

Comparison of the Ability of Double-Robust Estimators to Correct Bias in Propensity Score Matching Analysis. A Monte Carlo Simulation Study.

Tri-Long Nguyen Pharm.D. M.Sc.,¹ Gary S. Collins Ph.D.,² Jessica Spence M.D.,³
P.J. Devereaux M.D. Ph.D.,⁴ Jean-Pierre Daurès M.D. Ph.D.,⁵ Paul Landais M.D. Ph.D.,⁶
Yannick Le Manach M.D. Ph.D.⁷

¹ Pharm.D., M.Sc., Ph.D. candidate, Laboratory of Biostatistics, Epidemiology, Clinical Research and Health Economics, EA2415, Montpellier University, Montpellier, France; Departments of Anesthesia & Clinical Epidemiology and Biostatistics, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University and the Perioperative Research Group, Population Health Research Institute, Hamilton, Canada.

² Ph.D., Professor, Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, United Kingdom.

³ M.D., M.Sc., Ph.D. candidate, Departments of Anesthesia & Clinical Epidemiology and Biostatistics, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University and the Perioperative Research Group, Population Health Research Institute, Hamilton, Canada.

⁴ M.D., Ph.D., Professor, Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Perioperative Medicine and Surgical Research Unit & Departments of Clinical Epidemiology and Biostatistics and Medicine, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, Canada.

⁵ M.D., Ph.D. Professor, Laboratory of Biostatistics, Epidemiology, Clinical Research and Health Economics, EA2415, Montpellier University, Montpellier, France.

⁶ M.D., Ph.D. Professor and Chairman, Montpellier University, Faculty of Medicine, EA2415, Department of Biostatistics, Clinical Research and Medical Informatics, Nîmes University Hospital, France.

⁷ M.D., Ph.D., Assistant Professor, Departments of Anesthesia & Clinical Epidemiology and Biostatistics, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University and the Perioperative Research Group, Population Health Research Institute, Hamilton, Canada.

Corresponding author:

Yannick LE MANACH, MD, PhD;

Departments of Anesthesia & Clinical Epidemiology and Biostatistics; Michael DeGroote School of Medicine; Faculty of Health Sciences, McMaster University. Hamilton, Ontario, Canada.

Population Health Research Institute; David Bradley Cardiac, Vascular and Stroke Research Institute; Perioperative Medicine and Surgical Research Unit; Hamilton, Ontario, Canada.

Email: yannick.lemanach@phri.ca

Phone: (905) 527-7327

Author contributions:

Study concept and design: Nguyen, Collins, Le Manach.

Analysis and interpretation: Nguyen, Collins, Le Manach.

Drafting of the manuscript: Nguyen, Collins, Le Manach.

Critical revision of the manuscript for important intellectual content: Nguyen, Collins, Spence, Devereaux, Landais, Le Manach.

Word count:

Abstract: 192 words,

Text: 3,236 words.

ABSTRACT

Objective: As covariates are not always adequately balanced after propensity score matching and double-adjustment can be used to remove residual confounding, we compared the performance of several double-robust estimators in different scenarios.

Methods: We conducted a series of Monte Carlo simulations on virtual observational studies. After estimating the propensity scores by logistic regression, we performed 1:1 optimal, nearest-neighbor and caliper matching. We used four estimators on each matched sample: i) a crude estimator without double-adjustment, ii) double-adjustment for the propensity scores, iii) double-adjustment for the unweighted unbalanced covariates, and iv) double-adjustment for the unbalanced covariates, weighted by their strength of association with the outcome.

Results: The crude estimator led to highest bias in all tested scenarios. Double-adjustment for the propensity scores effectively removed confounding only when the propensity score models were correctly specified. Double-adjustment for the unbalanced covariates was more robust to misspecification. Double-adjustment for the weighted unbalanced covariates outperformed the other approaches in every scenario and using any matching algorithm, as measured by the mean squared error.

Conclusion: Double-adjustment can be used to remove residual confounding after propensity score matching. The unbalanced covariates with the strongest confounding effects should be adjusted.

KEYWORDS: Causal inference, propensity score, adjustment, confounding.

INTRODUCTION

In observational studies, the treatment allocation relies on a clinical decision rather than on randomization. Therefore, the treated and control arms are unlikely to be balanced with regard to their baseline covariates, resulting in confounding bias – the sample average difference in outcome across the two arms is not *caused* by the treatment effect, but confounded with the effects of the unbalanced covariates. To restore the covariate balance, numerous approaches are described ¹.

Of these, the propensity score (PS) matching analysis is a popular method for estimating the treatment effect in medical observational studies ²⁻⁴. The PS is defined as the conditional probability of receiving the treatment given a set of confounding covariates ⁵. Under the strong unit treatment value assumption, the positivity assumption and the strong ignorability assumption (*i.e.* assumption of no unmeasured confounders), Rosenbaum and Rubin showed that conditioning on the PS restored the counterfactual framework: the outcome distribution observed in the treated arm is similar to the one that would have been potentially observed in the control arm had it received the treatment, and reciprocally, the outcome distribution observed in the control arm is similar to the one that would have been potentially observed in the treated arm had it received the control ⁵. Formally, this is denoted as $(Y_1, Y_0) \perp Z | PS$, where Y_1 and Y_0 are the potential outcomes (or ‘counterfactuals’) under treatment and control, respectively, and Z is the treatment status ⁵. Matching on the PS pairs control units to treated units with similar value of PS, a framework which allows unbiased estimation of the causal effect, by balancing covariates across groups on average ⁵. Because the treated units here refer to as the ‘reference’ population (*i.e.* controls that are not good matches for the treated units are discarded), the quantity that is estimated is the marginal average treatment effect in the treated population: $ATT = E(Y_1|Z = 1) - E(Y_0|Z = 1)$ on a linear scale ($E(Y_1|Z = 1)/E(Y_0|Z = 1)$ on a relative scale).

