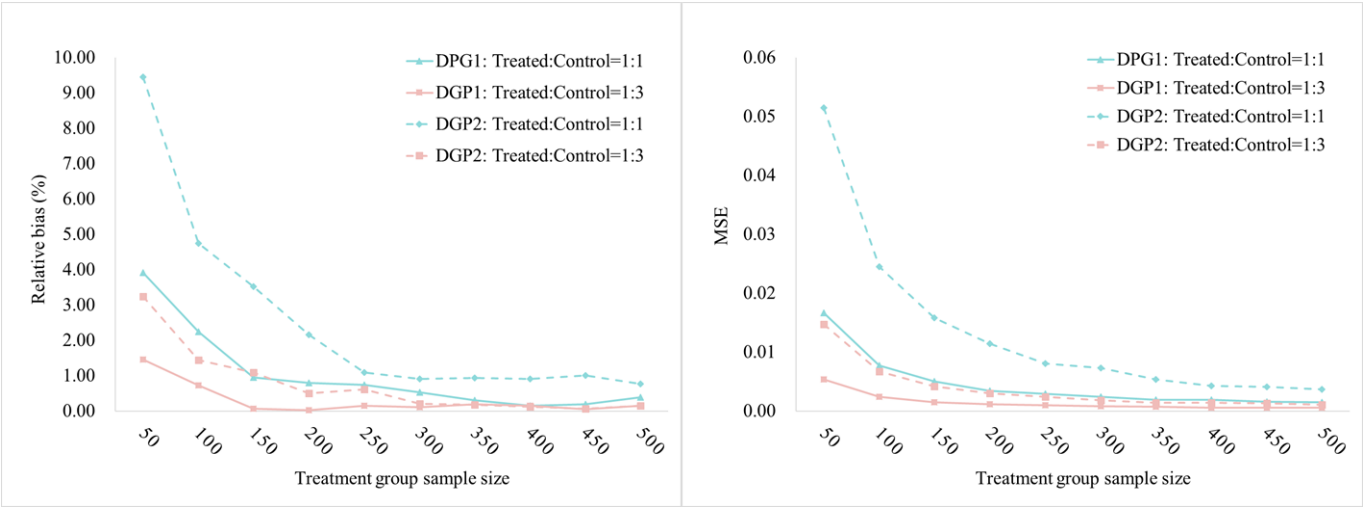


Fig. 2. Relative bias (%) and MSE of treatment effect estimates in *scenarios 1* and *2*.

Table 2

Results for *scenario 3* (true treatment effect $\tau = 0.800$).

Sample	Model 1 (X_1 and X_2)				Model 2 (X_1, X_2, X_3 and X_4)			
	Mean	RB (%)	Var	MSE	Mean	RB (%)	Var	MSE
100	0.8136	1.7030	0.0146	0.0147	0.8058	0.7246	0.0157	0.0158
200	0.8098	1.2246	0.0072	0.0073	0.8044	0.5561	0.0064	0.0064
300	0.8118	1.4730	0.0044	0.0045	0.8086	1.0697	0.0044	0.0045
400	0.8106	1.3223	0.0034	0.0035	0.8052	0.6500	0.0036	0.0036
500	0.8085	1.0570	0.0025	0.0026	0.8030	0.3770	0.0027	0.0027
Sample	Model 3 (X_1, X_2, X_5 and X_6)				Model 4 (full model)			
	Mean	RB (%)	Var	MSE	Mean	RB (%)	Var	MSE
100	0.8208	2.5969	0.1008	0.1011	0.8009	0.1175	0.0945	0.0944
200	0.8014	0.1774	0.0316	0.0316	0.8084	1.0440	0.0376	0.0376
300	0.8143	1.7900	0.0319	0.0320	0.8028	0.3463	0.0260	0.0260
400	0.8104	1.2959	0.0298	0.0299	0.8010	0.1244	0.0246	0.0246
500	0.8047	0.5875	0.0144	0.0144	0.8005	0.0660	0.0221	0.0221

