A comparison of approaches to implementing propensity score methods following multiple imputation

Bas BL Penning de Vries[1], *Rolf HH Groenwold*[1]

(1) Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

CORRESPONDING AUTHOR: *Bas BL Penning de Vries - Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, The Netherlands. - T: +(31)(0)88 75 681 81 - F: +(31)(0)88 75 680 99 - E: B.B.L.Penning_de_Vries@lumc.nl*

DOI: ***10.2427/12630***
Accepted on September 29, 2017

ABSTRACT

Background: In observational research on causal effects, missing data and confounding are very common problems. Multiple imputation and propensity score methods have gained increasing interest as methods to deal with these, but despite their popularity methodologists have mainly focused on how they perform in isolation.
Methods: We studied two approaches to implementing propensity score methods following multiple imputation, both of which have been used in applied research, and compared their performance by way of Monte Carlo simulation for a continuous outcome and partially unobserved covariate, treatment or outcome data. In the first, so-called Within, approach, propensity score analysis is performed within each of *m* imputed datasets, and the resulting *m* effect estimates are averaged. In the Across approach, for each subject the *m* estimated propensity scores are averaged first, after which the propensity score method is implemented based on each subject's average propensity score. Because of its common use, complete case analysis was also implemented. Five propensity score estimators were studied, including regression, matching, and inverse probability weighting.
Results: The Within approach was found to be superior to the Across approach in terms of bias as well as variance in settings with missing covariate data, when missing data were missing at random as well as when they were missing completely at random. In settings with incomplete treatment or outcome values only, the Within and Across approaches yielded similar results. Complete case analysis was generally least efficient and unbiased only in scenarios where missing data were missing completely at random.
Conclusion: We advise researchers not to use the Across approach as the default method, because even when data are missing completely at random, this may yield biased effect estimates. Instead, the Within is the preferred approach when implementing propensity score methods following multiple imputation.

Key words: causal effects, confounding, propensity scores, missing data, multiple imputation

INTRODUCTION

Establishing causal associations between risk factors, treatments, or other exposures, and outcomes is a key aim in many epidemiologic studies. However, in observational studies, the attempt is often hampered by

missing data and confounding.

Simply 'ignoring' missing data, which typically means that a complete case analysis is performed, often is inappropriate, because the conditions under which it is unbiased are very restrictive [1-4]. Even when these conditions are met, for example because the complete cases are truly a random subset of the study sample, discarding incomplete records may render the estimator unnecessarily inefficient [1-3]. An alternative to complete case analysis is multiple imputation (MI), in which missing data are filled in with random draws from their predictive distributions based on the observed data, thereby producing multiple plausible datasets. Inferences are typically made using a simple set of rules known as Rubin's Rules [1]. MI is a popular method for dealing with missing data, because it is flexible, relatively easy to implement with readily available statistical packages, and often provides valid estimates of effect and standard errors in situations where simpler techniques, including complete case analysis, fail [1-3].

To address the problem of confounding, researchers have traditionally used multivariable regression for data analysis. More recently, the use of propensity score methods has gained increasing interest [5]. A subject's propensity score is the conditional probability of being assigned to treatment given their measured covariates [6-7]. Among those subjects with the same propensity score, the distribution of measured covariates is expected to be the same between treated and untreated individuals [6]. Thus, by conditioning on the propensity score treatment status becomes independent of covariates. Several propensity score methods have been described: stratification, matching on the propensity score, inverse probability of treatment weighting (IPW), and covariate adjustment in multivariable regression [6-7]. However, despite increasing popularity, it is largely unclear how these perform in the presence of missing data.