1
2
3 Recently, King and Nielsen showed that, in contrast with coarsened exact matching, matching
4
5 on the PS was suboptimal for balancing every covariate perfectly across the groups ⁶. PS
6
7 matching approximates a completely randomized controlled trial which, by balancing
8
9 covariates in average, is likely to be concerned by residual imbalance if the sample size is
10
11 small ^{7,8}. Some authors have proposed using particular regression adjustment, called “double-
12
13 adjustment”, to remove any residual confounding remaining after matching ⁹⁻¹¹. These
14
15 “double-robust” methods take into account the residual confounding effect due to covariates
16
17 that are not properly balanced within the matched sample, and allows a marginal treatment
18
19 effect to be estimated. Austin suggested adjusting using the estimated PS ¹⁰, whereas Abadie
20
21 and Imbens proposed adjusting using the unbalanced covariates ⁹. These two methods have
22
23 not yet been compared.
24
25
26

27
28 Adjusting for the PS is attractive when regression adjustment is limited by a small number of
29
30 events ¹²⁻¹⁴. This approach considers the dimension-reductive PS to be the only variable that
31
32 should be adjusted for. However, we hypothesize that the bias will be inefficiently corrected if
33
34 the PS model is misspecified. We also hypothesize that adjusting for unbalanced covariates is
35
36 more robust to misspecification, but requires a sensible strategy for entering the covariates
37
38 into the regression adjustment ¹¹. Caruana *et al.* cautioned against considering all confounders
39
40 to be equally important ¹⁵. They highlighted the importance of considering the confounders in
41
42 order of both their imbalance and their strength of association with the outcome, by using
43
44 weighted balance metrics ¹⁵.
45
46
47

48
49 We compared the ability of several double-robust approaches to remove residual confounding
50
51 after matching from scenarios with misspecification. We tested adjustment for the PS,
52
53 adjustment for unbalanced covariates chosen with unweighted balance metrics, and
54
55 adjustment for unbalanced covariates chosen with weighted balance metrics.
56
57
58
59
60

METHODS

Data generation

We conducted a series of Monte Carlo simulations on virtual populations that mimicked perioperative settings. Large observational studies have been recently conducted in perioperative medicine^{16,17}; they inspired the design of the following simulations in terms of outcome and covariate distributions. To generate the data, we used a method similar to Setoguchi’s approach¹⁸, with changes adapted to our settings. We generated a set of 14 standard normal variables, which were correlated by different coefficients (*Supplementary Figure S1*). These covariates were dichotomized to obtain realistic distributions that agreed with preoperative biological and physiological statuses reported in observational studies^{16,17}. A 15th covariate was created as a combination of the others, mimicking the cardiac risk index for non-cardiac surgery¹⁹. The 15 covariates were entered into logistic models to define the treatment exposure (*i.e.* the true propensity score model) and the outcome occurrence (*i.e.* the true prognosis model) within three scenarios.

Scenario A – linearity and additivity in both models:

$$\begin{aligned} \text{logit}[p(Z)] &= \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_6 W_6 + \beta_7 W_7 + \beta_{11} W_{11} + \beta_{13} W_{13} \\ \text{logit}[p(Y)] &= \alpha_0 + \alpha_1 W_1 + \alpha_2 W_2 + \alpha_3 W_3 + \alpha_4 W_4 + \alpha_5 W_5 + \alpha_7 W_7 + \alpha_9 W_9 + \alpha_{10} W_{10} + \alpha_{11} W_{11} \\ &\quad + \alpha_{12} W_{12} + \alpha_{13} W_{13} + \alpha_{15} W_{15} + \gamma_Z Z \end{aligned}$$

Scenario B – non-linearity and non-additivity in the true propensity score model:

$$\begin{aligned} \text{logit}[p(Z)] &= \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_6 W_6 + \beta_7 W_7 + \beta_{11} (W_{11}^{1/2} + 0.01 W_{11}^2) + \beta_{13} (W_{13})^{1/2} \\ &\quad + \beta_1 (0.4) W_1 W_2 + \beta_7 (0.5) W_7 W_1 + \beta_2 (0.7) W_7 W_2 + \beta_{11} (0.7) (W_{11})^{1/2} (W_{13}/10) \\ \text{logit}[p(Y)] &= \alpha_0 + \alpha_1 W_1 + \alpha_2 W_2 + \alpha_3 W_3 + \alpha_4 W_4 + \alpha_5 W_5 + \alpha_7 W_7 + \alpha_9 W_9 + \alpha_{10} W_{10} + \alpha_{11} W_{11} \\ &\quad + \alpha_{12} W_{12} + \alpha_{13} W_{13} + \alpha_{15} W_{15} + \gamma_Z Z \end{aligned}$$

