Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Understanding how balance and sample size impact bias in the estimation of causal treatment effects: a simulation study

Andreas Markoulidakis (MarkoulidakisA@cardiff.co.uk)
Cardiff University
Peter Holmans
Cardiff University
Philip Pallmann
Cardiff University
Beth Ann Griffin
RAND Corporation

Research Article

Keywords: propensity score, balancing weights, balance threshold, variable selection, sample size

Posted Date: June 16th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1742290/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

RESEARCH

Understanding how balance and sample size impact bias in the estimation of causal treatment effects: a simulation study

Andreas Markoulidakis[1,2,5*], Peter Holmans[3,6], Philip Pallmann[2,7] and Beth Ann Griffin[4,8]

*Correspondence:
MarkoulidakisA@cardiff.co.uk
[1]School of Medicine, Cardiff University, Cardiff, UK
Full list of author information is available at the end of the article

Abstract

Observational studies are often used to understand relationships between exposures and outcomes. When analyzing such data, propensity score (PS) and balance weighting are commonly used techniques that aim to reduce the imbalances between exposure groups by weighting the groups to look alike on the observed confounders. There is now a plethora of available methods to estimate PS and balancing weights as well as rich guidance on how to properly employ these in an analysis. In such studies, unbiased and robust estimation of the causal treatment effect is not guaranteed unless several conditions hold. The literature has shown that accurate inference requires these key criteria: 1) the treatment allocation mechanism is known (e.g., no unobserved confounders), 2) the relationship between the baseline covariates and the outcome is known (if one is to rely on the outcome model itself), 3) adequate balance (comparability) of baseline covariates is achieved post-weighting, 4) a proper set of covariates to control for confounding bias is known, and 5) a large enough sample size is available. In this article, we use simulated data of various sizes to investigate the influence of these five criteria on statistical inference. We have two notable new findings to help improve practice. First, our findings provide important new evidence that the maximum Kolmogorov-Smirnov statistic is the proper statistical measure to assess balance on the baseline covariates, in contrast to the mean standardised mean difference used in many applications, and 0.1 is a suitable threshold to consider as acceptable balance. Second, we also find a clear recommendation that $60 - 80$ observations, per confounder per treatment group, are required to obtain a reliable and unbiased estimation of the causal treatment effect.

Keywords: propensity score; balancing weights; balance threshold; variable selection; sample size

1 Introduction

Randomized controlled trials (RCTs) are the gold standard in the estimation of causal treatment effects in clinical research, because randomization of a large enough sample usually creates two groups that are well-balanced on both observed and unobserved pre-treatment confounders. Observational studies on the other hand compare two or more groups (e.g., *treatments* and *control*) where the allocation mechanism of individuals to groups is unknown, and typically not random. Observational studies are employed in a range of circumstances, such as where an RCT design is not feasible or where there are routine datasets or cohorts providing valuable information about outcomes in relation to certain exposures. Group assignment

might be due to factors that the researcher cannot control, such as underlying conditions of the individual. These differences between the groups could induce bias in terms of estimated effects of treatment. For example, consider a study that aims to estimate the causal effect of physical activity on disease progression among a sample of individuals living with a particular disease using observational study data. The likelihood that an individual engages in regular physical activity at study entry might in part be influenced by their disease severity (e.g., those with more severe disease may struggle with motivation and/or ability to engage in regular physical activity) and any subsequent comparison between those who did and did not exercise will be distorted, as it will compare groups with different pre-treatment levels of disease severity, potentially overstating the benefits of exercise for individuals living with the disease.

The estimation of accurate causal treatment effects [1] in observational studies is heavily discussed in the literature, with propensity score (PS) [2] techniques being extensively used to reduce the confounding bias. The PS [2] is the probability of an individual's allocation to the treatment group, given their observed baseline (pre-treatment) characteristics. The PS can be used to create pseudo-randomized comparable treatment groups by either weighting, matching, adjusting, or stratifying on the PS. PS methods reduce the bias in the estimation of the causal treatment effect due to known and observed confounders by minimizing the imbalance of these between the treatment groups.

We focus our current study on PS and balancing weighting, although the conclusions are likely to generalize to other uses of the PS (such as matching or stratifying) and balancing weights. Despite the wealth of investigation into PS weighting [3, 4, 5, 6], there are still a number of uncertainties that limit best practice recommendations. Accurate inference based on observational data is conditional on several assumptions (e.g., positivity and unconfoundedness) which, if met, guarantee an unbiased and robust estimation of the causal treatment effect. The majority of methods [2, 3, 4, 5, 6, 7] rely in some way on the following key pieces to be understood: 1) the treatment allocation mechanism is known, 2) the relationship between the baseline covariates and the outcome is known, 3) adequate balance of baseline covariates is achieved post-weighting, 4) a proper set of covariates to control for confounding bias is known, and 5) a large enough sample size is available. In this article, we use simulated data, based on two previous studies [8, 9], to examine what extent of compliance with these five conditions is necessary for unbiased causal effect estimation. We find important new recommendations for the field regarding how best to assess balance and the required sample sizes needed to properly employ balancing weights.

In terms of assessing the impact of sample size on the accuracy of the estimation of the causal treatment effect, we note that there has been a rich body of literature using the same simulation models as used here to investigate the performance of different PS and balancing weights estimation algorithms [8, 10, 11, 12, 13, 14, 15, 16]. In each case, the studies typically only used two different sample sizes ($n = 2000$ and $n = 10000$ observations in total). Related work [12, 15, 16, 17] investigating the relative strength of different PS and balancing weights estimation algorithms also focused on the performance of the algorithms for large samples,

giving little attention to issues arising in the analysis of small samples. Based on these studies, a sample size of $n = 2000$ appears sufficient to effectively estimate the causal treatment effect, even when the true outcome model deviates significantly from the one used in the statistical analysis. There is little discussion in the literature about the potential performance of different PS and balancing weight methods when applied to smaller samples [7], which are usual when studying rare diseases or when there are limited means to collect data. Here, we simulated samples of several sizes (from $n = 40$ to 2000), in order to determine the minimum sample size needed per confounder and treatment group that allows for weighting methods to achieve a reasonably small bias in the estimated causal treatment effect. In the occasions that small sample sizes are considered [7, 18], the true propensity score and outcome models (the true underlying relationship between the baseline covariates and the treatment/outcome) are simple logistic regression and/or linear regression with the main effects only. In such cases, there is evidence that even a regression model with no control for confounding bias could yield an estimation of the causal treatment effect with low relative bias [8, 9]. In this paper, we consider more complicated treatment allocation and outcome models, in a way that the logistic and linear regression models — that are traditionally used for PS and outcome estimation, respectively — do not reflect the truth. Our inspiration for exploring such small sample sizes is the limitations that often arise in real-world data, especially when they are related to rare diseases [19, 20] and there is a large number of observed confounding variables.

Additionally, even though PS and balancing weights have been used for decades [8, 9, 21], the definition of adequate balance is not yet well established across the community. The most common measure of balance used in the literature is the standardized mean difference (SMD) between treatment and control groups for each of the observed pre-treatment confounders used in the analysis [21, 22, 23, 24, 25, 26]. In this study, we investigate the strengths and limitations of using the SMD versus the Kolmogorov-Smirnov (KS) statistic [27], which is less frequently used. We also investigate optimal thresholds to minimise bias in treatment effect estimation for both statistics. Typically, a 0.2 [10] balance threshold for the SMD is suggested as sufficient to guarantee unbiased estimation of the causal treatment effect, but our work shows that stricter rules should apply [8, 21]. Notably, our findings provide important new evidence that the maximum KS statistic is a more appropriate statistical measure to assess balance on the baseline covariates, in contrast to the mean SMD used in many applications, and 0.1 is a suitable threshold to consider as acceptable balance for the KS statistic.

Selecting which covariates to control for in the PS and balancing weights is an issue that has been well discussed previously, with most authors arguing in favor of controlling for the true confounders as well as covariates causally linked only to the outcome since the inclusion of covariates related only to the treatment allocation could inflate the bias of the estimate and the mean squared error (MSE) [28, 29, 30, 31]. We also study the role of variable selection in our simulations with more restricted sample sizes by evaluating the impact of every possible combination of confounders on the outcome estimation, making suggestions regarding which sets of covariates should be prioritized for controlling.

Our study considers different sets of covariates to control for confounding bias, thus revealing the caveats that occur when *wrong* covariates are chosen to control for confounding bias and also how misleading sometimes the balance measures could be, if covariates not related to the outcome are deployed. All these sets of covariates are evaluated across datasets of different size, to provide approximate guidelines regarding the number of participants one needs for each confounder to achieve adequate balance.

The remainder of the article is organised as follows. *Section* 2 provides a brief overview of the basic principles for estimation of the causal treatment effect. In *section* 3 we briefly review the algorithms that we use to obtain PS and balancing weights and the statistics to evaluate the balance. *Section* 4 discusses the choice of baseline covariates used in the estimation of the PS and balancing weights. In *section* 5 we define the different sets of confounders used in the simulations and present the data simulation framework. We report and discuss our findings in *sections* 6 and 7, respectively.

2 Overview of Causal Modeling

In the causal inference framework, we are often interested in evaluating the performance of a new treatment compared to one traditionally used or an untreated situation, which is referred to as reference or control treatment. The measure of the difference between the two is usually called the *average treatment effect* [1]. In this article, we focus on studies with two groups, namely control and treatment.

In Rubin's causal model (RCM), a formal mathematical framework for causal inference [2, 32, 33, 34, 35], every individual $i = 1, ..., n$ has a potential outcome conditional on the presence of treatment ($\{Y_{i|T=1}\}$), and a potential outcome conditional on the absence of treatment ($\{Y_{i|T=0}\}$). The treatment effect for each individual is defined as the difference between the two potential outcome effects

$$\tau_i = Y_{i|T=1} - Y_{i|T=0}. \tag{1}$$

It is impossible to observe both outcomes for the same individual since every individual is exclusively assigned either to the control or treatment group. Thus, we are able to observe the outcome in the presence of treatment $Y_{i|T=1}$ only for individuals in the treatment group ($Y_{i|T=1}|T_i = 1$) and the outcome in the absence of treatment $Y_{i|T=0}$ only for individuals in the control group ($Y_{i|T=0}|T_i = 0$).

