ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(4); 374-387

Covariate Balance as a Quality Indicator for Propensity Score Analysis

Yusuf KARA* Akihito KAMATA** Elisa GALLEGOS***
Chalie PATARAPICHAYATHAM**** Cornelis J. POTGIETER******

Abstract
Propensity score analysis, such as propensity score matching and propensity score weighting, is becoming increasingly popular in educational research. When a propensity score analysis is conducted, examining the covariate balance is considered to be crucial to justify the quality of the analysis results. However, it has been pointed out that solely considering how covariates balance after matching may not be enough for justifying the quality of the propensity score analysis results. Suitable covariate balance may still yield biased estimates of treatment effects. The current study aimed to systematically demonstrate this problem by a series of simulation studies. As a result, it was revealed that a good covariate balance on the mean and/or the variance does not guarantee reduced bias on an estimated treatment effect. It was also found that estimation of the treatment effect can be unbiased to some degree, even with a lack of balance under specific conditions.

Key Words: Propensity score analysis, covariate balance, unbiased treatment effect.

INTRODUCTION

Propensity score (PS) analysis is becoming increasingly popular in educational research that adopts quasi-experimental design. PS analysis allows researchers to create a balance between treatment and control groups in order to estimate unbiased treatment effect when a randomized control design is not possible or is considered unethical/impractical (Guo & Fraser, 2015; Rosenbaum & Rubin, 1983). A challenge with the application of PS analysis is how to ensure that we have obtained an improved treatment effect estimate, ideally an unbiased estimate of the population treatment effect.

Typically, a researcher will go through a series of steps, and back steps, when conducting PS analysis. First, researchers identify the covariates to be included in the PS model and select a method (e.g., logistic regression-LR) for obtaining the propensity scores (PSs). Second, researchers conduct the analysis to estimate PSs and use them to balance the treatment and control groups in terms of covariate distributions. Third, researchers examine the quality of covariate balance and either go back to the first step and/or account for any insufficiently balanced covariates in the outcome analysis. Fourth, researchers conduct the outcome analysis for treatment effect estimation.

An important factor affecting PS analysis results is the type of covariates (i.e., associated only with the outcome, treatment assignment, or both) used to estimate PSs. Researchers have investigated the effects of including different covariate types in PS models in an effort to identify the most appropriate covariates

* Senior Data Analyst, Center on Research and Evaluation, Southern Methodist University, Dallas, TX, USA, ykara@smu.edu, ORCID ID: 0000-0003-0691-0630
** Professor, Department of Education Policy & Leadership and Department of Psychology, Southern Methodist University, Dallas, TX, USA, akamata@smu.edu, ORCID ID: 0000-0001-9570-1464
*** Senior Program Specialist, Center on Research and Evaluation, Southern Methodist University, Dallas, TX, USA, elisa@smu.edu
**** Research Assistant Professor, Department of Education Policy & Leadership, Southern Methodist University, Dallas, TX, USA, cpatarapichy@smu.edu
***** Assistant Professor, Department of Mathematics, Texas Christian University, Fort Worth, TX, USA, c.potgieter@tcu.edu, ORCID ID: 0000-0002-1995-6817

To cite this article:
Kara, Y., Kamata, A., Gallelogos, E., Patarapichayatham, C., & Potgieter, C.J. (2021). Covariate balance as a quality indicator for propensity score analysis . *Journal of Measurement and Evaluation in Education and Psychology, 12*(4), 374-387. doi: 10.21031/epod.993571

Received: 10.09.2021
Accepted: 19.12.2021

to include. Covariates that yield balanced treatment and control groups and ultimately unbiased treatment effects would be the desired ones to use in PS analysis. There have been some studies that revealed the importance of including covariates strongly associated with the outcome and not with the treatment assignment (Brookhart et al., 2006; Myers et al., 2011). Some studies also highlighted the negative effect of intrinsic covariates (mostly associated with the treatment assignment and showed little association with the outcome) on PS models and recommended avoiding them due to inconsistencies in the PS analyses (Bhattacharya & Vogt, 2007) or increased bias in the estimation of treatment effects (Patrick et al., 2011). Similarly, What Works Clearinghouse-WWC (2017) also highlighted the negative effects of using intrinsic covariates on the estimation of the treatment effect. Researchers also explored that including covariates those strongly associated with both treatment assignment and outcome, yields the least bias (Brookhart et al., 2006; Hong, Aaby, Siddique, & Stuart, 2018; Myers et al., 2011; Steiner, Cook, Shadish, & Clark, 2010).

Another important factor is the method utilized for obtaining PSs. The standard method is traditional LR in most applications. There have been more advanced methods that involve a combination of adjustments within the calculations/algorithms for obtaining PSs. For example, researchers have found that non-parametric and adaptive approaches (e.g., generalized boosted regression-GBR, classification and regression trees-CART, nearest neighbor matching-NNM, etc.) are more promising than LR in terms of covariate balance and treatment effect estimation (Cannas & Arpino, 2019; Lee, Lessler, & Stuart, 2010; McCaffrey, Ridgeway, & Morral, 2004; Setoguchi, Brookhart, Glynn, & Cook, 2008; Westreich, Lessler, & Funk, 2010). Other researchers found that certain PSs, such as the covariate balancing propensity scores (CBPS), perform very well in terms of balancing treatment and control groups (Kainz et al., 2017). In the current study, we consider multiple methods for estimating the PSs, which are elaborated in the methods section.

The most important overall goal for PS analysis is the reduction of confounding by balancing the treatment and control groups. In other words, PS analysis aims to reduce the confounding effect of external variables on the estimation of the true treatment effect. For this reason, examining covariate balance is crucial in justifying PS analysis results (Kainz et al., 2017). Researchers commonly evaluate the first moment of the covariate distributions between treatment and control groups by using the standardized absolute mean difference (SAMD). As also implied by Stuart, Lee, and Leacy (2013), SAMD is the most common measure of balance that is calculated similarly to effect size. It simply checks the magnitude of the mean differences in absolute scale compared to standardized mean difference (SMD). According to Rubin (2001), SAMD values from 0.1 to 0.25 are considered to be acceptable as indicators of good mean balance. Nevertheless, it was seen that applied researchers generally follow a stricter criterion as 0.05, which was suggested by WWC (2017) accessible through the Institute of Education Sciences as part of the US Department of Education.