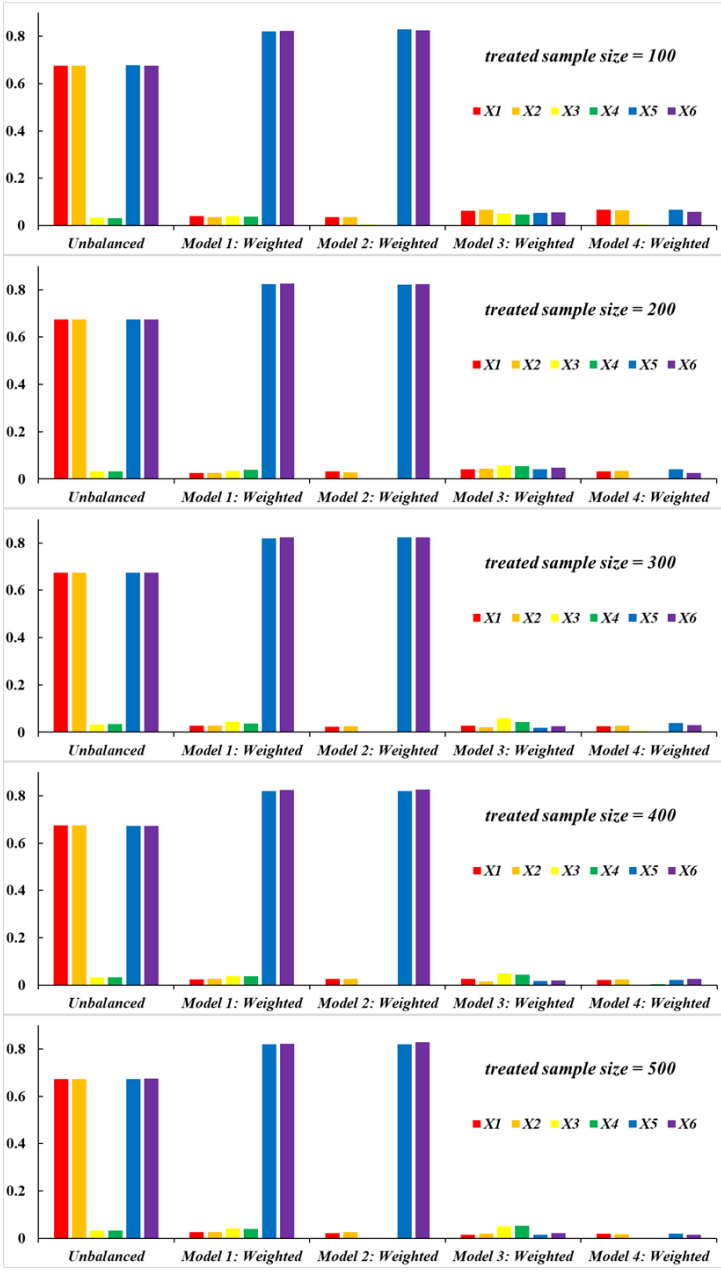
Note: 'Sample' column shows the number of the selected treated units.

Table 3

Results for *scenario 4* (true treatment effect $\tau = 0.800$).

DGPs	Model 1 (X_1 and X_2)				Model 2 (X_1, X_2, X_3 and X_4)			
	Mean	RB (%)	Var	MSE	Mean	RB (%)	Var	MSE
4.1	0.8297	3.7119	0.0030	0.0039	0.8027	0.3360	0.0028	0.0028
4.2	0.8652	8.1552	0.0032	0.0075	0.8012	0.1522	0.0031	0.0031
4.3	0.8240	2.9985	0.0027	0.0033	0.8186	2.3284	0.0026	0.0029
4.4	0.8643	8.0360	0.0031	0.0073	0.8571	7.1412	0.0030	0.0063
DGPs	Model 3 (X_1, X_2, X_5 and X_6)				Model 4 (full model)			
	Mean	RB (%)	Var	MSE	Mean	RB (%)	Var	MSE
4.1	0.8287	3.5906	0.0145	0.0154	0.8060	0.7487	0.0275	0.0275
4.2	0.8682	8.5194	0.0131	0.0177	0.8116	1.4481	0.0310	0.0311
4.3	0.8066	0.8242	0.0211	0.0211	0.7954	0.5700	0.0126	0.0126
4.4	0.8060	0.7485	0.0172	0.0172	0.8069	0.8685	0.0168	0.0168

Notes: (1) Treated sample size = 500, (2) treated: control ratio = 1:3.