Few have investigated approaches that combine missing data techniques with methods for confounding. Mitra and Reiter studied two approaches that combine multiple imputation with propensity score matching [8]. In both, missing covariate data are imputed m times through multiple imputation. For each of the completed datasets, a propensity score is then estimated for each subject. In the so-called Within approach, propensity score analysis is performed within each of m imputed datasets, and the resulting m effect estimates are averaged. In the Across approach, for each subject the m estimated propensity scores are averaged first, after which the propensity score method is implemented based on each subject's average propensity score. While both approaches were shown to be superior to complete case analysis in terms of bias, it was found that the Across method was less biased than the Within method, especially in the presence of missing covariate data [8]. However, as with any simulation study, these results may not extend beyond the settings that were considered. For example, while it was assumed that the treatment and outcome variables were fully observed, none have compared the approaches in settings with incomplete data for one or both of these variables. Furthermore, although it has been argued that often the outcome should be included in the imputation model [3,9,10] it was excluded from the imputation model in the previous study. With the outcome included, subsequent simulations have found the Within approach to be preferred [11,12]. Nevertheless, in applied research, the Across approach appears to have gained interest since its introduction [13-21].

Our aim was therefore to provide further insight into how propensity scores analysis should be applied in combination with multiple imputation. Specifically, we compared the Within and Across approaches in settings with missing covariate data, missing treatment indicators, and missing outcomes. Because of their common use, complete case analyses were also studied. The remainder of this article is structured as follows. The notation and set-up for the simulations are detailed in Sections 2 and 3. Section 4 details anticipated sources of bias. Results are presented in Section 5 and discussed in Section 6. Finally, we conclude with a summary in Section 7.

METHODS

Notation

Suppose the random vector $\boldsymbol{Z} = (X_1, X_2, \ldots, X_g, T, Y)$ is observed on n subjects. The first g variables of $\boldsymbol{Z}$ represent covariates, whereas T and Y refer to a binary treatment indicator variable and a continuous outcome, respectively. Realisations are printed in lower case letters. We denote an $n \times (g+2)$ matrix by $\boldsymbol{Z}$, whose ith row $\boldsymbol{Z}_i = (X_{i1}, X_{i2}, \ldots, X_{ig}, T_i, Y_i)$ represents the ith $(i = 1,2,\ldots,n)$ subject's record. For each i,j element in $\boldsymbol{Z}$, define a missing indicator variable M_{ij} that takes the value of 1 if it is observed and 0 otherwise. Further, we write $\mathbf{z} = (\mathbf{z}^{obs}, \mathbf{z}^{mis})$ to indicate that z can be partitioned into an observed part $\mathbf{z}^{obs}$ and a missing part $\mathbf{z}^{mis}$. In multiple imputation, values of $\mathbf{z}^{mis}$ are imputed m times by drawing from posterior predictive distributions, resulting in m completed datasets $\mathbf{z}^{(k)}$, $k = 1,2,\ldots,m$ that may be subjected to propensity score analysis. A detailed description of the Across and Within approaches are given in the Supplementary Material.

Simulation methods

We used a series of Monte Carlo simulations to examine the performance of the Across, the Within, and complete case approaches under various missing data mechanisms. The simulations were carried out in several stages. In the first stage, complete data are generated following one of the data generating mechanisms detailed

below. These were chosen for comparability with Mitra and Reiter [8]. Second, missing data are introduced into one of the variables. Third, a number of approaches are applied to estimate the treatment-outcome effect. For each scenario (combination of complete data generating mechanism and missing data mechanism), this process was repeated 1000 times. A full factorial design was used. All simulations were conducted with R Statistical Software version 3.1.1 [22]. For multiple imputation we used the mice package [23]. Continuous and binary variables were imputed using the norm and logreg options, respectively. The number of imputations was set to $m = 5$ for efficiency. For any incomplete variable, all other variables, including the outcome, were included in the imputation model. In all simulations, we used correctly specified propensity score and imputation models.

Data generating mechanisms

We considered $g = 2$ covariates, a binary treatment indicator variable (T_i) and a continuous outcome (Y_i) for $n = 1100$ subjects. Data were simulated by sequentially drawing (X_{i1},X_{i2}), T_i, and Y_i for $i = 1,2,\ldots,n$ from the respective distributions. Let $(X_{i1},X_{i2})\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (10,10)$ and $\boldsymbol{\Sigma}$ has variances equal to 5 and covariance 2.5 (correlation 0.5).