Scenario C – non-linearity and non-additivity in both models:

$$\begin{aligned} \text{logit}[p(Z)] = & \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_6 W_6 + \beta_7 W_7 + \beta_{11}((W_{11}-50)^2) + \beta_{13}(W_{13}) + \beta_6(0.8)W_1 W_6 \\ & + \beta_1(0.7)W_7 W_1 + \beta_7(0.9)W_7 W_2 \end{aligned}$$

$$\begin{aligned} \text{logit}[p(Y)] = & \alpha_0 + \alpha_1 W_1 + \alpha_2 W_2 + \alpha_3 W_3 + \alpha_4 W_4 + \alpha_5 W_5 + \alpha_7 W_7 + \alpha_9 W_9 + \alpha_{10} W_{10} \\ & + \alpha_{11}(W_{11}/10)^2 + \alpha_{12}(W_{12})^{1/2} + \alpha_{13} W_{13} + \alpha_{15} W_{15} + \gamma_Z Z + \alpha_{10}(0.2)W_{10} W_7 \\ & + \alpha_4(0.7)W_4 W_2 + \alpha_1(0.6)W_1 W_3 W_7 \end{aligned}$$

As described by these models, we used five true confounders (W_1 , W_2 , W_7 , W_{11} and W_{13}), seven potential confounders (W_3 , W_4 , W_5 , W_9 , W_{10} , W_{12} and W_{14}), one treatment predictor (W_6) and two unrelated covariates (W_8 and W_{14}). The treatment was considered to be assigned ($Z = 1$) and/or the outcome to occur ($Y = 1$) if $p(Z)$ and/or $p(Y)$ were greater than a random uniformly distributed number, $u \sim U(0,1)$. In all of the scenarios, the treatment exposure was set to 40% of the population and the outcome prevalence to 15%. The coefficients used for data generation are reported in *Supplementary Table S1*. The treatment effect was considered to be protective, with a conditional odds ratio of 0.80, which agrees with perioperative randomized controlled trial results²⁰⁻²². In all of the scenarios, the coefficient values, along with $\gamma_Z = \log(0.80)$, were chosen to set the true marginal effect in the treated populations (*i.e.* the average treatment effect in the treated, ATT) to be equal to -0.03 on the absolute risk difference scale. This true marginal effect corresponded to the empirical expectation, computed across an iterative process, of $ATT = E(Y_1|Z = 1) - E(Y_0|Z = 1)$, and was the quantity to be estimated by the PS analysis. (For each unit, Y_1 and Y_0 were generated similarly to Y by setting $p(Y_1) = p(Y|do(Z = 1))$ and $p(Y_0) = p(Y|do(Z = 0))$.) One thousand populations were generated for each scenario, each consisting of 1,000 units. The size was chosen for agreement with similar published studies³ and because small samples are more likely to be affected by residual imbalance.

Propensity score analysis

The individual propensity scores were estimated for each sample using logistic regression that included all 15 covariates. We believe that the models are reflective of current practices^{2,3} rather than optimal²³⁻²⁵. Indeed, including instrumental variables (*i.e.* variables that are only related to the treatment exposure) is recognized to inflate the bias²³⁻²⁵. To attempt to remove confounding, treated units were matched with control units according to their propensity scores within 1:1 ratios without replacement, using either optimal matching, nearest-neighbor matching (NNM) or NNM caliper matching²⁶. In NNM caliper matching, a caliper of width equal to 0.2 standard deviations of the logit-propensity scores was used²⁷. After matching, the covariate balance was checked by calculating the standardized absolute mean difference^{28,29}:

$$SMD = 100 \frac{|\overline{W}_{i1} - \overline{W}_{i0}|}{\sqrt{\frac{s_{i1}^2 + s_{i0}^2}{2}}}$$

\overline{W}_{i1} and \overline{W}_{i0} denote the means (proportions for discrete variables), and s_{i1}^2 and s_{i0}^2 denote the variances in the treated and control groups, respectively.

We then compared the treatment effects estimated for each matched sample by four estimators described in the literature.

We first tested a crude estimator that did not use the double-robust approach on the matched samples. For J matched-pairs:

$$\widehat{ATT}_{crude} = \frac{1}{J} \sum_1^J [(Y_j|_{Z=1}) - (Y_j|_{Z=0})].$$

Second, we tested the double-robust estimator proposed by Austin¹⁰. Two logistic regressions that included Y as the outcome and the PS as the unique explanatory variable were fitted on

each matched arm: $\text{logit}[p(\hat{Y}_1)] = \hat{\gamma}_0 + \hat{\gamma}PS$, to the treated arm; $\text{logit}[p(\hat{Y}_0)] = \hat{\delta}_0 + \hat{\delta}PS$, to the control arm.

If $\hat{p}(Y_1|Z = 1, PS)$ denotes the predicted outcome probability in the treated units, according to the model derived on the treated arm, and $\hat{p}(Y_0|Z = 1, PS)$ denotes the predicted outcome probability in the control units, according to the model derived on the control arm, then:

$$\widehat{ATT}_{DR-PS} = \frac{1}{J} \sum_{j=1}^J [\hat{p}_j(Y_1|Z = 1, PS) - \hat{p}_j(Y_0|Z = 1, PS)].$$

Third, we tested the double-robust estimator described by Abadie and Imbens⁹. Again, two logistic regressions were fitted on each arm. However, the unbalanced covariates were included as explanatory variables rather than the PS.

$$\widehat{ATT}_{DR-PS} = \frac{1}{J} \sum_{j=1}^J [\hat{p}_j(Y_1|Z = 1, W) - \hat{p}_j(Y_0|Z = 1, W)].$$