In addition to checking the balance in terms of means, researchers may aim to evaluate other characteristics of covariate distributions in treatment and control groups. Along with SAMD, the literature has suggested the use of a combination of several criteria, including goodness-of-fit measures for covariate distributions (e.g., Kolmogorov-Smirnov test: Austin, 2009; Kainz et al., 2017; Stuart, 2010) and variance ratio (Kainz et al., 2017). Variance ratio (VR) is simply the ratio of a covariate's variance in the treatment and control group with a value of one indicating identical variances. According to Rubin (2007) VR values lower than 0.5 or higher than two are considered to be indicators of variance imbalance. We consider SAMD and VR as two widely-used standard measures of covariate balance in the current study. Readers are referred to Austin (2009) for a detailed overview of common balance measures and to Stuart, Lee, and Leacy (2013) for alternative balance measures such as prognostic score-based solutions.

Purpose of the Study

Obtaining good balance, such as a SAMD of 0.05 or less (What Works Clearinghouse, 2017), is standard practice when estimating treatment effects in PS analysis. Nevertheless, solely considering how covariates balance after matching may not be enough for justifying the quality of the PS analysis results.

Suitable covariate balance may still yield biased treatment effects (Lee, Lessler, & Stuart, 2010; Stuart, Lee, & Leacy, 2013). More specifically, Belitser et al. (2011) showed that following various balance diagnostic approaches might result in different levels of treatment effect bias. More interestingly, having measurement error in covariates might lead to a problematic estimate of the treatment effect even with a good level of covariate balance (Hong, Aaby, Siddique, & Stuart, 2018). Lastly, it was also shown that the balance of specific covariates could have more influence on the treatment effect bias (Stuart, Lee, & Leacy, 2013). All mentioned evidence from the literature points out the same conclusion: obtaining a good level of overall covariate balance might not be enough for estimating an unbiased treatment effect. There can be several other factors that can deteriorate the estimation of the true treatment effect even with a good amount of overall covariate balance.

Although there have been some studies that have discussed a potential lack of the direct relation between covariate balance and treatment effect bias (Hong, Aaby, Siddique, & Stuart, 2018; Lee, Lessler, & Stuart, 2010; Stuart, 2013), no study to our knowledge has demonstrated this problem by systematically examining it in the context of the aforementioned factors that are important to PS analysis, as well as other key factors such as sample size, the proportion of treatment group, and the association between covariates. Therefore, the current study aims to investigate the inconsistent relation between covariate balance and bias in treatment effect estimation. In other words, this study aims to systematically explore the effects of covariate balance on treatment effect estimation by considering many conditions that applied researchers encounter in their PS analyses. To facilitate a better examination of the current study results, we utilized the PSs estimated from different methods as weights for determining the treatment and control groups across all conditions (Guo & Fraser, 2015; Rosenbaum & Rubin, 1983).

METHOD

A simulation study was conducted to evaluate the performance of three PS estimation methods for recovering the population treatment effect and establishing the covariate balance in terms of means and variances. The three PS methods considered were the traditional LR, GBR, and CBPS. LR is the widely-used method among educational researchers and predicts the PSs through a logistic regression model. The GBR method uses a nonparametric, automated machine learning technique to estimate the PSs and associated weights (Ridgeway, McCaffrey, Morral, Griffin, & Burgette, 2017). The GBR method can predict the treatment assignment using a large number of covariates and is flexible in that it can handle nonlinear relationships between PSs and covariates (McCaffrey, Ridgeway, & Morral, 2004). The CBPS method simultaneously derives the PSs and weights for observations to optimize covariate balance between the treatment and control groups (Fong, Ratkovic, & Imai, 2019). Readers are referred to the cited literature for detailed explanations of the mentioned PS estimation methods.

Simulation Design

In addition to the three PS estimation methods, the current study also investigated the effect of using different types of covariates on the estimation quality of the treatment effect in relation to balance. Twenty-four covariates were classified into three different types (eight covariates per type) depending on their relationship with the treatment assignment and the outcome variable. Types of covariates were referred to as 1) type-W: correlated with both treatment assignment and the outcome variables, 2) type-X: correlated only with the outcome variable, and 3) type-Z: correlated only with the treatment assignment variable. A total of 108 simulation conditions were considered by crossing three PS estimation methods (LR, GBR, and CBPS), three covariate types (type-W, type-X, and type-Z), three sample sizes (500, 1,000, and 5,000), two proportions of treatment group (0.25 and 0.45), and two scenarios for the correlations among the covariates (uncorrelated and correlated).

Simulation conditions were mainly identified based on practical considerations that aim to guide applied researchers. LR was selected to represent the simplistic yet widely-used method among practitioners. CBPS and GBR were selected to represent more advanced methods compared to LR. It is known that CPBS is also a popular method in applied PS analysis studies and GBR is being recognized by many

practitioners who aim to use more advanced methods, namely machine learning-based approaches. Sample sizes were identified to reflect small to extremely large conditions. Group proportions were identified considering scenarios with low and medium/ levels of treatment group availability. It wouldn't be wrong to say that PS analysis studies mostly have larger sample sizes for the control group rather than the treatment group. Thus, we limited the maximum proportion of the treatment group to be 45% to represent a more realistic condition.

Some other magnitudes and/or parameters were fixed in the current simulation. The number of the covariates per type (fixed to eight) was identified randomly yet as a typical size of covariate availability in applied PS analysis studies. The magnitudes of the correlations between covariates were fixed to 0.05 and 0.15 for the uncorrelated vs. correlated covariates conditions. These values were identified in reference to the magnitudes used in other PS analysis simulation studies as well as thinking realistic magnitudes relative to the correlation values between covariates and treatment/outcome variables. We avoided high correlations among the covariates themselves in order to better reveal the effect of covariate and treatment/outcome relation. Lastly, the effect size was fixed to 0.8 and not varied in the current simulation. We chose a relatively high effect size in order to eliminate the side effects of having a small effect during the estimation phase. In other words, we intended to examine the performance of the methods under various conditions when the effect size is already known to be large.

Data Generation

Consider observations of the form $(\mathbf{W}_i, \mathbf{X}_i, \mathbf{Z}_i, B_i, Y_i),\ i = 1, \dots, n$ where $\mathbf{W}_i, \mathbf{X}_i, \mathbf{Z}_i$ are type-W, type-X, and type-Z covariates with size m_W, m_X, and m_Z respectively. As explained above, the number of the covariates for each covariate type was fixed to eight, thus $m_W = m_X = m_Z = 8$. Additionally, B_i is an indicator as to whether an observation belongs to the control ($B_i = 0$) or treatment ($B_i = 1$) group, and Y_i is the outcome of interest. As described earlier, it is assumed that covariates $\mathbf{W}_i$ and $\mathbf{Z}_i$ affect the probability of treatment group membership, while $\mathbf{W}_i$ and $\mathbf{X}_i$ affect the outcome after the group membership has been determined. For notational convenience, let $\mathbf{\Sigma}_W$, $\mathbf{\Sigma}_X$, and $\mathbf{\Sigma}_Z$ denote the covariance matrices of $\mathbf{W}$, $\mathbf{X}$, and $\mathbf{Z}$. Also, let $\mathbf{\Sigma}_{WX}$, $\mathbf{\Sigma}_{WZ}$, and $\mathbf{\Sigma}_{XZ}$ denote the cross-covariance matrices. Note that former matrices contain the covariances between the same-type covariates, whereas the latter ones contain the covariances between different covariate types. For example, $\mathbf{\Sigma}_W$ is the 8x8 covariance matrix of eight type-W covariates. $\mathbf{\Sigma}_{WZ}$ is the 8x8 covariance matrix of eight type-W covariates and eight type-Z covariates. Then $\mathbf{\Sigma}_W$ and $\mathbf{\Sigma}_{WZ}$ matrices are combined for the generation of the treatment group membership. Other matrices can be interpreted in a similar way.