In this article, we estimate the average treatment effect on the treated population (ATT) [2, 36], which is defined as

$$ATT = E[Y_{i|T=1}|T = 1] - E[Y_{i|T=0}|T = 1],$$

where $E[Y_{i|T=1}|T = 1]$ is the potential outcome under treatment, given that the individual receives treatment, while $E[Y_{i|T=0}|T = 1]$ is the potential outcome under control, given that the individual receives treatment. The latter quantity is not possible to be observed — it represents the hypothetical outcome of an individual of the control group, had it been assigned to the treatment group. Since it is impossible to observe this quantity, balancing weights are used to adjust the observed outcomes

of the control group, supposed they received treatment, to match the baseline characteristics of the treatment group — in case of ATT estimation, balancing weights are only computed for the individuals of the control group [36].

ATT measures the effect of the treatment only among individuals similar to those in the treatment group. Other quantities of interest are the average treatment effect on the entire population (ATE) and the average treatment effect on the control population (ATC), which express the average causal treatment effect for the entire population of individuals in both the treatment and control groups and the effect of the treatment among only individuals similar to those in the control group, respectively.

In order to obtain accurate estimates of $E[Y_{i|T=1}|T=0]$ and $E[Y_{i|T=0}|T=1]$, RCM requires two assumptions to hold, named *strong ignorability* and *stable unit treatment value assumption* (SUTVA) [37].

Strong ignorability assumes that the treatment assignment mechanism is independent of the distributions of the outcome, given the observed covariates X on the baseline ($Y_{i|T=0}, Y_{i|T=1} \perp T|X$ — unconfoundedness). This assumption also requires that every individual, given the values of the covariates X on the baseline, has a positive probability greater than 0 and less than 1 to be assigned either on treatment or control group ($0 < P(T_i = 1) < 1$ for every i). In other terms, the individuals in the two treatment groups should overlap on the baseline characteristics, thus there are representatives from both groups for each value of baseline covariates [36]. If the control and intervention group are properly balanced after the weighting, this is a sufficient indication that strong ignorability has been achieved on the treatment assignment given the observed covariates.

SUTVA states that the outcome value $Y_{i|T=k}$ corresponding to individual i for treatment k ($k \in \{0, 1\}$) is unique. This assumption implies that the distribution of potential outcomes for each individual is independent of the potential outcome of another individual. Additionally, it implies that all possible values of treatment status are represented [38, 36].

Inverse probability weighting (IPW) [39] assigns a weight to each individual, and then the treatment effect is computed from the weighted mean effect on control and treatment group, as follows[1]:

$$E[\widehat{Y_0}] = \frac{\sum_{i=1}^{n_0} w_i^0 Y_{i|T=0}}{\sum_{i=1}^{n_0} w_i^0}, \quad E[\widehat{Y_1}] = \frac{\sum_{i=1}^{n_1} w_i^1 Y_{i|T=1}}{\sum_{i=1}^{n_1} w_i^1}. \tag{2}$$

These weights are often derived from an individual's PS [5, 40, 41], however, recent algorithms allow direct estimation of the weights [42], subject to a set of restrictions, to achieve better balance among the treatment groups — see section 3.1.

It is very common practice to use a multivariable regression to adjust for the weights, which ideally includes all of the observed confounders used in the estimation of the weights [41, 43] — and potentially other covariates that were not used to control for confounding bias. When using PS weights, this approach is called

[1] In case of ATT estimation, the quantity $E[\widehat{Y_1}]$ is the sample mean of the treatment group, as all the weights $\{w_i^1\}_{i=1}^n$ are equal to one [36].

a *doubly robust estimator* of the causal treatment effect [44, 45, 46, 47]. The estimated treatment effect is unbiased so long as either the PS weight model or the multivariable outcome model is correctly specified [36]. In this article, we use this method for the estimation of the causal treatment effect.

3 Weighting & Balance Evaluation

3.1 Propensity Score & Balancing Weights

We will use four main algorithms to compute PS and balancing weights [36], named Logistic Regression (LR) [48, 49], Covariate Balance Propensity Score (CBPS) [40], Generalized Boosted Model (GBM) [50] and Entropy Balance (EB) [42].

Logistic Regression (LR) LR is the simplest and most commonly used parametric method to estimate probabilities and model binary outcomes, thus it is helpful to estimate the PS of each individual [48] since treatment assignments are often binary. The LR model for estimating the PS assumes that the *logit* of the probability of receiving treatment is a linear combination of the covariates [2]. The main caveat with the LR is that the PS model can often be misspecified, leading to biased estimates of treatment effects. More complex relationships of the treatment with the baseline covariates could be considered, including higher orders and/or interactions, which would have to be specified in the model.

Covariate Balancing Propensity Score (CBPS) CBPS [40] is a parametric method which is deployed to overcome some of the potential misspecifications of the LR. The CBPS algorithm still assumes the same relationship between the treatment status and the baseline covariates as the LR, however, it adds further constraints to achieve a good balance between the treatment groups. The set of constraints considered in the basic version of CBPS aims to balance the first moment of the distributions of the baseline covariates of the two groups (the means). As a consequence, this method is robust to mild model misspecification about balancing confounders compared to standard LR [8, 11, 15, 16, 40].

Further extensions of CBPS can achieve balance of higher moments [36, 51]. It is possible to impose restrictions that require the first m moments to be matched, by replacing the original covariates with orthogonal polynomials of degree m. For instance, imposing restrictions to match the first moment is equivalent to match the mean value of the treatment groups, the second order corresponds to matching the means and variances, and so on. This transformation is possible only on continuous covariates, and it further increases the number of confounders that the PS model should control for — e.g. if we have five continuous confounders, setting $m = 3$ would increase the number of confounders from five to $3 \cdot 5 = 15$.

In this study, we consider three versions of the CBPS algorithm, controlling for $m = 1, 2$ and 3 moments (denoted by $CBPS\#1$, $CBPS\#2$ and $CBPS\#3$, respectively) [36].

[2] An extension of LR is Multinomial Logistic Regression which could be used to estimate generalized PS if there were more than two treatment conditions.

Generalized Boosted Model (GBM) GBM is a non-parametric machine learning approach, which is often used to estimating PS weights. It predicts the binary treatment indicator by fitting a piecewise-constant model, constructed as a combination of simple regression trees (Burgette, McCaffrey, and Griffin in press, [52, 53, 41]), namely *Recursive Partitioning Algorithms* and *Boosting*. To develop the PS model, GBM uses an iterative, *forward stagewise additive algorithm*. Starting with the PS equal to the average of treatment assignment on the sample, the algorithm starts by fitting a simple regression tree to the data to predict treatment from the covariates by maximizing the function

$$l(x) = \sum_{i=1}^{N} T_i g(X_i) - log(1 + exp(g(X_i))), \tag{3}$$

where $g(X_i)$ is the *logit* of treatment assignment. In each iteration the algorithm splits the nodes of the tree with respect to the criterion that one wishes to minimize [41] (this is usually the *mean SMD* or the *maximum KS* value). The algorithm will stop either when adequate balance has been achieved, or the maximum number of iterations has been reached.

Entropy Balancing (EB) EB [42] is a method used to estimate the balancing weights directly rather than the PS of the individuals. The method attempts to achieve exact balance (difference of moments across the treatment groups equal to 0) on as many moments as defined by the user. EB calculates weights through a re-weighting scheme until adequate balance in the pre-selected moments is achieved, attempting to match the first m moments of the distributions of the two groups.

In this study, we consider three versions of the EB algorithm, controlling for $m = 1, 2$ and 3 moments (denoted by $EB\#1$, $EB\#2$ and $EB\#3$, respectively) [36].

3.2 Balance Evaluation

Once we have an estimation of the PS and balancing weights, it is important to evaluate the balance on the two groups achieved. To do so we will use the *standardized mean difference* (SMD) and *Kolmogorov-Smirnov statistic* (KS).

The SMD [22, 23] is a measure of the distance of the means of the two groups. It is defined as the difference of the means, divided by an estimate of the standard deviation for a given covariate. We will concentrate on the *Absolute SMD*, whose values are non-negative, with lower values corresponding to better balance — the lower the value of the SMD, the closer the mean values of the confounders for the two treatment groups are. Initially, 0.2 was recommended as a threshold [10] to define groups as balanced. However, there is more recent evidence that a more conservative threshold (0.1) should be considered [21, 22, 24, 25, 26, 36].

KS is a test statistic used in a procedure [27, 36] which tests the hypothesis that two samples are from the same distribution. It takes values in $[0, 1]$, and lower values indicate that the two distributions are more similar. In contrast to SMD, KS quantifies the similarity of the entire distribution of the two groups, rather than the means only. There is no clear guidance on what is the best threshold for the KS but values over 0.1 would be considered notably large and thus, we propose to use 0.1

as the threshold for balance for the KS as well as the SMD. It is also expected that once balance across distributions of the groups has been achieved (KS value under 0.1), balance across the means is also held (SMD value under 0.1). Additionally, it is usually easier to achieve an acceptable level of balance on SMD rather than KS statistic, since the former concerns only the mean value of the groups (one-dimensional measure), while the latter concerns the entire distribution of the two groups.

Finally, we compute the *effective sample size* [41] (ESS) for each algorithm. This is a measure of the sample power lost by weighting, and corresponds to the sample size of an unweighted analysis that would give the same power as the weighted analysis.