Contrary to traditional regression, here the separate regressions are used only to estimate the counterfactual outcomes $p(Y_1)$ and $p(Y_0)$, given either the PS or the covariates. We therefore note that marginal effects are still estimable in case of non-collapsible effects (e.g.

$$\frac{[\frac{1}{J} \sum_{j=1}^J \hat{p}_j(Y_1|Z=1)] / [1 - \frac{1}{J} \sum_{j=1}^J \hat{p}_j(Y_1|Z=1)]}{[\frac{1}{J} \sum_{j=1}^J \hat{p}_j(Y_0|Z=1)] / [1 - \frac{1}{J} \sum_{j=1}^J \hat{p}_j(Y_0|Z=1)]}, \text{ for estimating an odds ratio}).$$

Recommendations are that there should be at least 5-10 events per variable when using logistic regression^{12,13}. We used two strategies to add the unbalanced covariates to the logistic regression in this final estimator, giving two variations of the estimator. We first considered the covariates to be equal and entered the covariates with the greatest imbalance, as measured by the SMD.

We then took into account the covariates' association with the outcome and entered the covariates with the greatest weighted SMD, as proposed by Caruana *et al.*¹⁵. For each covariate W_i , if $\hat{\alpha}_i$ denotes the estimated regression coefficient obtained from a full prognosis model fitted on the initial sample and \widehat{SD}_i is the estimated standard deviation in the initial sample, then:

$$weighted\ SMD_i = SMD_i \times \hat{\alpha}_i \times \widehat{SD}_i.$$

Therefore, in each matched sample, four estimators were computed: the crude, the one using double-adjustment for the PS, the one using double-adjustment for unbalanced covariates as measured in SMD and the one using double-adjustment for unbalanced covariates as measured in weighted SMD. The two latters limited the number of covariates to be included, with respect to the common rule-of-thumb of events per variable: in no cases could the number of covariates exceed $\frac{Number\ of\ events}{10}$. Therefore, the separate regression models predicting the counterfactual outcomes could be different across the simulations, depending on the number of events and on which covariates were the most unbalanced. Unlike Caruana *et al.*, we used the weighted SMD as a criterion to enter covariates in double-adjustment, not to summarize the average covariates balance¹⁵.

The four tested estimators' performance was measured with the relative bias and mean squared error. For each estimator and n simulations:

$$Relative\ bias_n\ (\%) = 100 \times \frac{|\widehat{ATT}_n - ATT_{true}|}{ATT_{true}},$$

$$MSE = \frac{1}{N} \sum_{n=1}^N (\widehat{ATT}_n - ATT_{true})^2.$$

In this series of simulations, all PS matching methods were performed using the package 'MatchIt' in R.

RESULTS

In scenarios A, B and C, complete matching methods (*i.e.* nearest neighbor matching and optimal matching) led on average to matched sample sizes of 799, 800 and 805, respectively, which corresponded to the number of treated units. Incomplete matching led to average sample sizes of 688, 682 and 461, respectively.

Figure 1 depicts the SMD for each covariate after matching. As expected, caliper matching minimized the covariate imbalance in all three scenarios and no difference was found between optimal and NNM matching. The residual imbalance increased in scenarios B and C, whose PS models were misspecified as they included non-linearity and non-additivity.

The crude estimates were biased in all three scenarios due to residual confounding (*Figure 2*). As it excludes outliers, caliper matching resulted in less bias than optimal and NNM matching. When the PS models were correctly specified, the double-robust estimator proposed by Austin ¹⁰ consistently removed confounding, regardless of the matching algorithm used. However, this estimator increased the bias in the scenario with high misspecification (scenario C). The double-robust estimator proposed by Abadie and Imbens ⁹ reduced the bias in all three scenarios, even when the functional forms of the covariates entered into the logistic regression were not correctly specified. This estimator performed slightly better when the covariates were chosen for the regression adjustment according to their weighted SMD, as described by Caruana *et al.* ¹⁵. The weighted SMD thus appeared to be a slightly more reliable metric than the unweighted SMD for assessing covariate imbalance.

We calculated the relative bias of each estimate and the percentage of estimated treatment effects that differed from the true ATT by more than 25%, 50%, 75% and 100% (*Figure 3*). When the PS was correctly specified, all of the double-robust approaches substantially

decreased the percentage of biased estimates after optimal and NNM matching. This benefit was weaker after caliper matching. When the PS was highly misspecified (scenario C), the double-adjustment based on the PS again increased the percentage of biased estimates, whereas the double-adjustment based on the unbalanced covariates decreased it. The double-robust estimator supported by the weighted SMD resulted in the least bias in all of the scenarios. It also outperformed the other estimators in all three scenarios and with all three matching algorithms, as measured by the mean squared error (*Figure 4*).

DISCUSSION

We found that double-adjustment for the unbalanced covariates removed residual confounding. Double-adjustment for the PS only reduced residual confounding if the PS models were correctly specified.

Although it attempts to remove confounding bias, matching on the PS can result in residual covariate imbalance ⁶. As in randomized controlled trials, regression adjustment for the unbalanced covariates can address the remaining confounding bias ³⁰⁻³². We compared two double-robust estimators described in the literature. Abadie and Imbens proposed adjusting for unbalanced covariates ⁹, whereas Austin suggested adjusting for the PS alone ¹⁰. In contrary to traditional regression adjustment that estimates conditional effects, these double-adjustment methods allow marginal treatment effects to be estimated. In this study we expressed the treatment effect as an absolute risk difference, which is collapsible, since this scale has been advocated for its convertibility to a number needed to treat – a clinically meaningful measurement. However, double-adjustment methods might also be of particular interest when working on treatment effect scales that are not collapsible (*e.g.* odds ratio).