	Unbalanced	Weighted 1	Weighted 2	Weighted 3	Weighted 4
X1	0.6752	0.0397	0.0348	0.0628	0.0663
X2	0.6761	0.0364	0.0360	0.0663	0.0637
X3	0.0324	0.0409	0.0049	0.0502	0.0040
X4	0.0320	0.0371	0.0014	0.0474	0.0002
X5	0.6770	0.8202	0.8293	0.0531	0.0672
X6	0.6751	0.8215	0.8245	0.0556	0.0574

	Unbalanced	Weighted 1	Weighted 2	Weighted 3	Weighted 4
X1	0.6743	0.0264	0.0328	0.0420	0.0317
X2	0.6749	0.0249	0.0277	0.0440	0.0339
X3	0.0315	0.0337	0.0019	0.0563	0.0021
X4	0.0332	0.0395	0.0014	0.0537	0.0011
X5	0.6750	0.8236	0.8209	0.0408	0.0406
X6	0.6733	0.8254	0.8231	0.0475	0.0257

	Unbalanced	Weighted 1	Weighted 2	Weighted 3	Weighted 4
X1	0.6742	0.0281	0.0236	0.0269	0.0256
X2	0.6745	0.0271	0.0259	0.0202	0.0273
X3	0.0325	0.0428	0.0018	0.0598	0.0047
X4	0.0337	0.0370	0.0028	0.0440	0.0041
X5	0.6736	0.8194	0.8245	0.0190	0.0381
X6	0.6733	0.8235	0.8237	0.0246	0.0301

	Unbalanced	Weighted 1	Weighted 2	Weighted 3	Weighted 4
X1	0.6750	0.0245	0.0263	0.0260	0.0223
X2	0.6739	0.0259	0.0268	0.0145	0.0235
X3	0.0326	0.0375	0.0006	0.0492	0.0006
X4	0.0328	0.0373	0.0013	0.0440	0.0029
X5	0.6725	0.8205	0.8208	0.0173	0.0217
X6	0.6735	0.8232	0.8255	0.0193	0.0267

	Unbalanced	Weighted 1	Weighted 2	Weighted 3	Weighted 4
X1	0.6732	0.0267	0.0216	0.0157	0.0184
X2	0.6734	0.0261	0.0254	0.0202	0.0162
X3	0.0320	0.0422	0.0017	0.0509	0.0030
X4	0.0333	0.0404	0.0000	0.0536	0.0015
X5	0.6723	0.8191	0.8208	0.0140	0.0194
X6	0.6744	0.8219	0.8280	0.0226	0.0146

Fig. 3. *Scenario 3*: tests of covariates balance (the absolute values of SMD are reported).