The value of T_i was assigned by drawing from a Bernoulli distribution with parameter (i.e. the probability of treatment assignment) defined as a function of the ith subject's covariate data. In particular, we let

$$\Pr(T_i = 1|X_{i1},X_{i2}) = \text{expit}(-7.8+0.255X_{i1}+0.255X_{i2})$$

where expit(η) is the inverse logit function $\exp(\eta)/(1+\exp(\eta))$. As such, the log odds of treatment increases with 0.255 for every unit increase in either X_1 or X_2. This mechanism assigns approximately 100 subjects to treatment (T=1) and 1000 subjects to the control group (T=0).

We defined the outcome Y_i such that, for all i,

$$Y_i = 2T_i+X_{i1}+0.5X_{i2} + \varepsilon_i, \qquad \varepsilon_i\sim N(0,\sigma^2)$$

where ε_i is independent of (T_i, X_{i1},X_{i2}). The interest lies in estimating treatment effect $\beta_{TY} = 2$, that is, the conditional treatment effect, which—because of homogeneity and the collapsibility of the causal difference in means—equal the marginal treatment effect. Clearly, both covariates serve as a confounder for the association between T and Y. We varied $\sigma^2 = 1,9$ to show that larger residual variances ($\sigma^2=9$) correspond with larger discrepancies between Across and Within estimates in the case of missing covariate data.

Missing data mechanisms

To aid understanding, we initially restricted ourselves to simple missing data mechanisms, namely univariate missing completely at random (MCAR) mechanisms, and finally considered univariate missing at random (MAR) settings. The mechanisms for generating missing data were as follows:

1. *MCAR covariate values*. Any subject's X_2 value was allowed to be missing with probability $Pr(M_{i2} = 1|Z) = p$, $p = 0.2,0.4,0.6,0.8$. For columns $j = 1,3,4$ in Z, we let $Pr(M_{ij} = 1|Z) = 0$.
2. *MCAR treatment indicator values*. We allowed for missing treatment status with $Pr(M_{i3} = 1|Z) = p$, $p = 0.2,0.4,0.6,0.8$, and let the missingness probability of the other variables equal 0.
3. *MCAR outcome values*. We considered the same missing data mechanisms as in *ii* except that we simulated missing outcomes as opposed to missing treatment indicator values.
4. *MAR covariate values*. Two MAR mechanisms were considered. Under mechanism MAR1, missing covariate values were simulated with $\Pr(M_{i2} = 1|Z) = \text{expit}(-8.2 + 0.8X_{i1})(1-T_i)$. Under mechanism MAR2, $\Pr(M_{i2} = 1|Z) = \text{expit}(-13 + 0.8Y_i)$. These mechanisms set approximately 40% of the subjects' X^2 values to missing.

Effect estimators

For all simulated datasets, the Within and Across estimates were obtained as described in the Supplementary Material. Because of their common use, complete case analyses were also performed for comparison. A number of propensity score methods were investigated. The regression estimates of the treatment effect were obtained by linearly regressing the outcome on treatment and the logit of the estimated propensity score. The term matching is used to refer to pair matching performed by selecting for each treated subject a single untreated control without replacement using a greedy nearest neighbour matching algorithm. No restrictions were placed on the maximum acceptable difference between the propensity scores of any two matched subjects. We also performed matching on the logit of the propensity score using a calliper distance of 0.05. A fourth effect estimator was obtained using inverse probability weighting (IPW) where treated subjects are weighted by the inverse of their propensity score and untreated subjects by the inverse of its complement. Finally, we applied iterative inverse probability weighting (IIPW) using a convergence threshold of 10^{-4} and a maximum of 100 iterations per dataset [24]. Calliper matching and iterative IPW were used because matching and IPW are sensitive to practical non-positivity [24,25]. More details on practical non-positivity and IIPW are given in the Supplementary Material.