As we hypothesized, using the PS as the only explanatory variable in the regression adjustment reliably corrected residual bias only when the PS model was correctly specified. If the model was incorrectly specified, this method added bias to the treatment effect estimate. This approach may be useful when regression adjustment is limited by few events. As the PS has reductive dimension properties, it removes the problem of the number of events per variable in logistic regression ^{12,13}.

Adjusting for unbalanced covariates is more robust to misspecification, but requires sufficient events per variable for logistic regression. If this condition limits the inclusion of all covariates, a strategy is needed to select the covariates for the regression. The covariate

1
2 imbalance must therefore be properly evaluated. Metrics such as the SMD assume that all
3
4 covariates have equal importance when measuring imbalance. Caruana *et al.* proposed
5
6 weighting covariate balance measures by the covariates' relative importance, as measured by
7
8 their association with the outcome (*i.e.* their potential confounding effects)¹⁵. This interesting
9
10 approach suggests that the imbalance of strong confounders is likely to lead to greater bias in
11
12 treatment effect estimation than the imbalance of weak confounders. We found that adjusting
13
14 the unbalanced confounders with strong confounding effects maximized the performance of
15
16 the double-robust estimator of Abadie and Imbens in all three scenarios⁹. However, the
17
18 improvement over the use of non-weighted SMD was only slight, and as such, one could
19
20 question the trade-off between the gain of performance and the additional analytic complexity
21
22 of this method.
23
24
25
26

27 Complete matching algorithms, such as optimal or NNM, are likely to leave residual
28
29 imbalance in the sample after matching¹⁰, but keep all of the treated units from the initial
30
31 sample. In contrast, caliper matching diminishes confounding bias, but excludes treated units
32
33 that do not have a PS sufficiently close to a control unit to be matched, which can introduce
34
35 what Rosenbaum and Rubin called "bias due to incomplete matching"³³. Strictly speaking,
36
37 complete matching methods estimate the sample ATT, while caliper matching estimates the
38
39 feasible sample ATT. This distinction, though being clear within an analytic sample, is more
40
41 confused at the population level, since it is data-driven and depends on the overlap in PS
42
43 between treated and control units within the sample. We found that the benefits of using
44
45 double-adjustment were maximized when using complete matching algorithms. Combining
46
47 caliper matching with double-adjustment also corrected confounding bias and minimized the
48
49 mean squared error estimates in all scenarios.
50
51
52
53

54 This study should be interpreted in light of its limitations. As matching analysis is used to
55
56 estimate the ATT rather than the average treatment effect in the population (ATE)³⁴, we only
57
58
59
60

explored the performance of double-adjustment on ATT estimators. Only double-adjustment methods for correcting bias after 1:1 matching have thus far been described. Thus, future studies should explore how double-robust estimators perform after sophisticated algorithms such one-to-many or full matching. Although full matching is not popular in medical observational studies ³, it allows the estimation of either the ATT or ATE ³⁵. As double-adjustment relies on two separate models, the calculation of their variance estimators is a challenge. In this series of simulations, we focused only on point estimators of the treatment effect. Further studies are needed to address the issue of standard error, type I and II error calculations. We therefore only measured estimator performance in terms of the mean squared error. As proposed by Austin and Small, approaches such as bootstrapping should be considered for standard error calculation ³⁶. The covariates could instead be adjusted for in a conditional model. However, conditional effects would then be estimated, which differ from marginal effects when considering a binary outcome ³⁷. Finally, double-adjustment is concerned by the issue of model dependence. Contrary to the property of covariate balance which is measurable, there is no means to verify that the double-adjustment models are correctly specified. Because the estimate depends on the model's specification, researchers might be tempted to report the results, which fit their "favorite hypothesis" ^{1,6}. As model dependence can be removed by covariates balancing, King and Nielsen recommend the use of other matching methods, such as coarsened exact matching ⁶. Indeed, optimal balance avoids the need for double-adjustment, and thus eliminates bias related to model dependence. It is important to note that the primary purpose of PS analysis is to provide a correct covariate balance. In this regard, if this goal is not achieved after matching, one should reiterate the PS specification to provide the best possible balance, before considering double-adjustment.

In conclusion, double-adjustment can be used to remove residual confounding after PS matching. Adjusting for unbalanced covariates is robust against misspecification, whereas

adjusting for the PS can introduce additional bias unless it is done carefully. If not all of the covariates can be included in the regression adjustment, we recommend including the variables with the greatest imbalance, weighted by their association with the outcome. Imbalance metrics such as the weighted SMD should be used to measure imbalance.

ACKNOWLEDGEMENTS

We would like to thank Dr Jennifer de Beyer (Centre for Statistics in Medicine, University of Oxford, UK) for proof reading the manuscript.