To simulate treatment group membership, let

$$\pi(\mathbf{W}, \mathbf{Z}) = \text{logit}(\alpha_0 + \boldsymbol{\alpha}_W^\top \mathbf{W} + \boldsymbol{\alpha}_Z^\top \mathbf{Z}). \quad (1)$$

Then, for the ith simulated case,

$$B_i \sim \text{Ber}[\pi(\mathbf{W}_i, \mathbf{Z}_i)]. \quad (2)$$

Note that π is the probability of being in the treatment group, and *Ber* stands for Bernoulli distribution. Also, terms with T superscripts refer to the transpose of a relevant matrix. The most important question here is how to choose the constant α_0 in (1), as this controls the proportion of cases that belong to the treatment and control groups. If $[\mathbf{W}_i, \mathbf{Z}_i]$ follows a zero-mean multivariate normal distribution (Φ is the inverse of the cumulative normal distribution function), then the choice

$$\alpha_0 = -\sigma_{WZ} \Phi^{-1}(1 - p) \quad (3)$$

with

$$\sigma_{WZ} = \sqrt{[\boldsymbol{\alpha}_W^\top, \boldsymbol{\alpha}_Z^\top] \begin{bmatrix} \boldsymbol{\Sigma}_W & \boldsymbol{\Sigma}_{WZ} \\ \boldsymbol{\Sigma}_{WZ} & \boldsymbol{\Sigma}_Z \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_W \\ \boldsymbol{\alpha}_Z \end{bmatrix}} \quad (4)$$

will result in a proportion p of the cases being associated with a success probability $\pi(\mathbf{W}, \mathbf{Z})$ greater than 0.5 and a proportion $1 - p$ with success probability less than 0.5. Once the treatment group memberships have been generated, the outcome Y_i is generated according to

$$Y_i = \beta_0 + \beta_T B_i + \boldsymbol{\beta}_W^\top \mathbf{W}_i + \boldsymbol{\beta}_X^\top \mathbf{X}_i + \varepsilon_i, \quad (5)$$

where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ and $\beta_T = D.\sigma_\varepsilon$ with D being the effect size, which is fixed to 0.8, assuming the existence of a large treatment effect. The population values in (4) and (5) were set to $\boldsymbol{\alpha}_W = -1.0$, $\boldsymbol{\alpha}_Z = 1.0$, $\boldsymbol{\beta}_W = 1.0$, and $\boldsymbol{\beta}_X = 2.0$. Note that varying the population values of these parameters regulates the level of association between the covariates and treatment assignment/outcome. We selected these values somewhat arbitrarily yet in the light of the other PS simulation studies, including Brookhart et al. (2006).

The covariance matrices between the same- ($\boldsymbol{\Sigma}_W$, $\boldsymbol{\Sigma}_X$, and $\boldsymbol{\Sigma}_Z$) and cross-type covariates ($\boldsymbol{\Sigma}_{WX}$, $\boldsymbol{\Sigma}_{WZ}$, and $\boldsymbol{\Sigma}_{XZ}$) were varied depending on the magnitude of the relationship among the covariates as a simulation condition. Since all covariates were assumed to have a zero mean and a unit variance, covariance matrices were also the correlation matrices. For the conditions that assumed no relationship among the covariates, all correlations were fixed to zero. Thus, the same- and cross-type matrices were 8x8 identity and 8x8 zero matrices, respectively. For the conditions that assumed a relationship among the covariates, 0.15 and 0.05 were assigned to the correlations among the same- and cross-type covariates and were used to create the relevant covariance matrices.

Two hundred data sets were generated per simulation condition with R (R Core Team, 2018) and analyzed with the relevant PS analysis method, as elaborated in the next section. Note that the data were generated by considering all three types of covariates as predictors of the treatment assignment and outcome variables. In other words, the treatment assignment was generated by using the W- and Z-type covariates (8+8=16 predictors in total), and the outcome measure was generated by using the W- and X-type covariates (8+8=16 predictors in total). During the PS model fitting procedure, however, only one type of covariate (8 in total) was used for each analysis in order to examine the effect of covariate type as a simulation condition.

Analysis

PS weights were computed to estimate the average treatment effect for the treated (ATT), utilizing weighting by the odds (Hirano, Imbens, & Ridder, 2003). Therefore, weights for observations in the treatment group were fixed to be $w_{ti} = 1.0$ and weights for observations in the control group were computed by $w_{ci} = (ps_i)/(1 - ps_i)$, where ps_i is the estimated PS for the ith observation. w_{ti} and w_{ci} values were then used to compute weighted standardized treatment effect for the outcome variable. For covariates, the weights were applied to compute the SAMD and VR as the indicators of covariate balance per covariate. SAMD and the VR values (after adjusting for the treatment/control group as the denominator) for eight covariates were further averaged to obtain the overall balance indicators per simulated data set, which are referred to as average SAMD (ASAMD) and average VR (AVR) in the following sections.

We intended to use the default PS estimation options as much as possible for the three methods in their relevant R functions, considering a typical user without deep knowledge. Nevertheless, we modified some options for GBR method in order to prevent masking its performance compared to simpler methods. The traditional LR method was conducted by the *MatchIt* R package (Ho, Imai, King, & Stuart, 2008), adopting the default specifications for PS estimation with NNM. The GBR method was utilized by the *twang* R package (Ridgeway, McCaffrey, Morral, Griffin, & Burgette, 2017) with default specifications except for the stopping method, which was modified to assess the maximum balance matrix for Kolmogorov-Smirnov statistic. We also modified the estimand option to be ATT (as it was

the adopted method in the current study), which was ATE-Average Treatment Effect by default. The *CBPS* R package (Fong, Ratkovic, & Imai, 2019; Imai & Ratkovic, 2014) was used to estimate the PSs by the CBPS method with default options. Readers are referred to relevant resources for more details about the R functions of each method.