Variance estimation

An appealing property of the standard multiple imputation approach is that it facilitates estimation of standard errors that reflect both the variability in the data and the uncertainty in the imputations.1 However, for the Across approach, the between-imputation variance component of Rubin's multiple imputation variance estimator cannot fully capture the uncertainty of the imputations. For example, when there are only missing covariates, the between-imputation variance would be zero, because the same set of propensity scores is used for each dataset.

As an alternative to Rubin's rules for variance estimation, analysts can implement a bootstrapping procedure that is akin to the full mechanism bootstrapping approach described by Efron [26]. Here, bootstrapping is implemented as follows:

1. Sample with replacement n rows from the incomplete dataset $\boldsymbol{z}$ to obtain a bootstrapped dataset $\boldsymbol{z}_b$.
2. Impute missing values m times through multiple imputation, producing for $k = 1,2,\ldots,m$ an imputed dataset $\boldsymbol{z}_b^{(k)}$.
3. Apply the analysis procedure (e.g. Within or Across approach) to the m imputed datasets to obtain a single effect estimate for the bootstrapped dataset.
4. Repeat steps 1-3 B times to obtain B bootstrap replicates.

The bootstrap variance and confidence interval for the effect estimate can be obtained from the bootstrap replicates using standard formulae.

For the scenarios with MAR missingness, we estimated variances and confidence intervals using Rubin's rules and the bootstrapping procedure outline above. As discussed, the former can be expected to yield too narrow standard errors and therefore suboptimal coverage. To illustrate this, we applied Rubin's rules for the regression estimators using the modified degrees of freedom formula detailed elsewhere to obtain 95% confidence intervals [27,28]. As for bootstrapping, we calculated bootstrap sample variances and 95% percentile confidence intervals, using the 2.5th and 97.5th percentiles as the lower and upper bounds, based on 1000 bootstrap samples.

Performance measures

The primary performance measure of interest is bias, estimated by the mean deviation of the estimated effect from the true effect of treatment on the outcome (β_{TY}) across all 1000 simulations, but we also provide empirical variances and mean squared errors (MSE).

For the MAR scenarios, coverage probabilities and the mean estimated variances relative to the corresponding empirical variances are also provided. Based on 1000 simulations, the Monte Carlo standard error for the true coverage probability of 0.95 is $\sqrt{(0.95(1-0.95)/1000)} \approx 0.0069$, implying that the estimated coverage probability is expected to lie with 95% probability between 0.936 and 0.964.29 Empirical coverage rates outside this interval provide evidence against the true coverage probabilities being equivalent to the nominal level of 0.95.

Potential sources of bias

To see how the Within approach following multiple imputation might avoid bias due to missing data, it is instructive to consider large samples, so that uncertainty in imputation model parameters may be ignored. Under correct model specification and missingness that is at random (strictly, 'ignorable' in the sense defined by Rubin1), multiple imputation allows for the information lost because of missingness to be restored in such a way that the imputed datasets follow closely the distribution of the full data. Therefore, any analysis procedure may be anticipated to give similar results when applied to the imputed data and to the unobserved full data.

Note that only the Within approach fully adheres to Rubin's original MI algorithm, where averaging across imputations is deferred until the last step. In the context of propensity score matching, this may seem unsatisfactory. Untreated subjects who would be considered unsuitable matches based on their 'true' propensity scores, may be included in the matched set because by random variability their estimated propensity scores, based on the imputed data, better resemble the treated subjects' propensity scores. Untreated subjects whose propensity scores are overestimated tend to be included in the matched set; conversely, untreated subjects whose propensity scores are underestimated tend to be left out. This may then lead to bias by a systematic lack of exchangeability between treatment groups of the matched pseudopopulations. Intuitively, the Across approach may be preferable because of the lesser reliance on random variability. This problem of random variability is due to the nature of propensity score matching-based estimators, and is not expected to introduce bias when for example regression adjustment is used.

The Across approach would appear more robust against the aforementioned source of bias, because matched pseudopopulations are formed after the pooling of propensity scores across imputed datasets. However, for large m the Across approach is comparable to conditional mean imputation, which, as illustrated in the Supplementary Material, may also introduce bias. Given its resemblance, the Across approach is also expected to be biased in the case of missing covariate data.