In this paper, we use Logistic Regression (LR), Generalized Boosted Model minimizing the mean SMD value (GBM_{ES}), Generalized Boosted Model minimizing the maximum KS value (GBM_{KS}), Covariate Balancing Propensity Score controlling for moments $m = 1, 2, 3$ ($CBPS\#1$, $CBPS\#2$, $CBPS\#3$, respectively), and Entropy Balancing controlling for moments $m = 1, 2, 3$ ($EB\#1$, $EB\#2$, $EB\#3$, respectively). $CBPS\#2$, $CBPS\#3$, $EB\#2$ and $EB\#3$, attempt to control for higher moments than just the mean, so this could match the higher-order moments existing on the true relationship between baseline covariates and the treatment allocation. Since our aim is to provide general guidelines for balance, we will use all the algorithms used in [36] in our analyses. We do not intend to compare the performance of the algorithms, and thus we do not explicitly recommend one algorithm over the others. Instead, we encourage users to try more than one algorithm to evaluate balance on the baseline covariates and use the one that achieves the best trade-off between balance and ESS to model the outcome [36].

4 The Impact of Variable Selection

The lack of random allocation mechanism among the treatment groups in observational studies [1] incurs confounding bias, which PS and balancing weights are intended to minimize [2]. A key consideration when deploying PS and balancing weights to control for confounding bias is the selection of variables that will be considered as confounders (predictors of both the outcome and the treatment status) [54]. The impact of variable selection on the estimation of causal treatment effects in observational studies, has been widely discussed in the literature [8, 9, 54, 28, 30, 29, 31, 55, 56, 57], however, these studies typically discuss it in the context of a large sample. Here, we will use different combinations of sets of covariates to verify these findings on smaller samples.

The covariates that affect both the treatment allocation and the outcome value [54] are the *true confounders* and should always be included in the estimation of the PS and balancing weights [54, 28, 29, 31].

In the early stages of using PS and balancing weights to control for confounding bias, it was made explicit that all true confounders should be included in the treatment allocation model estimation [28, 29, 31] — the model used for the estimation of the PS or balancing weights. However, recent literature suggests that including covariates that are highly related to the treatment allocation but not to the outcome could increase the bias of the causal treatment effect estimation [30].

Also, covariates related to the outcome, but not to the treatment allocation should be included in the model, even though they would not decrease the bias, but they could rather increase the precision of the estimated treatment effect [28, 55]. Thus, it is suggested that it is preferable to control for covariates that are related to the outcome only, rather than related only to the treatment allocation [56] — of course, true confounders should always be prioritized for inclusion in the treatment allocation model.

Others suggest including in the PS model all the baseline covariates that are not balanced [57] — this is all the baseline covariates with an SMD over 0.1. This suggestion is dubious, since it is possible that covariates related to the treatment allocation but not the outcome have SMD over 0.1. Including such a covariate could increase the bias [30], which is not a desirable outcome. Furthermore, including all unbalanced covariates is not always practical, especially when the sample size imposes limitations as to how many covariates the PS and balancing weights could attempt to balance. As a consequence, general guidelines are required as to which covariates to prioritize in the PS model.

It is still unclear whether the covariates used in the treatment allocation model and the outcome model should be the same [1, 2], although in most applications this is the case [8, 9]. Hirano and Imbens [29] made explicit that the covariates of the two models need not necessarily be the same. This is also implied by their proposed procedure to choose covariates for each regression model. They propose a step-wise method for variable selection, where covariates are added to and removed from each model until only statistically significant covariates are included in each model. Since the procedure is performed separately for the treatment allocation and the outcome model, and the choice of covariates is only based on their statistical significance as this is measured by the model, it is apparent that the covariates included in the two models are not necessarily the same.

5 Overview of Simulation

To investigate the features we are interested in, we use simulated data based on a simulation study performed by Setoguchi and colleagues [9]. The analysis considers a binary treatment variable T and ten pre-treatment covariates $(X_1, X_2, ..., X_{10})$.

To generate the ten pre-treatment covariates $(X_1, X_2, ..., X_{10})$, initially we generate six independent covariates following a standard normal distribution $(X_1, X_2, X_3, X_4, X_7, X_{10})$, with zero mean and unit variance. The remaining four covariates (X_5, X_6, X_8, X_9) are generated using the correlation matrix:

$$\rho = \begin{bmatrix} 1 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0.9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0.2 & 0 & 0 \\ 0.2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0.9 & 0 \\ 0 & 0.9 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0.9 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Finally, six covariates are converted to binary variables $(X_1, X_3, X_5, X_6, X_8, X_9)$ — using random numbers between 0 and 1 (see the *Appendix*).

5.1 Treatment Allocation

The binary treatment variable, T is modeled using *logistic regression* as a function of X_i. The formula used to compute the true propensity score ($P(T|X_i)$) is

$$P[T = 1|X] = \frac{1}{1 + e^{-(A+B+C)}} \tag{4}$$

$$\begin{aligned}
A &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7, \\
B &= 0.5 \cdot \beta_1 X_1 X_3 + 0.7 \cdot \beta_2 X_2 X_4 + 0.5 \cdot \beta_3 X_3 X_5 + 0.7 \cdot \beta_4 X_4 X_6 + 0.5 \cdot \beta_5 X_5 X_7 \\
&\quad + 0.5 \cdot \beta_1 X_1 X_6 + 0.7 \cdot \beta_2 X_2 X_3 + 0.5 \cdot \beta_3 X_3 X_4 + 0.5 \cdot \beta_4 X_4 X_5 + 0.5 \cdot \beta_5 X_5 X_6, \\
C &= \beta_2 X_2 X_2 + \beta_4 X_4 X_4 + \beta_7 X_7 X_7,
\end{aligned}$$

corresponding to *scenario G* of Setoguchi *et al.* [9], with strong non-linearity and non-additivity. The true values of the coefficients used are:

$$(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7) = (0, 0.8, -0.25, 0.6, -0.4, -0.8, -0.5, 0.7).$$

Setoguchi *et al.* [9] and others [10, 11, 12, 13, 14, 15, 16] typically used seven treatment models, each expressing a different strength of linearity and additivity, since the main goal of those articles was to explore the relative performance of several different PS and balancing weights estimation algorithms. Our goal is not to determine which algorithm performs best in terms of PS estimation, but to assess the overall quality of the results produced by the different candidate methods based on different levels of balance, sample size, and sets of confounders. Thus, we focus on only one treatment allocation mechanism — one that is non-linear and non-additive, in order to ensure our findings reflect potential modeling issues encountered in real world applications.

For the simple linear and additive treatment allocation model (this corresponds to *Scenario 1* of Setoguchi *et al.* [8, 9]), every algorithm can provide a valid model for the estimation of PS and balancing weights. The doubly-robust method traditionally used for the estimation of the causal treatment effect assumes that at least one of the treatment allocation and outcome models is properly specified. In real-world study data, it is often impossible to know the exact relationship between the confounders and the treatment and outcome covariate. Thus, one typically uses a simple linear (or logistic) regression model to express the relation of the outcome with the confounders when the outcome is continuous (binary, respectively) — unless more information on the mathematical relationship between the outcome and the baseline covariates is available. Modern algorithms provide a wider range of parametric [5, 48, 40, 42] and non-parametric [41] functions to model the relationship between the confounders and the treatment, however, it is still not possible to test whether the assumptions of the treatment allocation model hold.

5.2 Outcome Models

In this article, we consider four different models to generate the outcome. We do so to understand the sample size and balance level needed to obtain a good estimation of the causal treatment effect under a range of outcome models. In most applied analyses, researchers will tend to use additive, parametric models to estimate the causal effects of a binary treatment on outcomes (e.g., linear regression models to model continuous outcomes and logistic regression to model binary outcomes which control for main effects of the key confounders) [9, 39, 58] and we want to understand how the performance of the estimation depends on the true underlying relationship between the outcome and the pre-treatment confounders. When one is working with real data, it is almost impossible to know the true relationship between the outcome and the predictors, thus these models are sensible candidates to help us quantify the impact the true underlying outcome model form may have on inferences.

Here, we use one binary and three continuous models to generate the true outcome, which deviate from the models used to estimate the treatment effect in the outcome analysis to varying degrees.

The binary model fits a *logistic regression* as a function of X and T as follows:

Outcome 1.

$$P(Y=1|T,X) = \frac{1}{1+e^{-(\alpha_0+\alpha_1 X_1+\alpha_2 X_2+\alpha_3 X_3+\alpha_4 X_4+\alpha_5 X_8+\alpha_6 X_9+\alpha_7 X_{10}+\gamma T)}} \tag{5}$$

where the continuous value of $P(Y=1|T,X)$ is dichotomized — using random numbers between 0 and 1 (see the *Appendix*). This corresponds to the original outcome model proposed by Setoguchi *et al.* [9]. *Outcome model* 1 is a truly additive model and thus our estimation of the treatment effect using simple main effects logistic regression should perform well so long as we include the correct set of pre-treatment covariates as independent variables in the regression model for Y.

The three continuous outcome models are

Outcome 2.

$$Y = \alpha_0 + \gamma T + e^{\alpha_1 X_1+\alpha_2 X_2+\alpha_3 X_3} + \alpha_4 e^{1.3 X_4} + \alpha_5 X_8 + \alpha_6 X_9 + \alpha_7 X_{10} \tag{6}$$

Outcome 3.

$$Y = \alpha_0+\gamma T+4sin\left(\alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_8 + \alpha_6 X_9 + \alpha_7 X_{10}\right) \tag{7}$$

Outcome 4.

$$Y = \alpha_0+\gamma T+\alpha_1 X_1+\alpha_2 X_2^2+\alpha_3 X_3+\alpha_4 e^{1.3 X_4}+\alpha_5 X_8+\alpha_6 X_9+\alpha_7 X_{10} \tag{8}$$

The true values of the coefficients used are:

$$\begin{aligned} \gamma &= -0.4, \\ (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7) &= (-3.85, 0.3, -0.36, -0.73, -0.2, 0.71, -0.19, 0.26). \end{aligned}$$

Outcomes 2, 3 and 4 correspond to outcomes 4, 5 and 2 of Setodji *et al.* [8], respectively.