Evaluation Criteria

In an effort to derive a consistent measure of covariate balance and treatment effect, we utilized the *cobalt* R package (Greifer, 2019) to compute the ASAMD and AVR between treatment and control groups per generated data set. The *cobalt* package is commonly used as a supplement to the balance diagnostic tools and provides efficient summary tables of balance diagnostics for each covariate.

For the evaluation of overall covariate balance performance, ASAMD and AVR values were further averaged across 200 data sets generated per simulation condition. Average ASAMD values closer to 0 and average AVR values closer to 1 are considered to be good indicators of overall covariate balance in terms of mean and variance, respectively. Recovery of the population treatment effect ($D = 0.8$) was evaluated by the absolute bias (AB) and standard error (SE), which were calculated by equations (6) and (7), respectively.

$$AB(\widehat{D}) = \left|\overline{\widehat{D_m}} - D\right| \tag{6}$$

$$SE(\widehat{D}) = \sqrt{\frac{\sum_m^M (\widehat{D} - \overline{\widehat{D_m}})^2}{M}} \tag{7}$$

Note that D and $\widehat{D}$ are the population and estimated values of the treatment effect. Also, M represents the total number of replications (200 in our case), m is a specific step of those M replications, and $\overline{\widehat{D_m}}$ is the average of the estimated effect sizes across M replications.

RESULTS

Recovery of the Treatment Effect

It was confirmed that the use of type-W covariates provided the lowest ABs for the recovery of the treatment effect under all simulation conditions (Figure 1). This result is consistent with the existing literature that encourages researchers/practitioners to use covariates that correlate with both outcome and treatment indicators (Brookhart et al., 2006; Hong, Aaby, Siddique, & Stuart, 2018; Myers et al., 2011; Steiner, Cook, Shadish, & Clark, 2010). However, contrary to our expectation, the performance of type-X covariates (correlated with the outcome) was not as good as type-W covariates. Rather, their performance was close to type-Z covariates (correlated with the treatment condition), especially when there was no correlation among covariates. Nevertheless, the ABs were lower for type-X covariates (all below 0.6) than type-Z covariates for all simulation conditions. Lastly, the bias was slightly lower when the type-W covariates were correlated vs. they were not.

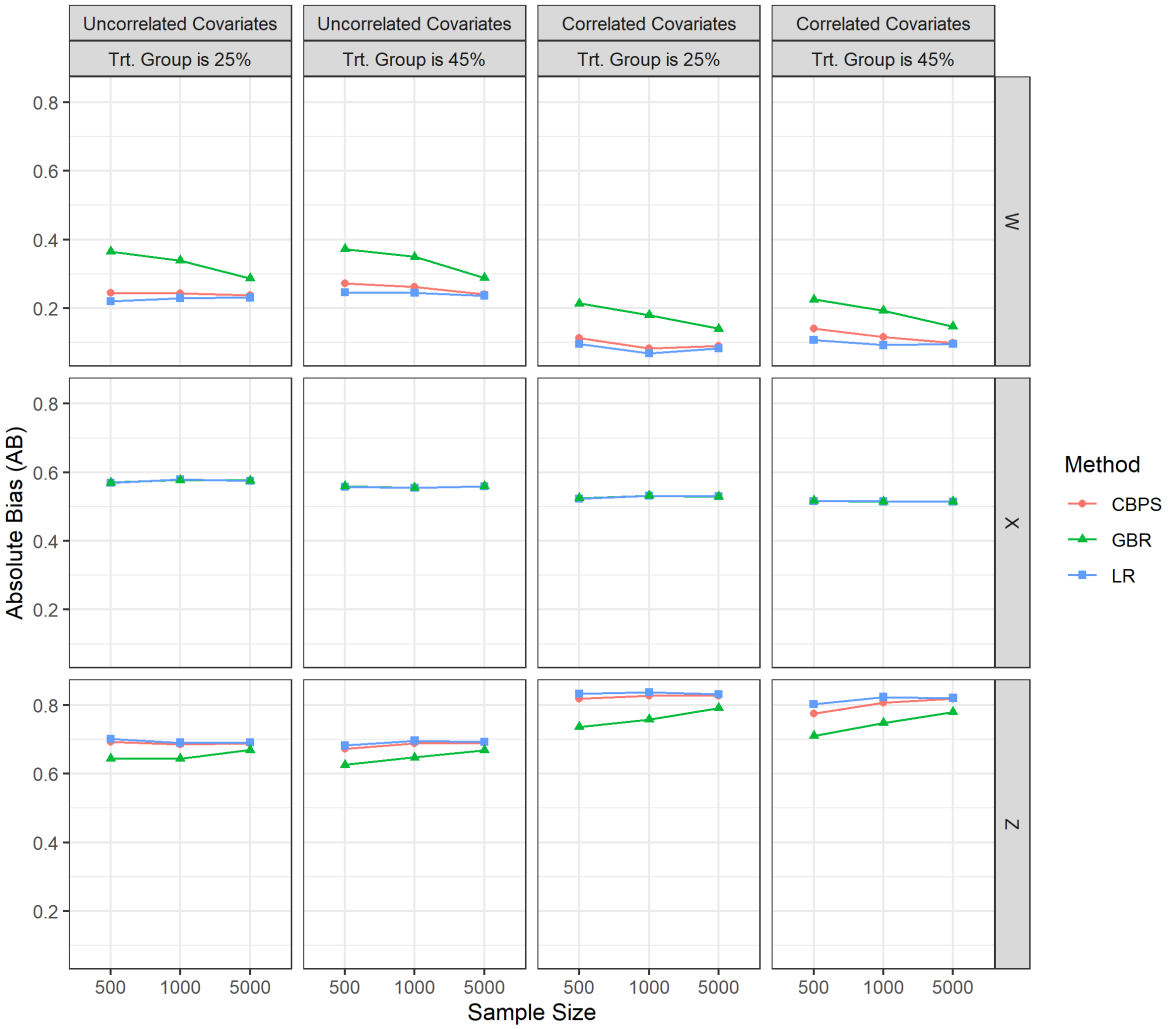

Figure 1. Absolute Bias of the Estimated Treatment Effect

In general, the three PS estimation methods performed similarly for recovering the population treatment effect under common conditions. This was especially true for conditions with type-X covariates. However, for conditions with type-W covariates, GBR did not perform as well as CBPS and LR, while CBPS and LR performed similarly. These tendencies were observed for all conditions, regardless of sample size, proportion of the treatment group, and correlation among covariates. Interestingly, GBR had the lowest AB values when type-Z covariates were used. However, the ABs were large, and the differences between GBR and the other two methods were not large enough to claim that GBR would be useful with type-Z covariates.