REFERENCES

1. Ho DE, Imai K, King G, Stuart EA. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*. 2007;15:199–236.
2. Ali MS, Groenwold RH, Belitser SV, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *J Clin Epidemiol*. Feb 2015;68(2):112-121.
3. Gayat E, Pirracchio R, Resche-Rigon M, Mebazaa A, Mary JY, Porcher R. Propensity scores in intensive care and anaesthesiology literature: a systematic review. *Intensive care medicine*. Dec 2010;36(12):1993-2003.
4. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *The Journal of thoracic and cardiovascular surgery*. Nov 2007;134(5):1128-1135.
5. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
6. King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. <http://gking.harvard.edu/files/gking/files/psnot.pdf?m=14566831912016>.
7. Lachin JM. Properties of simple randomization in clinical trials. *Controlled clinical trials*. Dec 1988;9(4):312-326.
8. Nguyen TL, Collins GS, Lamy A, et al. Simple randomization did not protect against bias in smaller trials. *J Clin Epidemiol*. Feb 28 2017;10.1016/j.jclinepi.2017.02.010.
9. Abadie A, Imbens GW. Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business & Economic Statistics*. 2011;29(1):1-11.

10. Austin PC. Double propensity-score adjustment: A solution to design bias or bias due to incomplete matching. *Statistical methods in medical research*. Jul 17 2014;10.1177/0962280214543508.

11. Nguyen TL, Collins GS, Spence J, et al. Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. *BMC medical research methodology*. Apr 28 2017;17(1):78.

12. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*. 1996;49(12):1373-1379.

13. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol*. Mar 15 2007;165(6):710-718.

14. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol*. Dec 1995;48(12):1503-1510.

15. Caruana E, Chevret S, Resche-Rigon M, Pirracchio R. A new weighted balance measure helped to select the variables to be included in a propensity score model. *J Clin Epidemiol*. Dec 2015;68(12):1415-1422 e1412.

16. Botto F, Alonso-Coello P, Chan MT, et al. Myocardial injury after noncardiac surgery: a large, international, prospective cohort study establishing diagnostic criteria, characteristics, predictors, and 30-day outcomes. *Anesthesiology*. Mar 2014;120(3):564-578.

17. Vascular Events In Noncardiac Surgery Patients Cohort Evaluation Study Investigators, Devereaux PJ, Chan MT, et al. Association between postoperative

- troponin levels and 30-day mortality among patients undergoing noncardiac surgery. *Jama*. Jun 6 2012;307(21):2295-2304.
18. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*. Jun 2008;17(6):546-555.
19. Lee TH, Marcantonio ER, Mangione CM, et al. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. *Circulation*. Sep 7 1999;100(10):1043-1049.
20. Poise Study Group, Devereaux PJ, Yang H, et al. Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. *Lancet*. May 31 2008;371(9627):1839-1847.
21. Pearse RM, Harrison DA, MacDonald N, et al. Effect of a perioperative, cardiac output-guided hemodynamic therapy algorithm on outcomes following major gastrointestinal surgery: a randomized clinical trial and systematic review. *Jama*. Jun 4 2014;311(21):2181-2190.
22. Schouten O, Boersma E, Hoeks SE, et al. Fluvastatin and perioperative events in patients undergoing vascular surgery. *The New England journal of medicine*. Sep 3 2009;361(10):980-989.
23. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in medicine*. Feb 20 2007;26(4):734-753.

24. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *Am J Epidemiol.* Jun 15 2006;163(12):1149-1156.

25. Pearl J. Invited commentary: understanding bias amplification. *Am J Epidemiol.* Dec 1 2011;174(11):1223-1227; discussion pg 1228-1229.

26. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Statistics in medicine.* Mar 15 2014;33(6):1057-1069.

27. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics.* Mar-Apr 2011;10(2):150-161.

28. Austin PC. Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. *Pharmacoepidemiology and drug safety.* Dec 2008;17(12):1218-1225.

29. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine.* Nov 10 2009;28(25):3083-3107.

30. Lavori PW, Louis TA, Bailar JC, 3rd, Polansky M. Designs for experiments--parallel comparisons of treatment. *The New England journal of medicine.* Nov 24 1983;309(21):1291-1299.

31. Senn S. Testing for baseline balance in clinical trials. *Statistics in medicine.* Sep 15 1994;13(17):1715-1726.

32. Senn SJ. Covariate imbalance and random allocation in clinical trials. *Statistics in medicine.* Apr 1989;8(4):467-475.

- 1
2
3 33. Rosenbaum PR, Rubin DB. The bias due to incomplete matching. *Biometrics*.
4
5 1985;41(1):103-116.
6
- 7
8 34. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity:
9 a review. *The Review of Economics and Statistics*. 2004;86(1):4–29.
10
- 11
12 35. Austin PC, Stuart EA. The performance of inverse probability of treatment weighting
13 and full matching on the propensity score in the presence of model misspecification
14 when estimating the effect of treatment on survival outcomes. *Statistical methods in*
15 *medical research*. Apr 30 2015;10.1177/0962280215584401.
16
17
- 18
19 36. Austin PC, Small DS. The use of bootstrapping when using propensity-score matching
20 without replacement: a simulation study. *Statistics in medicine*. Oct 30
21 2014;33(24):4306-4319.
22
23
- 24
25 37. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Systematic
26 differences in treatment effect estimates between propensity score methods and
27 logistic regression. *International journal of epidemiology*. Oct 2008;37(5):1142-1147.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

FIGURE LEGENDS

Figure 1. Covariate balance diagnostics after matching in scenarios (A), (B) and (C). SMD, standardized absolute mean difference.

Figure 2. Estimated average treatment effect in the treated (ATT) in scenarios (A), (B) and (C). In the double-robust estimators (DR), the regression adjustment included either the propensity score (PS) or the covariates as explanatory variables. Covariate imbalance was assessed by the standardized absolute mean difference (SMD) or the weighted SMD.