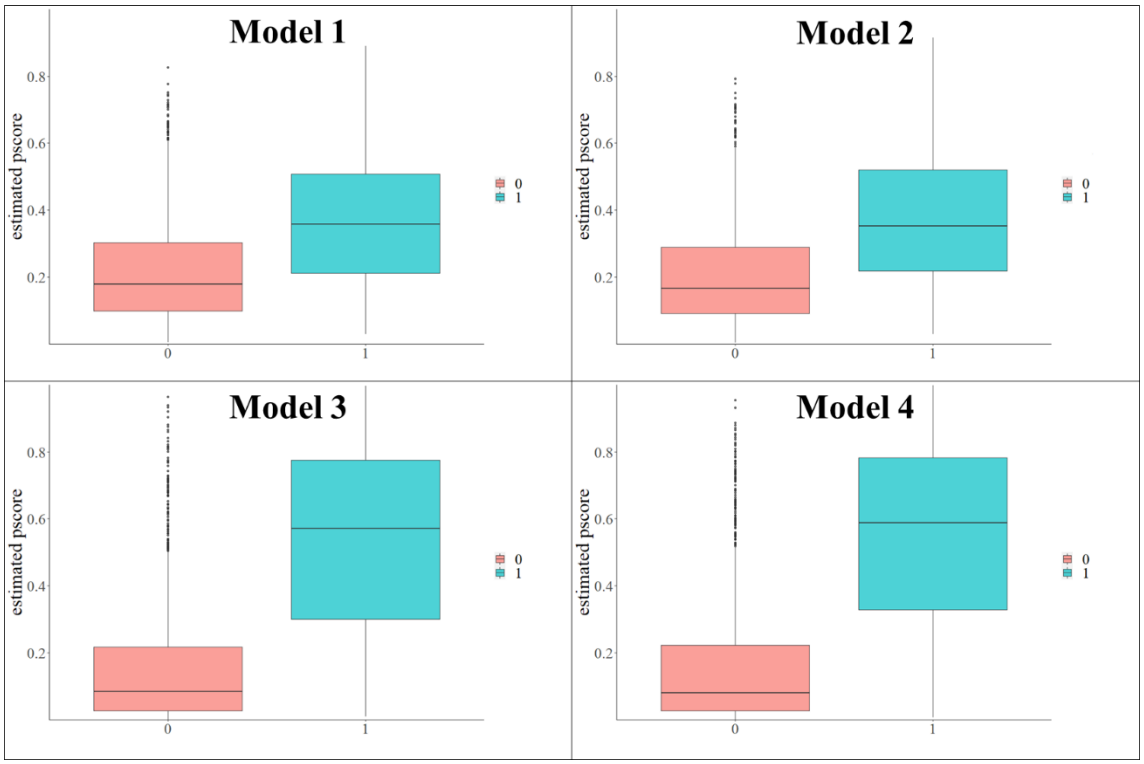


Fig. 4. Estimated PS box-plot (treated vs. control, treated sample size = 500).

4.3 The strategy

In practical applications, the procedures for using the IPW estimator to evaluate the effects of road safety measures can be illustrated as the following steps.

(1) Selection of treated units. The road segments, intersections, or traffic analysis zones (TAZs) allocated with the safety treatments of interest are selected and aggregated to construct a treated group.

(2) Pre-selection of control units. In most cases of road safety evaluations, control units are selected manually, which costs much time. A practical way is to select an equal sized group of control units preliminarily, that is, the primary treated-control ratio is 1:1.

(3) Data collection. The data for road safety related covariates (e.g., accident record, traffic volume) is collected for both treated and pre-selected control groups. According to the simulation results, covariates that greatly impact both the implementation of road safety measures and the road accidents, and covariates that greatly impact the road accidents should be included in the PS model. On the contrary, covariates that have limited impacts on the road accidents can be omitted. Previous studies, and handbooks for the road safety programme of interest can be served as guidelines for the selection of covariates. Also, a regression model is helpful to explore the importance of each covariate in the *SPFs*.

(4) Preliminary balancing test for the original sample. The SMD of each covariate is calculated to assess the differences in average covariate values by treatment status in the original sample. Large SMD values usually indicate that the corresponding covariates have great impacts on the implementation of road safety

measures, and also imply that the treatment assignment is strong.

(5) Simulation results suggest that if the treatment assignment is not strong in terms of the SMD metric, the original 1:1 sample may be sufficient to produce precise treatment effect estimates. However, extra control units are required to create a 1:*n* sample if the treatment assignment is strong.

(6) Weighting with estimated PSs. Once a sample is created, weighting can be conducted with the PSs estimated by any discreate choice model. The IPW approach weights the sample by the inverse of the conditional probability of the observed treatment status of each unit.

(7) Balancing test for the weighted sample. The SMD of each covariate for the weighted sample is calculated to check the balance between treatment groups. If SMD values for all the concerned covariates are smaller than the threshold set by researchers, the treatment effect can be subsequently estimated. Otherwise, extra control units are required or some treated units with extreme PS values should be discarded to satisfy the overlap assumption, and repeat steps (6) and (7).

(8) Estimating the treatment effect. Using the IPW estimator, treatment effect is computed as $\mathrm{ATE}_{IPW} = [N^{-1}\sum_i^N \frac{T_i Y_i}{p_i}]/[N^{-1}\sum_i^N \frac{(1-T_i)Y_i}{1-p_i}]$.

5 An application: UK's speed enforcement camera

In this section, the IPW approach discussed in the previous sections is applied to a road safety evaluation case. The safety effects of UK's speed enforcement cameras are estimated. We use the dataset collected by Li et al. (2013), which contains 771 camera sites and 4787 control sites randomly selected from the following eight English administrative districts: Cheshire, Dorset, Greater Manchester, Lancashire, Leicester, Merseyside, Sussex, and West Midlands. A geographical information systems (GIS) software, MapInfo, is used to create road segments manually on a map. The speed cameras included in the analysis were installed during the period of 2002 to 2004, and the research period covers three years before and after the camera installation for every camera site (1999 to 2007). For the purpose of illustrating the implementation procedures described in section 4.3, assume that the 771 treated sites have been selected to construct the treated group while the control group have not been constructed, and therefore, the first step as mentioned in the strategy is to create an equal sized control group (i.e., pre-selected control group) by randomly selecting control road sites on the same map.

The inclusion of covariates would be less complicated if clear criteria for treatment assignment were available. Therefore, we first review the handbooks for the speed camera programme. In the UK, fixed speed camera sites were selected primarily based on the following three guidelines (Department for Transport (DfT), 2004; DfT, 2005; Gains et al., 2004):

(1) Site length: between 400 and 1500 m.