Regarding the sample size, there was no clear effect on the AB values for any of the three PS estimation methods. Only a slight decreasing tendency of AB with increased sample size was observed with the use of GBR method in conditions with type-W covariates. On the other hand, a clear effect of sample size was observed for SE values (Figure 2), which were in decreasing trend with the increase of sample size as expected. The proportion of the treatment group also did not show a considerable effect for AB values. Lastly, the effect of the correlation among the selected covariates showed different trends depending on a specific condition. For example, the AB values were lower when type-W covariates were correlated regardless of the proportion of the treatment group. This was also the case for type-X covariates; however, the change was not as clear as for the type-W covariates. An opposite tendency was observed for type-Z covariates. Namely, the AB values were higher when the covariates were correlated for both proportions of the treatment group. Thus, the amount of correlation between type-W

covariates seems to provide extra information for a better recovery of the treatment effect. Conversely, intercorrelated type-Z covariates seem to deteriorate the estimation of the treatment effect. Although we don't have an exact explanation for this phenomenon, we suspect that the intercorrelated treatment assignment predictors led to problematic balancing hence to slightly higher bias in the treatment effect estimation. Moreover, it is known that the use of only type-Z predictors is expected to result in a higher bias in the treatment effect estimation (Patrick et al., 2011). Thus, with the intercorrelated type-Z covariates, this negative effect seemed to be slightly larger.

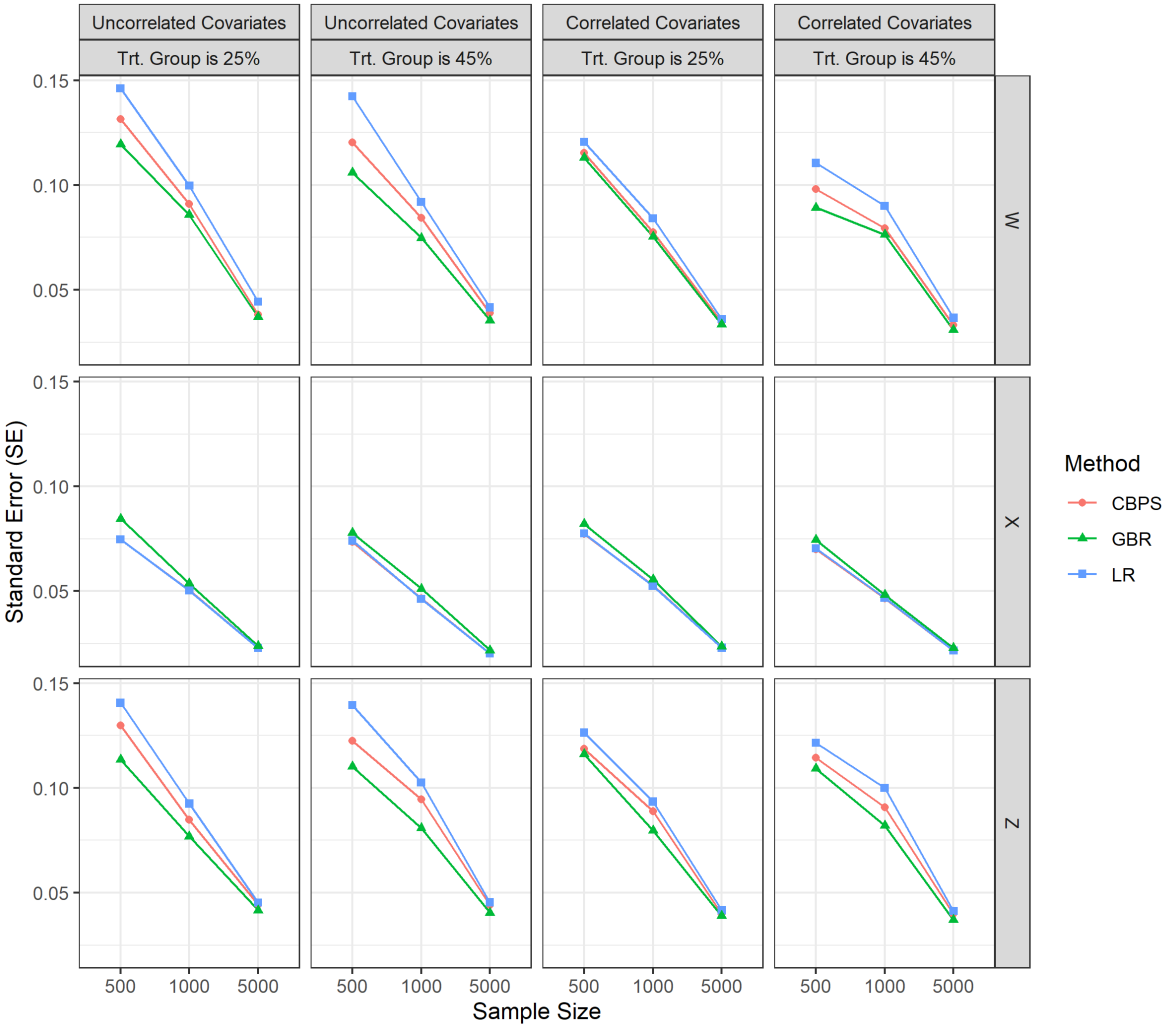

Figure 2. Standard Error of the Estimated Treatment Effect

In summary, our results demonstrated that the use of covariates that are related to both treatment indicator and the outcome resulted in a better recovery of the treatment effect nearly under all studied simulation conditions. Also, it was demonstrated that LR and CBPS produced better performance than GBR when type-W covariates were used. It is important to note that such a result might be explained by the data generation model, where we used only the first-level terms in the logistic regression of the treatment assignment. In other words, no higher-level terms (such as square of the predictors) or interactions were used in the generation of the treatment assignment. Thus, this is in line with the estimation of a plain logistic regression model adopted in LR method. GBR is known to examine also the prediction power of higher-level and interaction terms automatically. Thus, better performance of the LR compared to a more advanced method like GBR might be a result of this. Additionally, if selected

covariates are correlated only with the outcome, selection of the PS estimation method would not matter so much, as the three performed similarly.

Covariate Balance

Means

It was demonstrated that the mean balance was consistently better for the type-X covariates (correlated only with the outcome) under all conditions (Figure 3). Furthermore, the average ASAMD values for the type-X covariates were always below 0.05 regardless of the simulation condition, indicating that they met the threshold by WWC- What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education (2017). Therefore, a practitioner would interpret that the PS analysis went well. This result requires special attention, as it was demonstrated in the previous section that type-X covariates did not produce good AB for the recovery of the population treatment effect, compared to type-W covariates (correlated with the outcome and treatment condition). In other words, these results demonstrate that good covariate balance on means does not necessarily guarantee a good estimate of the treatment effect.