Figure 3. Percentage of biased estimates in scenarios (A), (B) and (C). In the double-robust estimators (DR), the regression adjustment included either the propensity score (PS) or the covariates as explanatory variables. Covariate imbalance was assessed by the standardized absolute mean difference (SMD) or the weighted SMD.

Figure 4. Mean squared error in scenarios (A), (B) and (C). In the double-robust estimators (DR), the regression adjustment included either the propensity score (PS) or the covariates as explanatory variables. Covariate imbalance was assessed by the standardized absolute mean difference (SMD) or the weighted SMD.

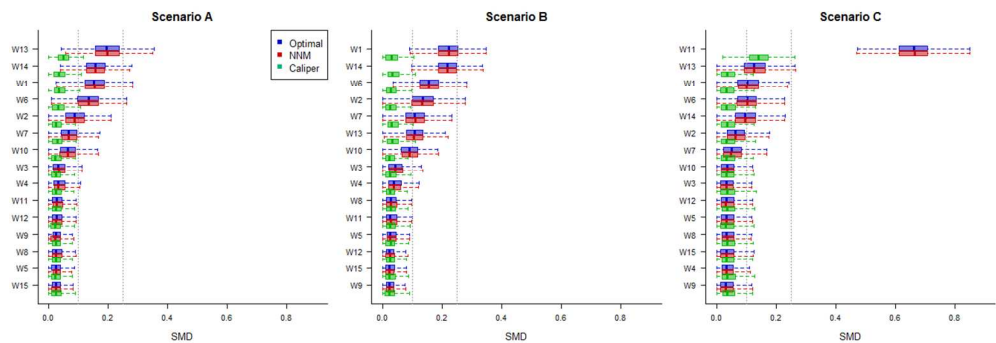


Figure 1. Covariate balance diagnostics after matching in scenarios (A), (B) and (C). SMD, standardized absolute mean difference.

474x162mm (72 x 72 DPI)

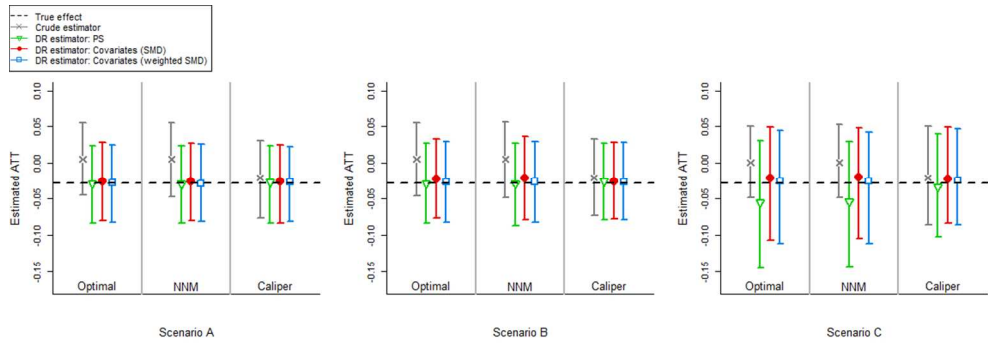


Figure 2. Estimated average treatment effect in the treated (ATT) in scenarios (A), (B) and (C). In the double-robust estimators (DR), the regression adjustment included either the propensity score (PS) or the covariates as explanatory variables. Covariate imbalance was assessed by the standardized absolute mean difference (SMD) or the weighted SMD.

411x142mm (72 x 72 DPI)

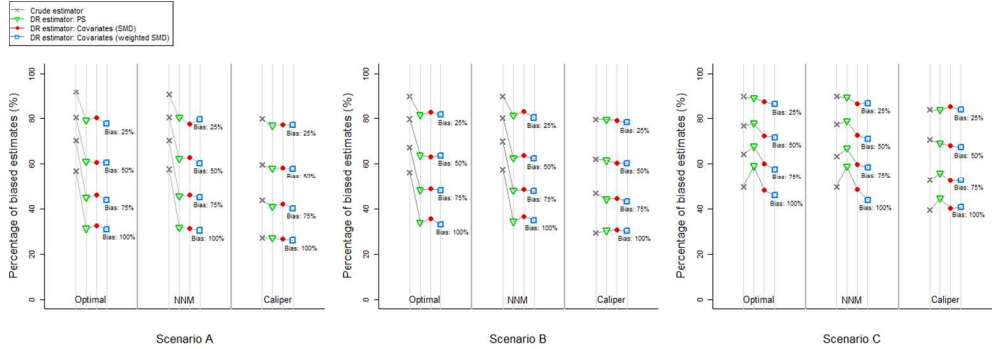


Figure 3. Percentage of biased estimates in scenarios (A), (B) and (C). In the double-robust estimators (DR), the regression adjustment included either the propensity score (PS) or the covariates as explanatory variables. Covariate imbalance was assessed by the standardized absolute mean difference (SMD) or the weighted SMD.