When treatment or outcome data are missing and covariate data are fully observed, the Within and Across approaches should yield similar results. Consider again a setting with MCAR missingness, now affecting

treatment indicator values only. With large samples, propensity score model estimates would be similar across imputations. Since all covariate values are observed, this implies that propensity score estimates too would exhibit little to no variation across the complete datasets, rendering the Within and Across approaches effectively indiscernible. With complete treatment and covariate data, the propensity score estimates would be exactly the same across imputed datasets, because the outcome does not enter the propensity score model fitting. Therefore, Within and Across estimates would be identical under missing outcome values only.

Daniel et al. showed how causal diagrams can be used to infer that in nearly all scenarios considered here, conditioning on the complete cases (i.e. prior to matching, IPW, or IIPW) does not itself induce bias.4 It follows that when other sources of bias, here practical non-positivity and confounding, are adequately addressed, the treatment effect can be validly estimated. Among the missing data mechanisms considered, it is only MAR2 that biases complete case analysis, namely by inducing a relation between treatment status and the (unmeasured) error on the outcome through collider stratification.

MAR1 is an example of a mechanism that accentuates practical non-positivity. Under this mechanism, untreated subjects with large X_1 values are more likely to be have missing X_2 values than others. When untreated subjects have systematically lower X_2 values even before the introduction of missingness, the consequence of this mechanism is that the propensity score distributions of groups of treated and untreated subjects become more distinct. As a result, estimators that are sensitive to practical non-positivity (e.g. matching and IPW) become more biased when incomplete records are discarded. Note that the matching and IPW methods described in Section 3.3 are estimators of the average effect on the treated (ATT) and the average effect (ATE) on all subjects, respectively [30]. A sufficient condition for these measures of effect to coincide is that of collapsibility and homogeneity of the treatment effect. This joint condition is met in our simulations. The assumptions of ATT and ATE estimators with respect to positivity are, however, not the same. ATE estimators require the covariate distributions of the treated and untreated to have common support, whereas ATT estimators require only the support of the treated to be shared by that of the untreated but not vice versa [31]. This may in part explain possible differences between matching versus IPW estimates.

RESULTS

Bias

In this section, we present graphically the estimated biases for the effect estimators of interest. Results on these and other performance measures are presented in tabular form in the Supplementary Material.

Missing (MCAR) covariate values

Figure 1 depicts the estimated biases for the scenarios with MCAR covariate data. Apart from those based on matching or IPW, the complete case and Within estimators were not identifiably biased. The Across approach, however, showed substantial bias, especially when either the missingness probability, the residual variance σ^2 or both were large. The regression-, matching-, calliper matching-, and IIPW-based estimators were all negatively biased for the Across approach. In contrast, Across IPW estimates were on average overestimated. Complete case matching and IPW estimates were also systematically overestimated, with the extent of bias increasing with the extent of missingness.

Missing (MCAR) treatment indicator values

Figure 2 depicts the estimated biases for the scenarios with MCAR treatment indicator values. The Across and Within estimates were on average highly similar. Apparent from the figure is also the trend that as the percentage of incomplete cases increases, the treatment effect becomes on average progressively more underestimated by both the Across and Within estimators. Conversely, the complete case matching and IPW estimators systematically overestimated the treatment effect, particularly for large missingness probabilities.