(2) Number of fatal and serious collisions (FSCs): at least 4 FSCs per km in the last three calendar years. The figure is changed to 3 FSCs per km in the 2005 version.

(3) Number of personal injury collisions (PICs): at least 8 PICs per km in the last three calendar years.

Site length is an important exposure variable in road safety analysis, and accident history record is also an important predictor of the road accidents in the post-treatment period. Hence, these three covariates should be included in the PS model. In addition, there are several secondary criteria for camera sites selection:

(1) 85th percentile speed at least 10% above the 'speed limit + 2 mph'.
(2) At least 20% of drivers exceeding the speed limit.
(3) Suitable site conditions.
(4) Community concerns.

However, the data for 85th percentile speed and speeding percentage are normally unavailable for the researchers. Also, the site condition and community concern issues are related to a variety of different factors (e.g., road level, road infrastructures, index of multiple deprivation (IMD), population density, employment, school zones) (Li et al., 2021). As noted in the previous section, only the covariates significantly affecting the road accidents should be included in the PS model, regardless of whether they affect the implementation of road safety measures. Therefore, according to the empirical evidence and previous research, the following data was collected as potential covariates included in the PS model: accident record, traffic volume (annual average daily traffic, AADT), site length, road type, speed limit, and the number of minor intersections. Descriptive statistics for road accidents and the relevant data are summarized in Table 4.

Subsequently, a standard negative binomial (NB) regression model based on the form proposed by Mountain et al. (1997, 2005) was used to develop *SPFs* and further determine the covariates selected into the PS model. A proper threshold of *z*-statistics can be set by researchers for discarding the covariates that have limited impacts on the road accidents. In this analysis, road type related covariates (*A road*, *B road*, and *M road*) are not included in the PS model. Table 5 shows the results of the preliminary balancing test on selected covariates. The SMDs vary across the selected covariates, with four out of nine SMDs larger than 0.30 in absolute value. These four are historical PICs and FSCs, the number of minor intersections, and the speed limit of 30 mph. The first two are listed in the handbook of speed camera programme as the primary criteria for camera installation. Therefore, the treatment assignment mechanism of UK's speed camera case can be considered as a strong one. We further increased the number of control units from 800 to 2000 to initially ensure a 1:2.5 treated-control ratio. Then the steps (6) and (7) mentioned in section 4.3 were repeated.

Table 6 reports the results of balancing tests for the weighted sample. The number of control units was increased from 2000 to 4500 by an increment of 500. It can be seen that a control group containing 3000 sites is able to achieve the goal of 'SMDs < 0.05'. That is, limited improvement can be achieved by selecting more sites. Based on the dataset with 771 treated sites and 3000 control sites, the estimated treatment effects on the treated units are -0.947 and -0.118 for PICs and FSCs respectively. In other words, the average reduction caused by the speed enforcement camera programme in annual PICs and FSCs per km is around 0.947 and 0.118 in absolute number. The results obtained from the dataset with 4500 control sites are -0.973 and -0.123,

similar to the results obtained from the dataset with 3000 control sites. Furthermore, we tested two alternative models. *Model 1* additionally includes the road type covariates (*A road*, *B road*, and *M road*), and *model 2* only includes the accident history record and the covariates used in the *SPF* proposed by Mountain et al. (1997) (*AADT*, *L*, and *I*), which are considered as the key covariates. The results obtained from *model 1* are -0.944 and -0.114 for PICs and FSCs respectively, which are similar to those from the original model, and the results from *model 2* are -0.834 and -0.109. Such results indicate that the covariates that have limited impacts on the road accidents (*A road*, *B road*, and *M road*) can be omitted, however covariates *SL30* and *SL40* (speed limit), which significantly affect the road accidents in this case, should not be excluded.

Table 4

Variable descriptive statistics.

Variables	Description	Mean	S.D.	Max.	Min.
Continuous variables					
PIC	Number of PICs from 1999 to 2001	8.943	11.879	123	0
FSC	Number of FSCs from 1999 to 2001	1.151	1.664	17	0
AADT	Annual average daily traffic	18162	10424	101221	399
L	Site length (m)	0.702	0.475	1.5	0.2
I	Number of minor intersections	3.792	4.646	65	0

Variables	Description	Proportion in the sample
Categorical variables		
A road	1 = A level road	78.859%
B road	1 = B level road	16.805%
M road	1 = Minor road	1.835%
SL30	1 = speed limit is 30 mph	57.588%
SL40	1 = speed limit is 40 mph	17.434%

Table 5

Results of preliminary balancing test.