In the *Appendix* we briefly present the data generation steps, as well as the steps followed to obtain the values of balance evaluation reported below.

For the continuous outcomes, outcomes 2 and 4 show a greater deviation from additivity compared to outcome 3. In these situations, it is likely to be more difficult to obtain an unbiased estimation of the causal treatment effect when analysing the data under a simple linear outcome model that only controls for the pre-treatment confounders using main effects.

5.3 Sets of Confounders

There is some discussion in the literature about the covariates one should consider as confounders and use to estimate the PS and balancing weights (see *section* 4). In order to test these suggestions, we consider several sets of confounders, each including different baseline covariates. These are:

Confounders-set 1: true_confounders. This is the set of true confounders only (X_1, X_2, X_3, X_4) — covariates related both to the treatment status and the outcome.

Confounders-set 2: treatment_all. This is the set of all covariates that are related to the treatment allocation $(X_1, X_2, X_3, X_4, X_5, X_6, X_7)$.

Confounders-set 3: all_covariates. This set includes all covariates, either related to the treatment or the outcome $(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10})$ — all the available covariates.

Confounders-set 4: outcome_all. This is the set of covariates that are related to the outcome $(X_1, X_2, X_3, X_4, X_8, X_9, X_{10})$.

Confounders-set 5: true_subset. This set includes only two, out of four in total, true confounders (X_1, X_2) — a restricted set of confounders.

Confounders-set 6: treatment_only. This set includes the three covariates related to the treatment allocation only (X_5, X_6, X_7) — no covariate related to the outcome is included.

Confounders-set 7: outcome_only. This set includes the three covariates related only to the outcome only (X_8, X_9, X_{10})— no covariate related to the treatment allocation is included.

Figure 1 depicts the original relationship between the baseline covariates and the treatment/outcome covariates, while the subplots represents the relations each set of confounders considers — only the solid-arrow covariates are considered in each case.

These seven sets of confounders represent all the realistic potential combinations of covariates that could be included in the estimation of PS and balancing weights — covariates related exclusively to the outcome (confounders-set 7: *outcome_only*), related exclusively to the treatment allocation (confounders-set 6: *treatment_only*),

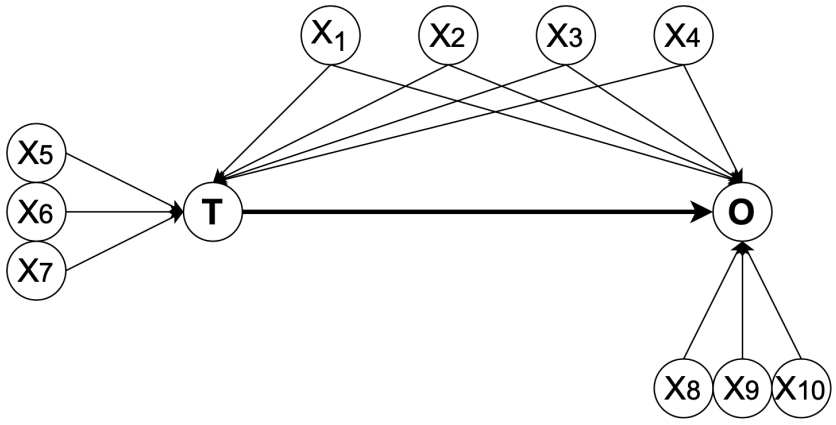

Figure 1 Diagram representing the relation of each covariate with treatment/outcome. The arrows represent which baseline covariates affect the treatment/outcome.

the true confounders (confounders-set 1: *true_confounders*), all the covariates related to the treatment allocation (confounders-set 2: *treatment_all*), all the covariates related to the outcome (confounders-set 4: *outcome_all*), covariates related either to the treatment allocation or the outcome (confounders-set 3: *all_covariates*) and a part of the true confounders (confounders-set 5: *true_subset*).

5.4 Sample Size

It is of key interest for us to understand the sample sizes required to draw unbiased inferences in the space of PS and balancing weights. No rules of thumb have been developed that help guide researchers as to how many units/individuals is sufficient. The rule of 10-to-1 (10 units/individuals per confounder per group) commonly used for regression models may not be appropriate in this situation and we therefore designed our simulations to allow for a more careful assessment of the impact of sample size on performance of PS and balancing weight methods. Specifically, we evaluate the performance of the algorithms on sample sizes equal to $n = 40, 80, 100, 200, 300, 400, 500, 600, 800, 1000, 1500, 2000$, for every set of confounders, for every outcome.

Software Implementation

The simulations were conducted in R version 4.0.3 [59]. The R-packages *stats* [60], *CBPS* [61], *twang* [41], and *entbal* [62], were used to compute PS and balancing weights from LR, CBPS, GBM and EB, respectively.

6 Results

We estimate PS and balancing weights for each dataset (for the 1000 replicates of each simulated scenario), and estimate the ATT using *doubly robust estimators* [44, 45, 46, 47] — this means the PS and balancing weights are used as weights in an augmented regression model (logistic for binary and linear for continues outcomes,

respectively). Outcomes 1 and 3 (see *section* 5.2) are very close to typical logistic and linear regression, respectively, thus it is possible to obtain a good estimation of the causal treatment effect using simple regression models without PS and balancing weights — here we consider a good estimator to be an estimator with low absolute relative bias. In such cases, an unweighted regression may be preferable as the ESS is not reduced, thus providing the maximum possible power. Of the remaining two outcomes (outcomes 2 and 4), outcome 2 shows the greater deviation from linearity. When analysing real data, the true relationship between the baseline covariates and the outcome is not known, and cannot be ascertained. We therefore report results obtained from outcome 2, as it represents a situation where there is extreme mis-specification of the outcome model, and thus a "worst-case" scenario. Results concerning outcome 4 lead to similar conclusions. We do not discuss results of outcomes 1, 3 and 4 in the main body of this article, however, we provide the equivalent of *Figure* 4 in the *Appendix*. Conclusions concerning outcome 4 are very similar to the findings about outcome 2.

We begin the presentation of the results by comparing the relative performance of different balance measures (*section* 6.1) — mean SMD, maximum SMD, mean KS and maximum KS — in predicting the absolute relative bias. We use these results to identify an appropriate threshold for the best balance statistic (*section* 6.2) which provides the most evidence that the causal treatment effect estimate will yield the lowest possible bias. Then, we discuss the performance of different sets of confounders with respect to absolute relative bias (*section* 6.3) for a range of sample sizes, and conclude with the number of units/individuals per covariate per group required (*section* 6.4) to obtain a treatment effect estimation with high accuracy and low bias.

6.1 Balance Statistic

To understand which balance measure is most suitable for providing evidence of balance achieved among the baseline covariates for the two groups, we examine the relationship between our four balance statistics (the mean SMD, maximum SMD, mean KS statistic, and maximum KS statistic) and the absolute relative bias of the estimated causal treatment effect. For the results reported in this section, we used confounders-set 3: *all_covariates* — all the available baseline covariates.

Figure 2 depicts mean SMD (bottom-right), maximum SMD (top-right), mean KS (bottom-left), maximum KS (top-left) on the $x-axis$, and the absolute relative bias on the $y-axis$.

From the figures depicting both maximum and mean SMD, it is apparent that some algorithms which achieve exact balance (exactly zero mean difference) fail to estimate the causal treatment effect with a reasonably small absolute relative bias — reported value over 0.5. Thus, low SMD is not a guarantee of low absolute relative bias. For example, Entropy Balancing controlling for the first moment ($EB\#1$), does not achieve low absolute relative bias, despite being an algorithm designed to achieve exact balance between the treatment groups on the mean value of each confounder. This seems to be the case independently of the sample size (we also tested sample sizes of 20000 and 50000). The reason for this is that $EB\#1$ produces weights that are extremely close to (or exactly) 0, to achieve balance on

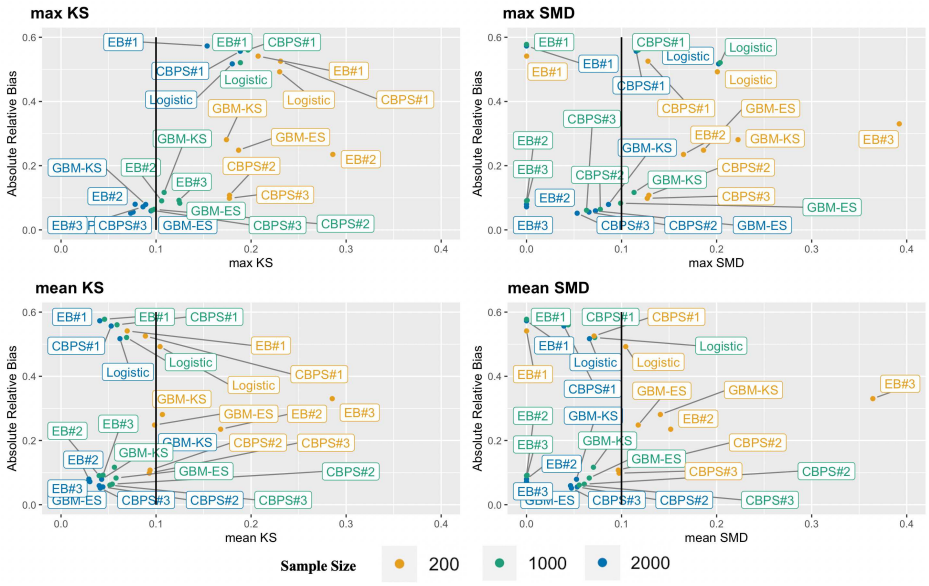

Figure 2 Maximum KS statistic (top-left), maximum SMD (top-right), mean KS statistic (bottom-left), mean SMD (bottom-right) vs absolute relative bias of the estimation of the causal treatment effect. Each color represents a different sample size $(200, 1000, 2000)$. The vertical black solid line represents the 0.1 level of balance, which is widely used as a threshold for acceptable balance. Each point corresponds to a different PS and balancing weights algorithm (as labeled).

the first moment between the treatment groups. By assigning weight of 0 to some individuals, these are also excluded from the estimation of the causal treatment effect, i.e. are considered as non-significant observations. Thus, excluding some individuals from the estimation of the treatment effect could bias the estimate. This is often acceptable, as it indicates regions where there is not much overlap in the baseline, however, in case of EB that occurs with many observations (having a zero weight) even in well overlapped regions. The same pattern is also observed when considering the mean KS statistic.