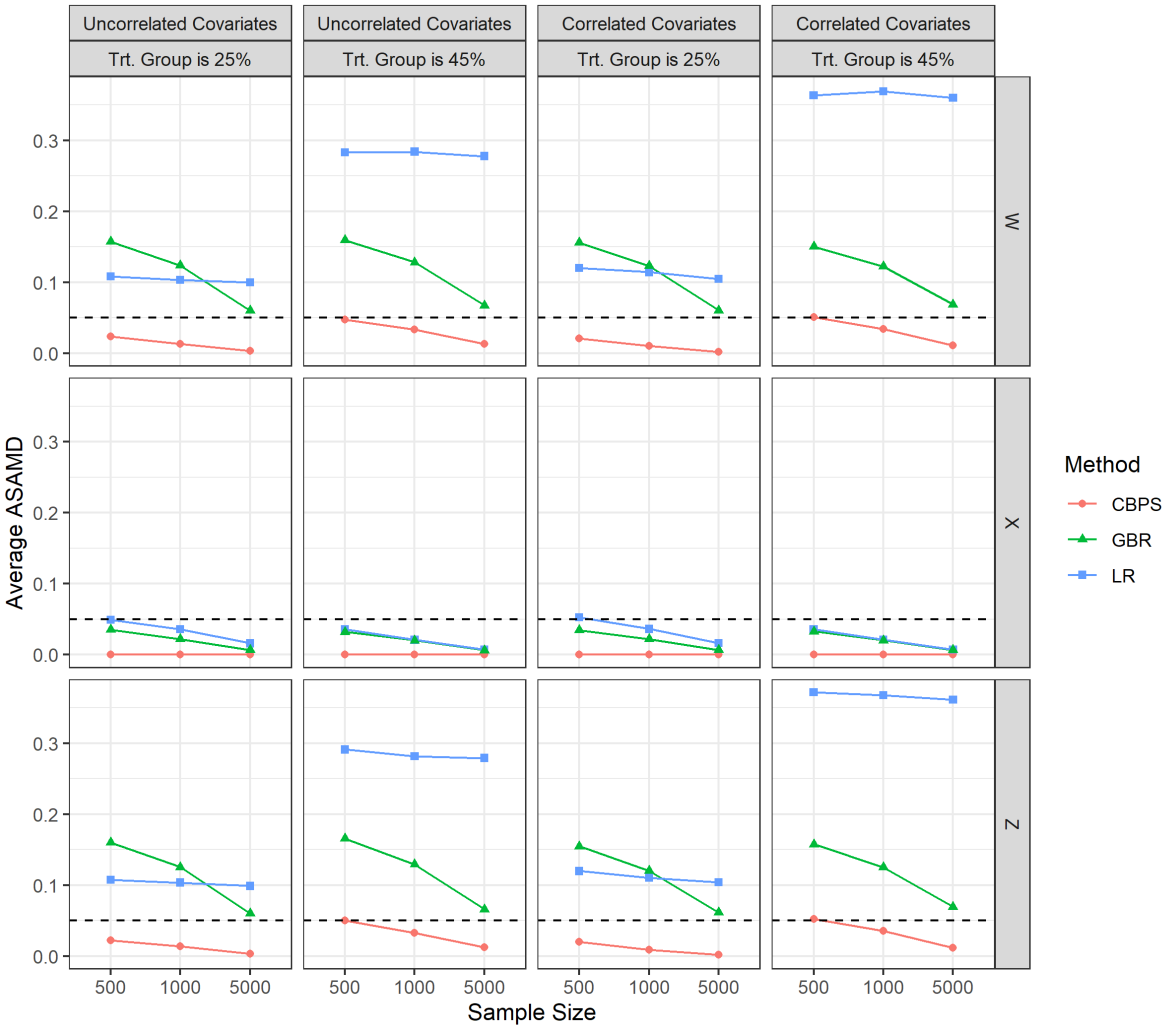

Figure 3. Covariate Balance on Means

On the other hand, when type-W covariates were used, the average ASAMD values met the WWC criterion of 0.05 or lower only in conditions with CBPS as the method used for PS analysis. It was also observed that conditions with GBR approached the WWC criterion with the largest sample size (n = 5,000) by the use of type-W covariates. Nevertheless, none of the two conditions resulted in low AB values, and GBR's AB values were the highest with the use of type-W covariates. Thus, these findings also support the fact that good mean-based covariate balance does not necessarily result in a less biased treatment effect estimate.

Looking at the results with the use of type-Z covariates (correlated only with the treatment condition) a similar trend was observed, where CPBS had the uniformly low average ASAMD values and GBR approached the WWC's criterion with the highest sample size. Nevertheless, the lowest AB was observed consistently for GBR, and surprisingly, the AB values were increased by the increase of the sample size. It is also worth noting that the average ASAMD values for LR were severely affected by the proportion of the treatment group when type-W and type-Z covariates were used. Related to this, LR was not the best performing method in terms of mean balance, compared to its good performance in terms of recovery of the treatment effect. This difference is more important for the use of type-W covariates. A potential explanation for this might be the simplicity of the LR method for computation of the PSs compared to CBPS and GBR. In other words, simple LR does not account for complex relationships during the balancing procedures as CBPS and GBR do.

Variances

Overall, GBR outperformed the other two PS estimation methods, producing better covariate balance with respect to variance ratio. This is not surprising because we set up GBR to derive PSs by evaluating the maximum of the balance matrices based on Kolmogorov-Smirnov statistic, which evaluates the difference in distribution shapes, as opposed to the difference in the means only. Exceptions were observed when the treatment group was 45% and n = 5,000 for type-W (correlated with the outcome and treatment condition) and type-Z covariates (correlated only with the treatment condition), where CBPS produced slightly better variance ratios than GBR. Unlike the mean balance of covariates, it was clear that the variance balance was affected by the sample size (Figure 4). All three PS estimation methods provided better variance ratios in conditions with larger sample sizes. It was demonstrated that sample size mattered more for CBPS than the other two methods.