Variable	Treated mean	Treated S.D.	Control mean	Control S.D.	|SMD|×100
PIC	12.722	13.759	8.256	11.147	35.669
FSC	1.843	2.263	0.986	1.392	45.610
AADT	19039	9540	18559	10701	4.733
L	0.712	0.481	0.702	0.477	1.968
I	5.458	6.399	3.445	4.066	37.545
SL30	0.757	0.429	0.524	0.500	50.187
SL40	0.122	0.327	0.180	0.384	16.266

Table 6

Weighted sample balancing tests for varied numbers of control units (|SMD|×100 is reported).

Variable	Control sample size					
	2000	2500	3000	3500	4000	4500
PIC	6.4642	5.8058	4.9628	4.4978	4.1974	3.8540
FSC	5.2520	4.6512	3.9017	3.4627	3.2248	2.9349
AADT	3.7023	3.3084	2.9237	2.6501	2.4696	2.2914
L	0.7389	0.5434	0.3712	0.2773	0.1887	0.0972
I	1.5453	1.0537	0.8007	0.6531	0.5301	0.5598
SL30	0.4264	0.3134	0.2246	0.1630	0.1213	0.0831
SL40	0.8808	0.7446	0.6616	0.5920	0.5252	0.4841

6 Conclusions

Propensity score based methods have been increasingly applied to evaluate the performance of road safety treatments. However, there are some questions regarding its implementation. This study investigates two common issues, sample size and covariates selection, in settings with different data conditions, and provides some practical suggestions. Also, a case study on the UK's speed enforcement camera programme is conducted for a better illustration.

Two major findings are reported in this section. First, the simulation results suggest that the bias and variance of the estimated treatment effect will remain stable when the sample size increases above a certain number. Therefore, once the key covariates have been balanced between the treated and control groups, limited improvement can be achieved by selecting more units. Second, in road safety studies, it is suggested that the covariates that greatly impact the road accidents should be included in the PS model, while the inclusion of covariates that have limited impacts on the road accidents may lead to increased variance of treatment effect estimates. Therefore, the optimal practice for constructing a PS model is to include the covariates that significantly affect the road accidents, regardless of whether they affect the implementation of road safety measures. The results from a previous study by Brookhart et al. (2006) point in the same direction.

Furthermore, based on these two findings, we provide a practical strategy for road safety applications with clearer covariates selection criteria and time-saving procedures. An evaluation on the UK's speed camera programme, which has been conducted previously in Li et al. (2013), was redone in this paper following the proposed strategy. Several modifications were made to simplify the procedures. For example, three road type related covariates (i.e., A level, B level, and minor road) were excluded due to their limited impacts on the road accidents. Moreover, the original dataset containing 4787 control sites manually selected on the map, while the results suggest that 3000 is large enough for creating a balanced sample weighted by the estimated

PSs, which save our time on selecting more road sites and collecting additional information.

There are also some limitations in this study. First, the PS is usually estimated by parametric models, but the functional form is not fully addressed in this study, which is another crucial issue for PS model specification in addition to the selection of covariates. In the existing applications, non-linear and non-additive relationships between covariates and the treatment assignment are often neglected, which could adversely affect the performance of the PS method. Second, the heterogeneity of treatment effect is not considered in the simulations for simplicity, despite the fact that the effects of road safety treatments may vary across locations in the real world. Finally, it should be noted that some relevant works published recently incorporated a variety of emerging machine learning approaches into the PS methods and the causal inference framework (e.g., Goller et al., 2020; Otok et al., 2020; Wager and Athey, 2018; Athey and Imbens, 2019), pointing out a future direction for the aforementioned issues.

Acknowledgement

This work was supported by the National Key R&D Program of China (No.2018YFE0102700), and the Key Project of National Natural Science Foundation of China (Grant No. 51638004).

Reference

Athey, S., Imbens, G.W., 2019. Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, 11, 685-725. https://doi.org/10.1146/annurev-economics-080217-053433

Augurzky, B., Schmidt, C.M., 2001. The Propensity Score: A Means to An End. IZA Discussion Paper No. 271.