Conversely, the maximum KS statistic is a very conservative measure, as it will report no balance if only one covariate reports a high value. However, it is the only balance statistic that has a consistent relationship with absolute relative bias — the higher the maximum KS statistic, the higher the absolute relative bias. In particular, algorithms achieving maximum KS < 0.1 also achieve absolute relative bias of < 0.1. This suggests that maximum KS < 0.1 is a sufficient criterion to indicate low absolute relative bias.

6.2 Balance Threshold

Next we will explore the validity of using 0.1 as a threshold for achieving balance for *max KS* across a range of sample sizes and weighting algorithms. In this section we again focus on the results produced based on using confounders-set 3: *all_covariates* — this is the case where all covariates are considered as confounders, and thus all available information is taken into account when we compute the PS and balancing weights. We utilise this set of confounders, since it is the one with the highest number of confounders, thus it would require a larger sample to achieve balance.

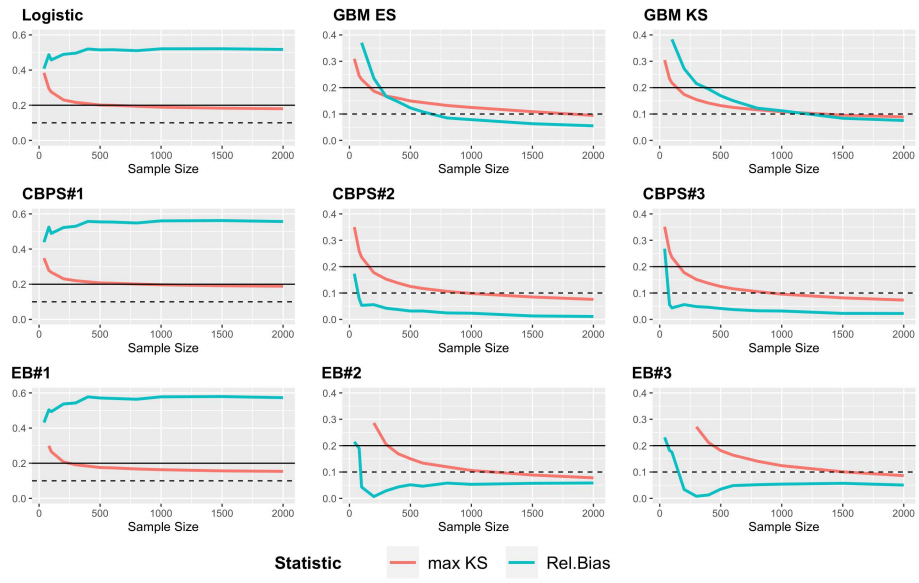

Figure 3 Relationship of maximum KS statistic (red line) and absolute relative bias (blue line) to sample size (40 to 2000). Each sub-figure corresponds to a different algorithm for the estimation of the PS and balancing weights. The horizontal black solid line corresponds to a balance (or bias) threshold of 0.2, while the horizontal black dashed line corresponds to a threshold of 0.1.

Figure 3 depicts the relationship of the maximum KS statistic and absolute relative bias to sample size (40 to 2000).

The horizontal solid black line corresponds to a balance threshold of 0.2 [10], and it is apparent that PS and balancing weights estimated with every algorithm manage to balance the baseline covariates on this level, given a sufficient sample size. However, only PS and balancing weight algorithms which eventually reduce the maximum KS statistic below 0.1 [8, 21, 18] (horizontal black dashed line) control the relative bias (i.e. < 0.1).

LR, $CBPS\#1$ and $EB\#1$, never produce PS and balancing weights with a maximum KS statistic less than 0.1, and the corresponding absolute relative bias is large for all sample sizes — this also true for sample sizes > 2000. These algorithms use parametric modeling to fit a model on first-order covariates. As a consequence, it is difficult to capture the higher-order relationships between the baseline covariates and the treatment allocation, which exist in the true propensity score model (see *section* 5.1). $CBPS\#2$, $CBPS\#3$, $EB\#2$ and $EB\#3$ manage to produce a maximum KS statistic below 0.1 for sample sizes greater than 1000, and low absolute relative bias (< 0.1) at lower sample sizes. This set of algorithms use parametric models with restrictions imposed on all m first moments of the covariates (in our case $m = 2, 3$). This helps these algorithms to fit the (true) underlying relationship between the baseline covariates and the treatment allocation. GBM_{ES} and GBM_{KS} both require a larger sample (about 2000) to achieve a low maximum KS statistic and absolute relative bias (below 0.1), which is expected considering the nature of the GBM algorithm: since it is fitting trees to predict the probability of allocation to the treatment group, and the total number of iterations is typically high, a large sample is required.

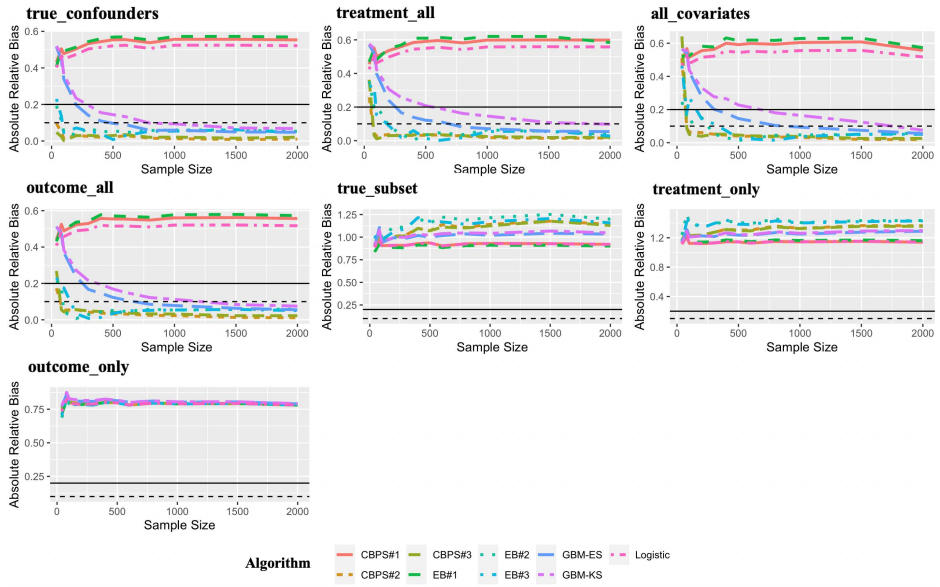

Figure 4 Absolute relative bias ($y - axis$) vs sample size ($x - axis$), per set of confounders (see section 5.3). Each colored line represents a different algorithm used for the estimation of the PS and balancing weights. The horizontal black solid line corresponds to a threshold of 0.2, while the horizontal black dashed line corresponds to a threshold of 0.1.

6.3 Variable Selection

Figure 4 shows the relationship between absolute relative bias of the causal treatment effect estimate and sample size (ranging from $n = 40$ to $n = 2000$), for each set of confounders separately. Each line represents a different algorithm. It is apparent that when the true confounders (X_1, X_2, X_3, X_4) are not all included in the PS and balancing weights model (and the outcome model) — as in confounders-sets *true_subset*, *treatment_only*, and *outcome_only* —, then the estimation of the causal treatment effect always shows large absolute relative bias for all sample sizes and algorithms. High levels of bias in the causal treatment effect are also observed when the algorithm used for the estimation of the causal treatment effect does not achieve adequate balance on the baseline covariates, independently of the covariates included in the PS and balancing weights model. Indeed, LR, $CBPS\#1$, and $EB\#1$ do not balance the maximum KS statistic on the baseline characteristics (the value remains always above 0.1), and show a large absolute relative bias, independently of the covariates treated as confounders.

Figures 5 and 6 show the relationship between sample size (ranging from $n = 40$ to $n = 2000$) and the absolute relative bias and precision (the *mean squared error* (MSE)), respectively, of the causal treatment effect estimate. These calculations were restricted to algorithms which achieve balance as assessed by the maximum KS statistic ($CBPS\#2$, $CBPS\#3$, $EB\#2$, $EB\#3$, GBM_{ES} and GBM_{KS}). Each sub-figure shows one of the sets of confounders which include the *true confounders* — confounders-sets: *true_confounders*, *treatment_all*, *all_covariates*, *outcome_all* (see section 5.3).