508x179mm (72 x 72 DPI)

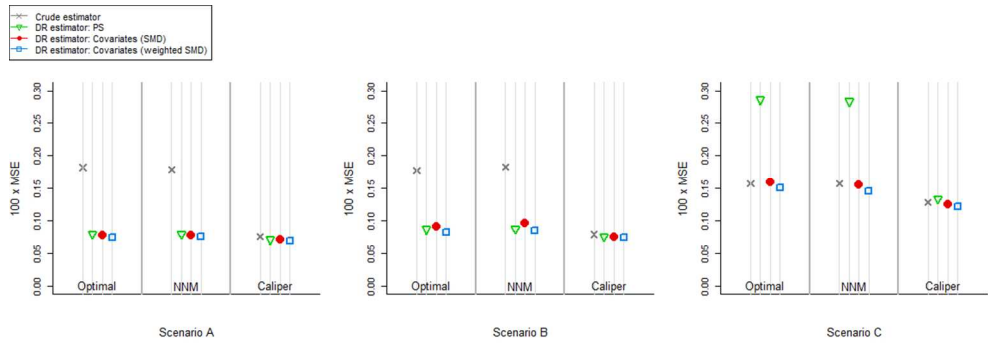


Figure 4. Mean squared error in scenarios (A), (B) and (C). In the double-robust estimators (DR), the regression adjustment included either the propensity score (PS) or the covariates as explanatory variables. Covariate imbalance was assessed by the standardized absolute mean difference (SMD) or the weighted SMD.

411x142mm (72 x 72 DPI)

Comparison of the Ability of Double-Robust Estimators to Correct Bias in Propensity Score Matching Analysis. A Monte Carlo Simulation Study.

Supplementary material

Table S1. Variable definitions and coefficients for data generation. The α and β coefficients define the true outcome model and true propensity score model, respectively.

Scenario A: linearity and additivity in both models

Variable name	Variable type	Related perioperative variable	α coefficients	β coefficients
<i>Intercept</i>			-2.57	0.16
W_1	Binary	Coronary artery disease	0.70	0.70
W_2	Binary	Chronic renal failure	0.64	0.63
W_3	Binary	Diabetes	0.31	0
W_4	Binary	Chronic heart failure	0.80	0
W_5	Binary	Chronic obstructive pulmonary disease	0.54	0
W_6	Binary	History of stroke	0	0.65
W_7	Binary	Hypertension	0.28	0.27
W_8	Binary	Obesity	0	0
W_9	Binary	History of cancer	0.64	0
W_{10}	Binary	Peripheral vascular disease	0.75	0
W_{11}	Continuous	Age (per year)	0.02	0.02
W_{12}	Continuous	Preoperative hemoglobin (per g/dL)	0.05	0
W_{13}	Continuous	Preoperative eGFR (per mL/min)	-0.03	-0.03
W_{14}	Ordinal (5 levels)	Revised cardiac risk index (per level)	0	0
W_{15}	Ordinal (3 levels)	Type of surgical procedure (per level)	0.59	0

Scenario B: non-linearity and non-additivity in the true propensity score model, linearity and additivity in the true outcome model

Variable name	Variable type	Related perioperative variable	α coefficients	β coefficients
<i>Intercept</i>			-2.57	1.48
W_1	Binary	Coronary artery disease	0.70	0.75
W_2	Binary	Chronic renal failure	0.64	0.70
W_3	Binary	Diabetes	0.31	0
W_4	Binary	Chronic heart failure	0.84	0
W_5	Binary	Chronic obstructive pulmonary disease	0.54	0
W_6	Binary	History of stroke	0	0.65
W_7	Binary	Hypertension	0.28	0.60
W_8	Binary	Obesity	0	0
W_9	Binary	History of cancer	0.64	0
W_{10}	Binary	Peripheral vascular disease	0.75	0
W_{11}	Continuous	Age (per year)	0.02	0.01
W_{12}	Continuous	Preoperative hemoglobin (per g/dL)	0.05	0
W_{13}	Continuous	Preoperative eGFR (per mL/min)	-0.03	-0.41
W_{14}	Ordinal (5 levels)	Revised cardiac risk index (per level)	0	0
W_{15}	Ordinal (3 levels)	Type of surgical procedure (per level)	0.59	0

Scenario C: non-linearity and non-additivity in both models

Variable name	Variable type	Related perioperative variable	α coefficients	β coefficients
<i>Intercept</i>			-1.88	-1.22
W_1	Binary	Coronary artery disease	0.77	0.80
W_2	Binary	Chronic renal failure	0.65	0.53
W_3	Binary	Diabetes	0.21	0
W_4	Binary	Chronic heart failure	0.34	0
W_5	Binary	Chronic obstructive pulmonary disease	0.24	0
W_6	Binary	History of stroke	0	0.70
W_7	Binary	Hypertension	0.54	0.30
W_8	Binary	Obesity	0	0
W_9	Binary	History of cancer	0.64	0
W_{10}	Binary	Peripheral vascular disease	0.55	0
W_{11}	Continuous	Age (per year)	0.01	0.01
W_{12}	Continuous	Preoperative hemoglobin (per g/dL)	0.10	0
W_{13}	Continuous	Preoperative eGFR (per mL/min)	-0.02	-0.02
W_{14}	Ordinal (5 levels)	Revised cardiac risk index (per level)	0	0
W_{15}	Ordinal (3 levels)	Type of surgical procedure (per level)	0.19	0

Figure S1. Variable definitions and data generation. We used five true confounders (W_1 , W_2 , W_7 , W_{11} and W_{13}), seven outcome predictors (W_3 , W_4 , W_5 , W_9 , W_{10} , W_{12} and W_{15}), one treatment predictor (W_6) and two unrelated variables (W_8 and W_{14}). Arrows represent causal effects. Arcs represent correlations and are accompanied by the correlation coefficients.