Brookhart, M.A., Schneeweiss, D., Rothman, K.J., Glynn, R.J., Avorn, J., Stürmer, T., 2006. Variable Selection for Propensity Score Models. *American Journal of Epidemiology*, 163(12), 1149-1156. https://doi.org/10.1093/aje/kwj149

Bryson, A., Dorsett, R., Purdon, S., 2002. The use of propensity score matching in the evaluation of active labour market policies. Working Paper No. 4, A study carried out on behalf of the Department for Work and Pensions. https://www.researchgate.net/publication/30524857

Department for Transport, 2004. Handbook of Rules and Guidance for the National Safety Camera Programme for England and Wales for 2005/06.

Department for Transport, 2005. Handbook of Rules and Guidance for the National Safety Camera Programme for England and Wales for 2006/07.

Elvik, R., 1997. A Framework for Cost-Benefit Analysis of the Dutch Road Safety Plan. Institute of Transport Economics, Oslo, Norway.

Gains, A., Heydecker, B., Shrewsbury, J., Robertson, S., 2004. The National Safety Camera Programme 3-Year Evaluation Report. Research Report, Department for Transport, London, U.K.

Goller, D., Lechner, M., Moczall, A., Wolff, J., 2020. Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany's programmes for long term unemployed. *Labour Economics*, 65, 101855. https://doi.org/10.1016/j.labeco.2020.101855

Graham, D.J., Naik, C., McCoy, E.J., Li, H., 2019. Do speed cameras reduce road traffic collisions? *PLoS ONE*, 14(9), e0221267. https://doi.org/10.1371/journal.pone.0221267
Hauer, E., 1992. Empirical Bayes approach to estimation of "unsafety": the multi-variate regression method. *Accident Analysis & Prevention*, 24(5), 457-477. https://doi.org/10.1016/0001-4575(92)90056-O
Hauer, E., 1995. On exposure and accident rate. *Traffic Engineering and Control*, 36(3), 134-138.
Hauer, E., 1997. Observational Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety. Pergamon, Oxford, U.K.
Hauer, E., Harwood, D.W., Council, F.M., Griffith, M.S., 2002. Estimating Safety by the Empirical Bayes Method: A Tutorial. *Transportation Research Record*, 1784, 126-131. https://doi.org/10.3141/1784-16
Holland, P.W., 1986. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396), 945-960. https://doi.org/10.1080/01621459.1986.10478354
Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685. https://doi.org/10.1080/01621459.1952.10483446
Imbens, G.W., 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1), 4-29. https://doi.org/10.1162/003465304323023651
Karwa, V., Slavković, A.B., Donnell, E.T., 2011. Causal inference in transportation safety studies: Comparison of potential outcomes and causal diagrams. *Annals of Applied Statistics*, 5, 1428-1455. https://projecteuclid.org/euclid.aoas/1310562728
Lechner, M., 1999. Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification. *Journal of Business & Economic Statistics*, 17, 74-90. https://doi.org/10.2307/1392240
Li, H., Graham, D.J., Majumdar, A., 2013. The impacts of speed cameras on road accidents: an application of propensity score matching methods. *Accident Analysis & Prevention*, 60, 148-157. https://doi.org/10.1016/j.aap.2013.08.003
Li, H., Graham, D.J., 2016. Quantifying the causal effects of 20 mph zones on road casualties in London via doubly robust estimation. *Accident Analysis & Prevention*, 93, 65-74. https://doi.org/10.1016/j.aap.2016.04.007
Li, H., Graham, D.J., Ding, H., Ren, G., 2019. Comparison of empirical Bayes and propensity score methods for road safety evaluation: A simulation study. *Accident Analysis & Prevention*, 129, 148-155. https://doi.org/10.1016/j.aap.2019.05.015
Li, H., Zhang, Y., Ren, G., 2020. A causal analysis of time-varying speed camera safety effects based on the propensity score method. *Journal of Safety Research*, 75, 119-127. https://doi.org/10.1016/j.jsr.2020.08.007
Li, H., Zhu, M., Graham, D.J., Ren, G., 2021. Evaluating the speed camera sites selection criteria in the UK. *Journal of Safety Research*. https://doi.org/10.1016/j.jsr.2020.11.013
Li, L., Donnell, E.T., 2020. Incorporating Bayesian methods into the propensity score matching framework: A no-treatment effect safety analysis. *Accident Analysis & Prevention*, 145, 105691. https://doi.org/10.1016/j.aap.2020.105691
Lord, D., Kuo, P., 2012. Examining the effects of site selection criteria for evaluating the effectiveness of traffic safety countermeasures. *Accident Analysis & Prevention*, 47, 52-63. https://doi.org/10.1016/j.aap.2011.12.008
Lu, D., Guo, F., Li, F., 2020. Evaluating the causal effects of cellphone distraction on crash risk using propensity score methods. *Accident Analysis & Prevention*, 143, 105579. https://doi.org/10.1016/j.aap.2020.105579
Mountain, L.J., Maher, M.J., Fawaz, B., 1997. The effects of trend over time on accident model predictions. In: Proceedings of the PTRC 25th European Transport Forum, 145-158.