Although all sets of confounders that include the true confounders (X_1, X_2, X_3, X_4) show similarly low absolute relative bias, it can be seen that confounders-sets

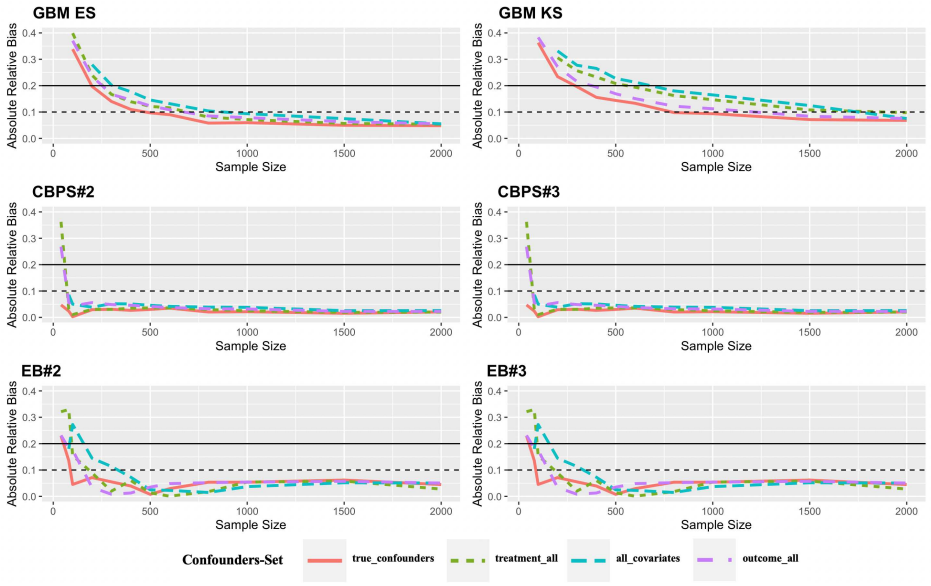

Figure 5 Absolute relative bias ($y - axis$) vs sample size ($x - axis$), per algorithm (see section 5.3). Each colored line represents a different set of confounders (*true_confounders*, *treatment_all*, *all_covariates*, *outcome_all*). The horizontal black solid line corresponds to a threshold of 0.2, while the horizontal black dashed line corresponds to a threshold of 0.1.

true_confounders and *outcome_all* show the lowest values of absolute relative bias, particularly for small sample sizes. Additionally, it is apparent that the sets of confounders that include the covariates related to the outcome (confounders-sets *all_covariates* and *outcome_all*) report lower MSE — thus, such models have better prediction ability. Inclusion of covariates related to the outcome (confounders-set: *outcome_all*) contributes more to lowering bias than the inclusion of covariates related to the treatment allocation (confounders-set: *treatment_all*), when added to the true confounders, which corroborates results in the literature [28, 55]. Adding covariates which are related only to the treatment allocation and not to the outcome to the set of covariates considered for confounding bias (confounders-set: *treatment_only*) inflates the bias and the MSE, and also makes it harder for any algorithm to balance the baseline covariates since the number of covariates increases [30] compared to confounders-set: *true_subset*. Overall, confounders-sets *all_covariates* (all the available covariates) and *outcome_all* (the true confounders and covariates related only to the outcome) achieve the best trade-off between low bias and low MSE, thus — if information regarding the relation of the baseline covariates to treatment/outcome is available, and sample size is sufficient — these are the covariates one should use to control for confounding bias.

6.4 Sample Size

Figures 7 and 8 show the relationship between maximum KS and absolute relative bias, respectively, with the sample size per confounder per treatment group. We focus on confounders-sets *true_confounders*, *treatment_all*, *all_covariates* and *outcome_all* since when the true confounders are not included in the set of covariates one uses to control for confounding bias, the estimate of the causal treatment

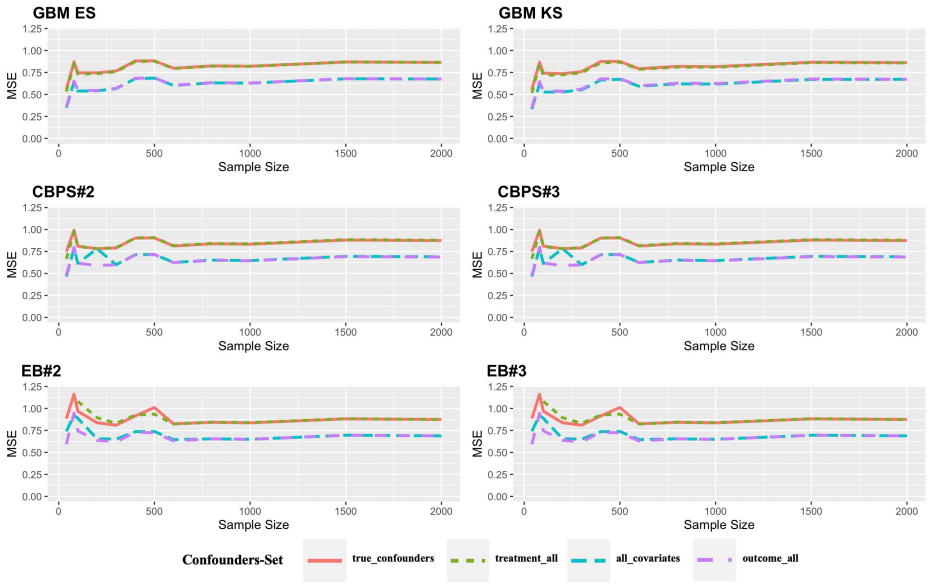

Figure 6 Mean squared error ($y-axis$) vs sample size ($x-axis$), per algorithm (see section 5.3). Each colored line represents a different set of confounders (confounders-sets: *true_confounders*, *treatment_all*, *all_covariates*, *outcome_all*).

effect shows extremely high absolute relative bias — each line represents a different confounders-set (see *section* 5.3). We consider $CBPS\#2$, $CBPS\#3$, $EB\#2$, $EB\#3$, GBM_{ES} and GBM_{KS}, as these algorithms achieve an adequate balance (maximum KS statistic below 0.1) and low absolute relative bias.

The required number of observations per covariate per group to achieve adequate balance (< 0.1), as assessed by the maximum KS statistic (*Figure* 7), is similar for all algorithms. Approximately $60-80$ observations per covariate per group are required to achieve a maximum KS statistic value below the threshold of 0.1, with the upper limit of this range required when the minimum number of covariates (the *true confounders* only — confounders-set: *true_confounders*) is included in the PS and balancing weights model. Slightly fewer observations per covariate per group are required when more covariates are incorrectly assumed to be true confounders — however, this increases the overall sample size.

Figure 8 shows a similar pattern, when PS and balancing weights are estimated by machine learning algorithms (GBM_{ES} and GBM_{KS}). Parametric algorithms ($CBPS\#2$, $CBPS\#3$, $EB\#2$ and $EB\#3$) require slightly lower numbers ($45-60$) of observations per covariate per group to achieve low bias (<0.1) of the causal treatment effect estimate. However, this advantage of the parametric models could be due to the similarity of the true PS model to the parametric models these algorithms fit. When both the true PS model and the true relationship between the baseline covariates and the outcome are mis-specified — an assumption which we are unable to verify when dealing with real data —, the balance statistic is the only measure we have to assess the similarity of the groups on the baseline characteristics, and thus to make inference with high confidence in low bias. In real data problems, it is not feasible to check whether the true relationship between the treatment status and the covariates we use to control for confounding bias is

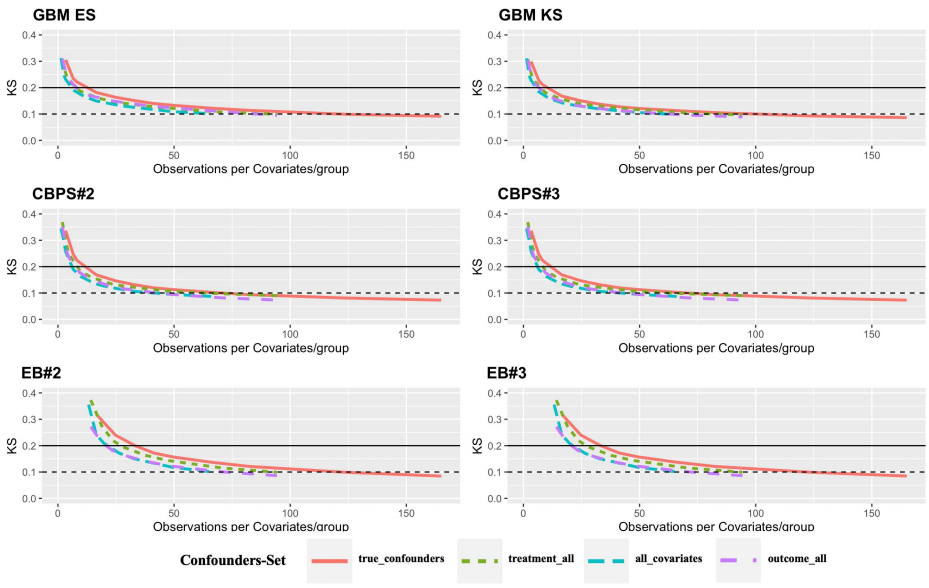

Figure 7 Sample size per confounder per set, vs maximum KS statistic. Each line represents a different set of confounders (see *section* 5.3). The horizontal black solid line corresponds to a threshold of 0.2, while the horizontal black dashed line corresponds to a threshold of 0.1.

close to the relationship assumed in any parametric model, no matter how many higher-order moments or interactions one could include in the model. It is therefore not feasible to know whether the bias of the obtained estimation of the causal treatment effect is low unless sufficient balance has been achieved. In view of this, we recommend the use of several algorithms to compute PS and balancing weights [36] — using both parametric and non-parametric algorithms — and to use for outcome analysis the weights that achieve the best trade-off of lower maximum KS statistic and larger ESS [36].

7 Discussion/Conclusion

Observational studies are becoming more and more popular given the greater use of routine data in applied health research. As such, PS and balancing weights are an area of active investigation. The estimation of the causal treatment effect using PS and balancing weights is a robust method to make inference based on data from observational studies, however, there are certain criteria that should be met to guarantee an unbiased and robust estimation of the desired estimand. Our study highlights several new insights into the role of balance and sample for estimating robust effects.