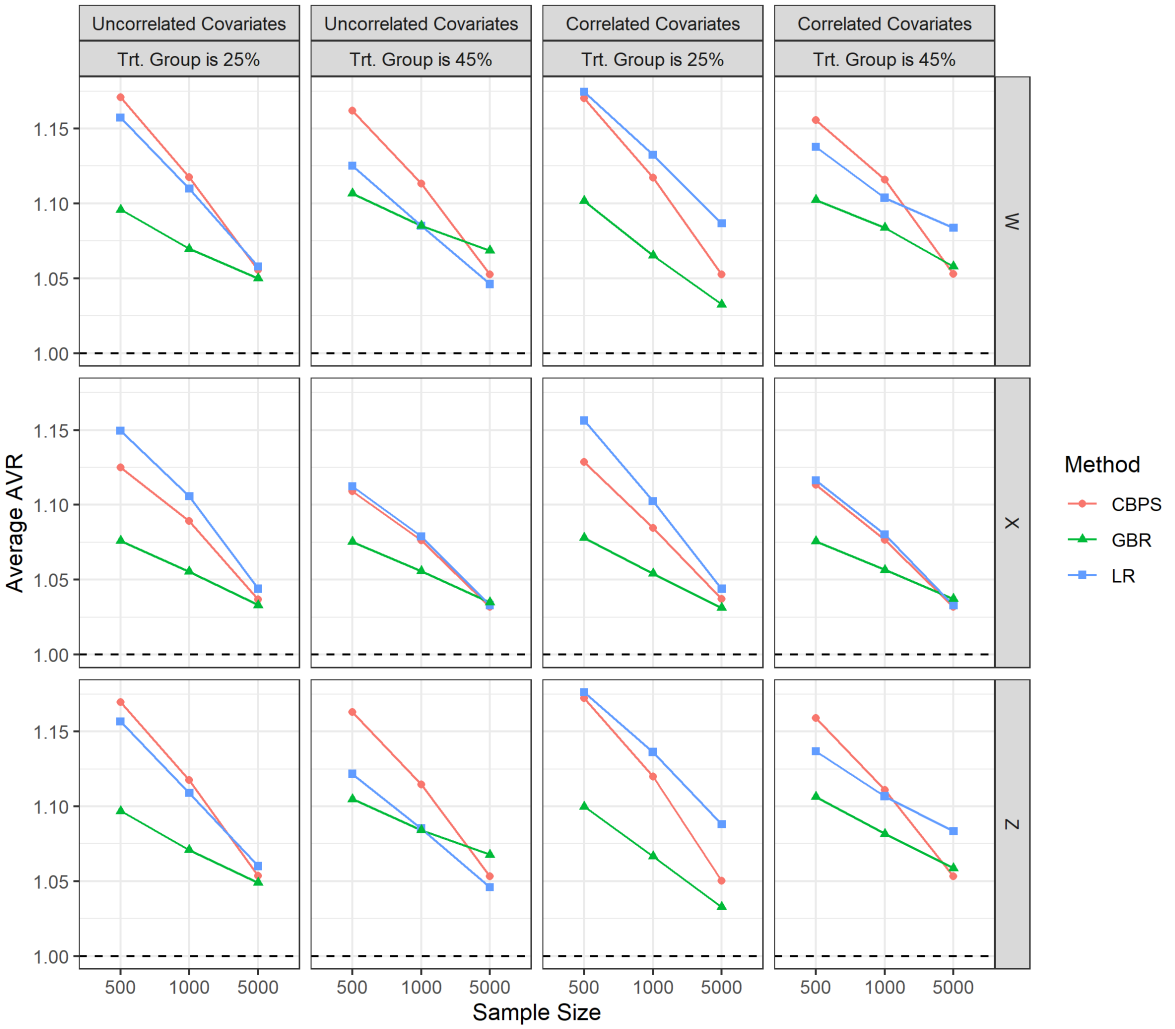

Figure 4. Covariate Balance on Variances

Similar to mean balance results, it was revealed that a PS estimation method that provided good variance balance did not necessarily do the same for the estimation of the treatment effect. Based on this determination, we can summarize our findings under two main statements. First, similar to mean balance results, the variance ratios were generally better for the type-X covariates (correlated only with the outcome) than the type-W and type-Z covariates across all conditions. However, as pointed out in the previous section, type-X covariates did not produce better AB values for the recovery of the treatment effect. Second, although GBR generally provided better variance ratios compared to the other two methods, it did not perform as well, in terms of AB, as CBPS and LR under conditions with type-W covariates. Also, LR provided the worst variance ratios for all conditions with type-X covariates; however, in terms of recovery of the treatment effect, LR performed as well as CBPS and GBR in conditions with type-X covariates. Therefore, based on our simulation results, we cannot conclude that variance ratios provided additional information to identify less biased treatment effect. Nevertheless, they provide more information about covariate balance in addition to mean balance.

DISCUSSION and CONCLUSION

Based on the simulation results, obtaining good covariate balance in terms of mean and variance ratio is likely when covariates are correlated only with the outcome variable (i.e., type-X covariates). The average mean balance across all conditions for this covariate type was below 0.05 on the standardized

scale, which meets the WWC criterion. The variance ratio for type-X covariates was consistently lower than other types of covariates in comparable conditions. However, the recovery of the treatment effect was not the best when only this type of covariates was used. Therefore, researchers/practitioners need to be cautious while using covariates that are mainly correlated with the outcome and not with the treatment assignment. Moreover, it is clear that in applied research that utilizes PS analysis, the true treatment effect will never be known. Thus, practitioners are encouraged to pay as much attention to the characteristics of the covariates as the level of balance they obtain after their selection. If the availability of the covariates is somewhat limited, practitioners can rely on the strengths of other PS estimation methods. For example, it was demonstrated that GBR provided slightly lower AB values when the covariates were the ones that only related to the treatment assignment. Nevertheless, we do not recommend solely relying on such an improvement since the bias values still were high in absolute values.

Also, obtaining a good mean covariate balance is more likely when the CBPS method is used regardless of the covariate type used, whether covariates are correlated or not, sample size, and the proportion of the treatment group. In all of the conditions this study investigated, the mean balance met the WWC criterion with the CBPS. However, it was revealed that the magnitude of AB was affected more by the types of covariates used, as already implied above. When the CBPS was used with covariates that are correlated with the outcome variable and treatment condition (i.e., type-W covariates), AB was quite small, especially when covariates were correlated with each other. However, it is not only for the CBPS; LR performed equally well, and the performance of the GBR was just slightly worse than the CBPS and LR.

Can an examination of the second moment (i.e., the variance ratio) help researchers/practitioners evaluate/predict the quality of an estimated treatment effect? Not likely. When the sample size was large, the variance ratio became very close to 1.05 for CBPS and GBR. However, this happened for all covariate types, including the ones that are correlated only with the treatment assignment (i.e., type-Z covariates). On the other hand, when the sample size was small ($n = 500$), variance ratios for CBPS and LR were large, up to 1.20 in the condition with correlated covariates with a 25% treatment group ratio. In conclusion, we can't recommend that using VR as a complement to ASAMD will be strong enough to predict the performance of the PS methods that would lead to less bias of treatment effect estimation. Rather, practitioners should pay more attention to the characteristics of the covariates they are planning to use for the estimation of PSs rather than solely relying on the level of balance.