Mountain, L.J., Hirst, W.M., Maher, M.J., 2005. Are speed enforcement cameras more effective than other speed management measures? The impact of speed management schemes on 30 mph roads. *Accident Analysis & Prevention*, 37, 742-754. https://doi.org/10.1016/j.aap.2005.03.017

Otok, B.W., Musa, M., Purhadi, Yasmirullah, S.D.P., 2020. Propensity score stratification using bootstrap aggregating classification trees analysis. *Heliyon*, 6, e04288. https://doi.org/10.1016/j.heliyon.2020.e04288

Persaud, B., Lyon, C., 2007. Empirical Bayes before-after safety studies: Lessons learned from two decades of experience and future directions. *Accident Analysis & Prevention*, 39, 546-555. https://doi.org/10.1016/j.aap.2006.09.009

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P.R., 2010. Design of Observational Studies. Springer, New York. https://doi.org/10.1007/978-1-4419-1213-8

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688-701. https://doi.org/10.1037/h0037350

Rubin, D.B., 1980. Comment of 'Randomization Analysis of Experimental Data: The fisher Randomization Test'. *Journal of the American Statistical Association*, 75, 591-593. https://doi.org/10.2307/2287653

Rubin, D.B., 1986. Comment: Which ifs have causal answers? *Journal of the American Statistical Association*, 81, 961-962. https://doi.org/10.1080/01621459.1986.10478355

Rubin, D.B., 1990. Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25, 279-292. https://doi.org/10.1016/0378-3758(90)90077-8

Rubin, D.B., Thomas, N., 1996. Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics*, 52, 249-264. https://doi.org/10.2307/2533160

Sasidharan, L., Donnell, E.T., 2013. Application of propensity scores and potential outcomes to estimate effectiveness of traffic safety countermeasures: Exploratory analysis using intersection lighting data. *Accident Analysis & Prevention*, 50, 539-553. https://doi.org/10.1016/j.aap.2012.05.036

Smith, H.L., 1997. Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology*, 27(1), 325-353. https://doi.org/10.1111/1467-9531.271030

Song, Y., Noyce, D., 2019. Effects of transit signal priority on traffic safety: Interrupted time series analysis of Portland, Oregon, implementations. *Accident Analysis & Prevention*, 123, 291-302. https://doi.org/10.1016/j.aap.2018.12.001

Stürmer, T., Rothman, K.J., Glynn, R.J., 2006. Insights into different results from different causal contrasts in the presence of effect- measure modification. *Pharmacoepidemiology and Drug Safety*, 15(10), 698-709. https://doi.org/10.1002/pds.1231

Tarko, A., Eranky, S., Sinha, K., 1998. Methodological considerations in the development and use of crash reduction factors. In 77th Annual Meeting of the Transportation Research Board, Washington, DC.

Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228-1242. https://doi.org/10.1080/01621459.2017.1319839

Wood, J.S., Donnell, E.T., 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 & Prevention*, 75, 144-154. https://doi.org/10.1016/j.aap.2014.11.019

Wood, J.S., Donnell, E.T., 2016. Safety evaluation of continuous green T intersections: A propensity scores-genetic matching-potential outcomes approach. *Accident Analysis & Prevention*, 93, 1-13. http://dx.doi.org/10.1016/j.aap.2016.04.015

Wood, J.S., Donnell, E.T., 2017. Causal inference framework for generalizable safety effect estimates.

Accident Analysis & Prevention, 104, 74-87. https://doi.org/10.1016/j.aap.2017.05.001

Wood, J.S., Gooch, J.P., 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 & Prevention*, 82, 180-191. https://doi.org/10.1016/j.aap.2015.06.002

Zhao, Z., 2004. Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence. *The Review of Economics and Statistics*, 86(1), 91-107. https://doi.org/10.1162/003465304323023705