In order to control for confounding bias, it is important to achieve a sufficient level of balance among the baseline covariates. The threshold of maximum KS statistic less than 0.1 seems to be adequate to guarantee low bias of the estimate of treatment effect, in contrast to 0.2, which is insufficient — assuming the correct set of covariates has been selected to control for confounding bias. Our simulation study used a wide range of models for treatment and outcome in order to provide guidance for achieving balance that will be widely applicable to real-world situations. However, it is of course not feasible to simulate every possible outcome and treatment

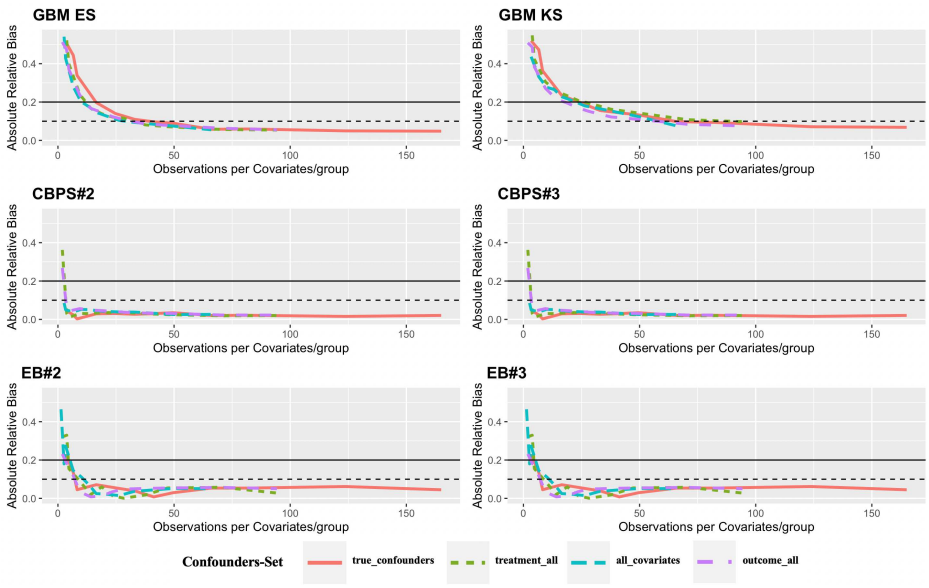

Figure 8 Sample size per confounder per set, vs absolute relative bias. Each line represents a different set of confounders (see *section* 5.3). The horizontal black solid line corresponds to a threshold of 0.2, while the horizontal black dashed line corresponds to a threshold of 0.1.

model, and it may therefore be the case that the thresholds we suggest may be insufficient to guarantee low bias in the treatment effect estimator for all datasets — although it provides important preliminary evidence that one needs at least 0.1 as a threshold on the KS statistic (not the SMD) to ensure adequate balance.

Additionally, special care should be taken when selecting variables to control for confounding bias. If the selected covariates are not predictors of the outcome, the estimation of the outcome could have a huge bias, even if seemingly adequate balance is achieved. The best trade-off between bias and the precision of the estimate (MSE) is achieved when covariates that are related both to the treatment allocation and the outcome (the true confounders), and the covariates related only to the outcome, are included in the PS and balancing weights model to control for confounding bias [28, 29, 31]. Inclusion of the covariates related only to the treatment could improve the balance on the baseline covariates but could increase the bias [28, 29, 31].

Finally, large sample sizes are required to guarantee the robustness of the estimation. It seems that an empirical rule is to have $60 - 80$ individuals per treatment group, per covariate that one wishes to control for — i.e. sample size $840 - 1120$ when controlling for 7 covariates.

The choice of algorithm to compute PS and balancing weights is an important part of the analysis. Since the true relationship between the baseline covariates and the treatment status is usually unknown, we strongly recommend using multiple algorithms for the estimation of the PS and balancing weights [36]. Of the algorithms that reduce the maximum KS statistic below 0.1, the one with with the highest ESS should be chosen, to maximise power. A potential advantage of GBM over the other algorithms is that it makes no parametric model assumptions for the relationship for the treatment allocation mechanism. As such, it is capable of estimating PS and balancing weights for more complex treatment allocation models.

Results on outcome model 3 indicate that when the true outcome model is close to the regression model used for the estimation of the causal treatment effect, smaller samples are acceptable, and a more liberal level of balance is not a restriction to obtain a valid estimate. In this case, even an unweighted estimate would perform well because the regression model is a good fit. However, for real data we are not able to know the true underlying relationship, thus all balance criteria should be strictly met to be confident about the validity of causal treatment effect estimates.

In real case data, the sample size is often a limiting factor on the number of covariates one could include in the PS and balancing weights model to control for confounding bias (see section 6.4). Additionally, since it is not known *a priori* which baseline covariates are related only to the treatment allocation, related only to the outcome, and which are related to both (the *true confounders*), we recommend discussing with subject-matter experts and utilise their prior knowledge of the relationship between the baseline characteristics and the treatment/outcome. Correlation coefficients (preferably *Spearman's correlation* [63, 64], as it makes fewer assumptions on the model relationships between the study covariates) can be calculated for each baseline covariate with the treatment and outcome variable. A combination of data-based evidence about the relationship of baseline covariates and treatment/outcome, and prior expert knowledge should be used to decide which covariates to treat as confounders. Correlation coefficients could also reveal underlying relationships between baseline covariates (e.g. X_2 and X_6 are highly correlated with correlation coefficient 0.9), and thus could be used to reduce the number of confounders included in the PS and balancing weights model, resulting in better balance.

Directed acyclic graphs (DAGs) [65] are another tool that could be used, supplementary to the correlation matrix and scientific advice, to decide which covariates should be included in the PS and balancing weights model. Causal diagrams [66] are graphs that depict the relation between confounders, treatment, outcome, instrumental variables, and more. DAGs [65, 66, 67] are graphical tools that work on causal diagrams, and could propose equivalent minimal sufficient diagrams, in the sense that any potential unnecessary paths could be removed. However, valid and adequate interpretation of these graphical tools is vital, and any decision about the final set of confounders should be communicated with scientific advisors of the study and clinicians, to guarantee that any modification will not affect the estimation (and the interpretation) of the causal treatment effect.

Previous work [29] suggests that goodness-of-fit could be deployed to select which covariates to use to control for confounding bias — this is the typical backward selection of covariates on regression models, starting from a full set of covariates, gradually removing one covariate at a time until only statistically significant covariates remain in the final set. The caveat of such a strategy is that one can only identify which covariates are related to the treatment and which are related to the outcome (separately), and this holds only as long as the regression model used for the significance test is close to the true relationship between the response and the regressors. As a consequence, the methods referred to above for variable selection seem more advisable, since they take into consideration the scientific knowledge and the inherent effect of the confounders on both the treatment allocation and the outcome.

In order to guarantee an estimation of the causal treatment effect with low (acceptable) bias, one needs a large enough sample, a proper set of covariates to control for confounding bias (the true confounders and the covariates related only to the outcome), and a strict threshold to evaluate balance (maximum KS less than 0.1 is recommended). If any of these three conditions is not met, then it is possible to obtain an estimate with large bias and MSE — indicating that this estimation is not robust.

Acknowledgements
We warmly thank Monica Busse, for the useful discussion, support, and comments and edits of the manuscript.

Funding
DOMINO-HD (the project that funds the first author's PhD) is funded though the EU Joint Program for Neurodegenerative Disease Research with UK funding from Alzheimer's Society and Jacques and Gloria Gossweiler Foundation. The Centre for Trials Research, Cardiff University receives infrastructure funding from Health and Care Research Wales. This work was also supported by Medical Research Council (UK) grant MR/L010305/1. Funding was also provided by grant R01DA045049 (PI Griffin) through the National Institute of Drug Abuse.

Abbreviations
PS, Propensity Scores; LR, Logistic regression; CBPS, Covariate Balancing Propensity Score; GBM, Generalized Boosted Model; EB, Entropy Balancing; SMD, Standardized Mean Difference; KS, Kolmogorov-Smirnov; ESS, Effective Sample Size; ATE, Average Treatment Effect on the Entire Population; ATT, Average Treatment Effect on the Treated Population; ATC, Average Treatment Effect on the Control Population; RCT, Randomized Controlled Trial;

Availability of data and materials
The datasets used and/or analysed during the current study, alongside relevant code, available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Authors' contributions
AM, PH, PP and BAG conceptualised the simulation study. AM wrote the code, performed the simulations, and prepared the figures. All authors analysed and interpreted the simulation results. AM and BAG drafted the manuscript. All authors revised and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Author details
[1]School of Medicine, Cardiff University, Cardiff, UK. [2]Centre for Trials Research, Cardiff University, Cardiff, UK. [3]Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK. [4]RAND Corporation, Arlington, VA, USA. [5]MarkoulidakisA@cardiff.ac.uk. [6]holmanspa@cardiff.ac.uk. [7]pallmannp@cardiff.ac.uk. [8]bethg@rand.org.