It would not be wrong to say that CBPS can be suggested as the optimal method considering the general simulation conditions since it showed the best performance for the mean balance and better or nearly equal performance with two other methods for the recovery of the treatment effect. On the other hand, practitioners who mainly use type-X covariates would feel better by the good mean balance and variance ratio diagnostics they get. Nevertheless, they should be cautious about the estimation of the treatment effect since the type-W covariates (correlated both with the outcome and the treatment condition) showed better recovery results. In conclusion, it can be suggested to use covariates that are equally relevant to the treatment assignment and the outcome.

Limitations and Future Research

As explained in the methods section, specifications of the population values for the coefficients of different covariate types were somewhat arbitrary. It is likely that the results may change based on different specifications of those population values during the data generation phase. This is true for changing either the LR coefficients for the generation of the treatment indicator or the linear model for the generation of the outcome variable. Nevertheless, we tried to assign reasonable values for those parameters depending on their relation with the treatment condition and the outcome variable. We also checked other studies (e.g., Brookhart et al., 2006) as references for identifying typical values that are expected to be encountered in applied research. Also, data generation models did not assume any higher-level terms (e.g., quadratic effects) or interactions between covariates.

The number of covariates were limited to eight in derivations of PSs. Although this number is realistic in many applications of PS analysis, a larger number of covariates may change the results. Also, our simulations investigated three different covariate types (i.e., type-W, type-X, and type-Z) in turn only, meaning none of our PSs were estimated using a combination of different covariate types. Although this does not sound realistic, we intentionally performed that in order to reveal the isolated effect of each covariate type. This also mimics the scenario where applied researchers miss using a specific type of covariates that potentially might change the results of the PS analyses.

Although our study indicates that practitioners utilizing PS analysis should not rely on mean-based covariate balance, it is still unclear which diagnostic measure is ideal when conducting PS analysis. It could be that the ideal diagnostic measure depends on the method used to estimate the PSs (e.g., LR, GBR, CBPS) or on the approach used to apply the PSs (e.g., weighting, subclassification, etc.) in balancing treatment and control groups. The *cobalt* R package is able to work with the PS methods we explored in this study to provide weights. It could be that the "power" behind each method relies on using the weighting values generated from within each method rather than pulling the PSs to generate weights outside of each method. Future studies could investigate how different PS methods in combination with the *cobalt* R package generate weights and how these might affect covariate balance diagnostics and treatment effect bias.

Last, this study systematically demonstrated the effect of various conditions on covariate balance and estimation of the treatment effect through a series of simulated data. While results were quite promising, an empirical data set was not analyzed. We believe that a real data set would be helpful to confirm our simulation findings. Therefore, an empirical data analysis can be considered in a future study by using various PS estimation methods.

REFERENCES

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, *28*(25), 3083–3107. https://doi.org/10.1002/sim.3697

Belitser, S. V., Martens, E. P., Pestman, W. R., Groenwold, R. H. H., Boer, A. de, & Klungel, E. H. (2011). Measuring balance and model selection in propensity score methods. *Pharmacoepidemiology and Drug Safety*, *20*(11), 1115–1129. https://doi.org/10.1002/pds.2188

Bhattacharya, J., & Vogt, W. B. (2007). Do instrumental variables belong in propensity scores? Cambridge, MA: National Bureau of Economic Research (NBER) Working Paper Series No. 343.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*(12), 1149–1156. https://doi.org/10.1093/aje/kwj149

Cannas, M., & Arpino, B. (2019). A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biometrical Journal*, *61*(4), 1049–1072. https://doi.org/10.1002/bimj.201800132

Fong, C., Ratkovic, M., & Imai, K. (2019). *CBPS: Covariate balancing propensity score*. Retrieved from https://CRAN.R-project.org/package=CBPS

Greifer, N. (2019). *Cobalt: Covariate balance tables and plots*. Retrieved from https://CRAN.R-project.org/package=cobalt

Guo, S., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications (ed. 2)*. Thousand Oaks, CA: Sage.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189. https://doi.org/10.1111/1468-0262.00442

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2008). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28. https://doi.org/10.18637/jss.v042.i08

Hong, H., Aaby, D. A., Siddique, J., & Stuart, E. A. (2018). Propensity score-based estimators with multiple error-prone covariates. *American Journal of Epidemiology*, *188*, 222–230. https://doi.org/10.1093/aje/kwy210

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society*, *76*, 243–263.

Kainz, K., Greifer, N., Givens, A., Swietek, K., Lombardi, B. M., Zietz, S., & Kohn, J. L. (2017). Improving causal inference: Recommendations for covariate selection and balance in propensity score methods. *Journal of the Society for Social Work and Research*, *8*, 2334–2351. https://doi.org/10.1086/sim.3782

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*, 337–346. https://doi.org/10.1002/sim.3782

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*(4), 403–425. https://doi.org/10.1037/1082-989X.9.4.403

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., … Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, *174*(11), 1213–1222. https://doi.org/10.1093/aje/kwr364

Patrick, A. R., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., Rothman, K. J., Avorn, J., & Sturmer, T. (2011). The implications of propensity score variable selection strategies in pharmacoepidemiology: An empirical illustration. *Pharmacoepidemiology and Drug Safety*, *20*(6), 551–559. https://doi.org/10.1002/pds.2098

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., & Burgette, L. (2017). *Twang: Toolkit for weighting and analysis of nonequivalent groups*. Retrieved from https://CRAN.R-project.org/package=twang

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*(3), 169-188. https://doi.org/10.1023/A:1020363010465

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine, 26*(1), 20–36. http://dx.doi .org/10.1002/sim.2739

Setoguchi, S., S amd Schneeweiss, Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluation uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology & Drug Safety*, *17*(6), 546–555. https://doi.org/10.1002/pds.1555

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selectin bias in observational studies. *Psychological Methods*, *15*(3), 250–267. https://doi.org/10.1037/a0018719

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313

Stuart, E. A, Lee, B. K., & Leacy, F. P. (2013). Prognostic score–based balance measures for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology*, *66*(8), S84–S90. https://doi.org/10.1016/j.jclinepi.2013.01.013

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: Machine learning and classification methods as alternatives to logistic regression. *Journal of Clinical Epidemiology*, *63*(8), 826–833. https://doi.org/10.1016/j.jclinepi.2009.11.020

What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. (2017). *What works clearinghouse: Procedures and standards handbook (version 4.0)*. Retrieved from http://whatworks.ed.gov