References
1. Holland, P.W.: Statistics and causal inference. Journal of the American statistical Association **81**(396), 945–960 (1986)
2. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology **66**(5), 688 (1974)
3. Elze, M.C., Gregson, J., Baber, U., Williamson, E., Sartori, S., Mehran, R., Nichols, M., Stone, G.W., Pocock, S.J.: Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. Journal of the American College of Cardiology **69**(3), 345–357 (2017)
4. Harder, V.S., Stuart, E.A., Anthony, J.C.: Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychological methods **15**(3), 234 (2010)
5. Olmos, A., Govindasamy, P.: A practical guide for using propensity score weighting in R. Practical Assessment, Research, and Evaluation **20**(1), 13 (2015)
6. Posner, M.A., Ash, A.S.: Comparing weighting methods in propensity score analysis. Unpublished working paper, Columbia University (2012)
7. Pirracchio, R., Resche-Rigon, M., Chevret, S.: Evaluation of the propensity score methods for estimating marginal odds ratios in case of small sample size. BMC medical research methodology **12**(1), 1–10 (2012)
8. Setodji, C.M., McCaffrey, D.F., Burgette, L.F., Almirall, D., Griffin, B.A.: The right tool for the job: Choosing between covariate balancing and generalized boosted model propensity scores. Epidemiology (Cambridge, Mass.) **28**(6), 802 (2017)

9. Setoguchi, S., Schneeweiss, S., Brookhart, M.A., Glynn, R.J., Cook, E.F.: Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiology and drug safety **17**(6), 546–555 (2008)
10. Abdia, Y., Kulasekera, K., Datta, S., Boakye, M., Kong, M.: Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. Biometrical Journal **59**(5), 967–985 (2017)
11. Choi, B.Y., Wang, C.-P., Michalek, J., Gelfond, J.: Power comparison for propensity score methods. Computational Statistics **34**(2), 743–761 (2019)
12. Gharibzadeh, S., Mansournia, M.A., Rahimiforoushani, A., Alizadeh, A., Amouzegar, A., Mehrabani-Zeinabad, K., Mohammad, K.: Comparing different propensity score estimation methods for estimating the marginal causal effect through standardization to propensity scores. Communications in Statistics-Simulation and Computation **47**(4), 964–976 (2018)
13. Lee, B.K., Lessler, J., Stuart, E.A.: Improving propensity score weighting using machine learning. Statistics in medicine **29**(3), 337–346 (2010)
14. Pirracchio, R., Petersen, M.L., van der Laan, M.: Improving propensity score estimators' robustness to model misspecification using super learner. American journal of epidemiology **181**(2), 108–119 (2015)
15. Wyss, R., Ellis, A.R., Brookhart, M.A., Girman, C.J., Jonsson Funk, M., LoCasale, R., Stürmer, T.: The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. American journal of epidemiology **180**(6), 645–655 (2014)
16. Xie, Y., Zhu, Y., Cotton, C.A., Wu, P.: A model averaging approach for estimating propensity scores by optimizing balance. Statistical methods in medical research **28**(1), 84–101 (2019)
17. Harvey, R.A., Hayden, J.D., Kamble, P.S., Bouchard, J.R., Huang, J.C.: A comparison of entropy balance and probability weighting methods to generalize observational cohorts to a population: a simulation and empirical example. Pharmacoepidemiology and drug safety **26**(4), 368–377 (2017)
18. Li, Y., Li, L.: Propensity score analysis methods with balancing constraints: A monte carlo study. Statistical Methods in Medical Research, 0962280220983512 (2021)
19. Mitani, A.A., Haneuse, S.: Small data challenges of studying rare diseases. JAMA network open **3**(3), 201965–201965 (2020)
20. Day, S., Jonker, A.H., Lau, L.P.L., Hilgers, R.-D., Irony, I., Larsson, K., Roes, K.C., Stallard, N.: Recommendations for the design of small population clinical trials. Orphanet journal of rare diseases **13**(1), 1–9 (2018)
21. Griffin, B.A., McCaffrey, D.F., Almirall, D., Burgette, L.F., Setodji, C.M.: Chasing balance and other recommendations for improving nonparametric propensity score models. Journal of causal inference **5**(2) (2017)
22. Austin, P.C.: Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in medicine **28**(25), 3083–3107 (2009)
23. Franklin, J.M., Rassen, J.A., Ackermann, D., Bartels, D.B., Schneeweiss, S.: Metrics for covariate balance in cohort studies of causal effects. Statistics in medicine **33**(10), 1685–1699 (2014)
24. Zhang, Z., Kim, H.J., Lonjon, G., Zhu, Y., Others: Balance diagnostics after propensity score matching. Annals of translational medicine **7**(1) (2019)
25. Griffin, B.A., Ramchand, R., Almirall, D., Slaughter, M.E., Burgette, L.F., McCaffery, D.F.: Estimating the causal effects of cumulative treatment episodes for adolescents using marginal structural models and inverse probability of treatment weighting. Drug and alcohol dependence **136**, 69–78 (2014)
26. Stuart, E.A., Lee, B.K., Leacy, F.P.: Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. Journal of clinical epidemiology **66**(8), 84–90 (2013)
27. Gail, M.H., Green, S.B.: Critical values for the one-sided two-sample Kolmogorov-Smirnov statistic. Journal of the American Statistical Association **71**(355), 757–760 (1976)
28. Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J., Stürmer, T.: Variable selection for propensity score models. American journal of epidemiology **163**(12), 1149–1156 (2006)
29. Hirano, K., Imbens, G.W.: Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. Health Services and Outcomes research methodology **2**(3-4), 259–278 (2001)
30. Adelson, J.L., McCoach, D.B., Rogers, H.J., Adelson, J.A., Sauer, T.M.: Developing and applying the propensity score to make causal inferences: variable selection and stratification. Frontiers in psychology **8**, 1413 (2017)
31. Perkins, S.M., Tu, W., Underhill, M.G., Zhou, X.-H., Murray, M.D.: The use of propensity scores in pharmacoepidemiologic research. Pharmacoepidemiology and drug safety **9**(2), 93–101 (2000)
32. Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)
33. Rubin, D.B.: Assignment to treatment group on the basis of a covariate. Journal of Educational Statistics **2**(1), 1–26 (1977)
34. Rubin, D.B.: Bayesian inference for causal effects: The role of randomization. The Annals of Statistics **6**(1), 34–58 (1978)
35. Rubin, D.B.: Using multivariate matched sampling and regression adjustment to control bias in observational studies. Journal of the American Statistical Association **74**(366a), 318–328 (1979)
36. Markoulidakis, A., Taiyari, K., Holmans, P., Pallmann, P., Busse, M., Godley, M.D., Griffin, B.A.: A tutorial comparing different covariate balancing methods with an application evaluating the causal effects of substance use treatment programs for adolescents. Health Services and Outcomes Research Methodology, 1–34 (2022)
37. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
38. Cox, D.R., Cox, D.R.: Planning of Experiments vol. 20. Wiley New York, ??? (1958)
39. Robins, J.M., Hernan, M.A., Brumback, B.: Marginal structural models and causal inference in epidemiology. LWW (2000)
40. Imai, K., Ratkovic, M.: Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B

(Statistical Methodology) **76**(1), 243–263 (2014)
41. Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., Griffin, B.A.: Toolkit for Weighting and Analysis of Nonequivalent Groups: A tutorial for the twang package. Santa Monica, CA: RAND Corporation (2017)
42. Hainmueller, J.: Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political Analysis **20**(1), 25–46 (2012)
43. Austin, P.C.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research **46**(3), 399–424 (2011)
44. Bang, H., Robins, J.M.: Doubly robust estimation in missing data and causal inference models. Biometrics **61**(4), 962–973 (2005)
45. Kang, J.D.Y., Schafer, J.L., Others: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical science **22**(4), 523–539 (2007)
46. Chattopadhyay, A., Hase, C.H., Zubizarreta, J.R.: Balancing vs modeling approaches to weighting in practice. Statistics in Medicine **39**(24), 3227–3254 (2020)
47. Zhao, Q., Percival, D.: Entropy balancing is doubly robust. Journal of Causal Inference **5**(1) (2016)
48. Agresti, A.: An Introduction to Categorical Data Analysis. John Wiley & Sons, ??? (2018)
49. Wright, R.E.: Logistic regression. (1995)
50. McCaffrey, D.F., Ridgeway, G., Morral, A.R.: Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological methods **9**(4), 403 (2004)
51. Huang MY, B.L.G.B.M.D. Vegetabile B: Balancing higher moments matters for causal estimation: Further context for the results of setodji et al. (2017). Epidemiology (To be submitted)
52. Burgette, L.F., McCaffrey, D.F., Griffin, B.A.: Propensity score estimation with boosted regression. Propensity Score Analysis: Fundamentals, Developments and Extensions. New York: Guilford Publications, Inc (2015)
53. Ridgeway, G.: The state of boosting. Computing Science and Statistics, 172–181 (1999)
54. Leite, W.: Practical Propensity Score Methods Using R. Sage Publications, ??? (2016)
55. Patrick, A.R., Schneeweiss, S., Brookhart, M.A., Glynn, R.J., Rothman, K.J., Avorn, J., Stürmer, T.: The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. Pharmacoepidemiology and drug safety **20**(6), 551–559 (2011)
56. Pearl, J.: Invited commentary: understanding bias amplification. American journal of epidemiology **174**(11), 1223–1227 (2011)
57. Nguyen, T.-L., Collins, G.S., Spence, J., Daurès, J.-P., Devereaux, P., Landais, P., Le Manach, Y.: Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. BMC medical research methodology **17**(1), 78 (2017)
58. Hernán, M.Á., Brumback, B., Robins, J.M.: Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology, 561–570 (2000)
59. Team, R.C.: R core team: A language and environment for statistical computing r foundation for statistical computing. Austria, Vienna (2018)
60. Team, R.C., Team, M.R.C., Suggests, M., Matrix, S.: R package "stats.". The R Stats Package (2018)
61. Ratkovic, M., Imai, K., Ratkovic, M.M.: R package 'cbps' (2013)
62. Vegetabile, B.G., Griffin, B.A., Coffman, D.L., Cefalu, M., Robbins, M.W., McCaffrey, D.F.: Nonparametric estimation of population average dose-response curves using entropy balancing weights for continuous exposures. Health Services and Outcomes Research Methodology **21**(1), 69–110 (2021)
63. de Winter, J.C., Gosling, S.D., Potter, J.: Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. Psychological methods **21**(3), 273 (2016)
64. Akoglu, H.: User's guide to correlation coefficients. Turkish journal of emergency medicine **18**(3), 91–93 (2018)
65. Textor, J., van der Zander, B., Gilthorpe, M.S., Liśkiewicz, M., Ellison, G.T.: Robust causal inference using directed acyclic graphs: the r package 'dagitty'. International journal of epidemiology **45**(6), 1887–1894 (2016)
66. Textor, J., Liskiewicz, M.: Adjustment criteria in causal diagrams: An algorithmic perspective. arXiv preprint arXiv:1202.3764 (2012)
67. Van der Zander, B., Liśkiewicz, M., Textor, J.: Efficiently finding conditional instruments for causal inference (2015)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixBMCtemplate.pdf