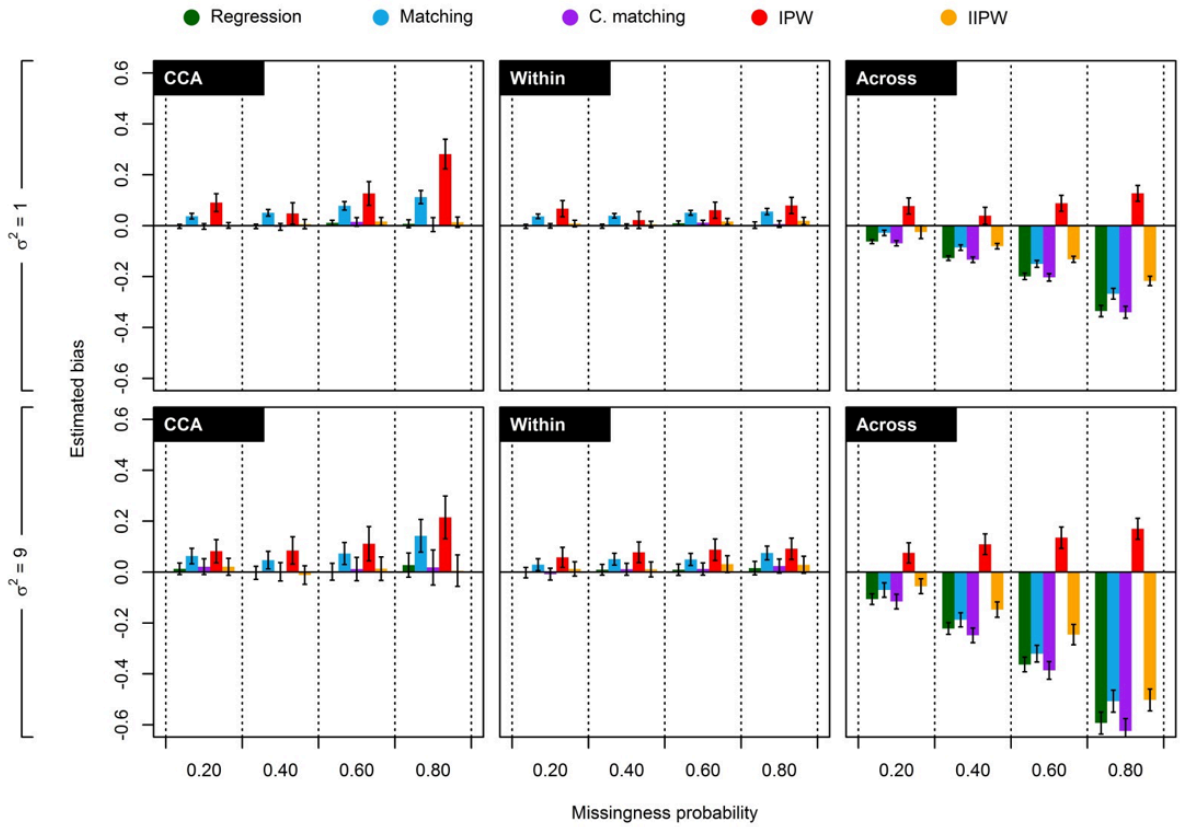
Missing (MCAR) outcome values

Figure 3 depicts the estimated biases for the scenarios with MCAR outcomes. For all propensity score methods, the Across and Within estimators yielded identical results. Again, the complete case matching and IPW estimators showed bias, particularly when the extent of missingness was large. The corresponding Within and Across estimators were less biased. The regression-, calliper matching-, and IIPW-based estimators resulted in minimal bias.

Missing (MAR) covariate values

Figure 4 depicts the estimated biases for the scenarios with MAR covariate data. The complete case matching and IPW estimators generally showed more bias than in the corresponding MCAR covariate settings with a comparable proportion of incomplete records (40%). The regression-, calliper matching-, and IIPW-based estimators showed minimal bias for both the complete

FIGURE 1. Biases of treatment effect estimators for various degrees of missing (MCAR) covariate data and residual variances σ^2.



Abbreviations: C. matching, calliper matching; IPW, inverse probability weighting; IIPW, iterative inverse probability weighting; CCA, complete case analysis. Error bars represent 95% confidence intervals for the simulation estimates of bias.

case analysis and Within approach when the missingness of X_2 depended on X_1 and T (mechanism MAR1). As for the scenarios where the missingness dependend on the outcome Y (MAR2), the Within but not the complete case approach yielded estimates close to the true treatment effect. As before, Across estimates were systematically too low for the regression-, matching-, calliper matching-, and IIPW-based estimators.

Other performance measures

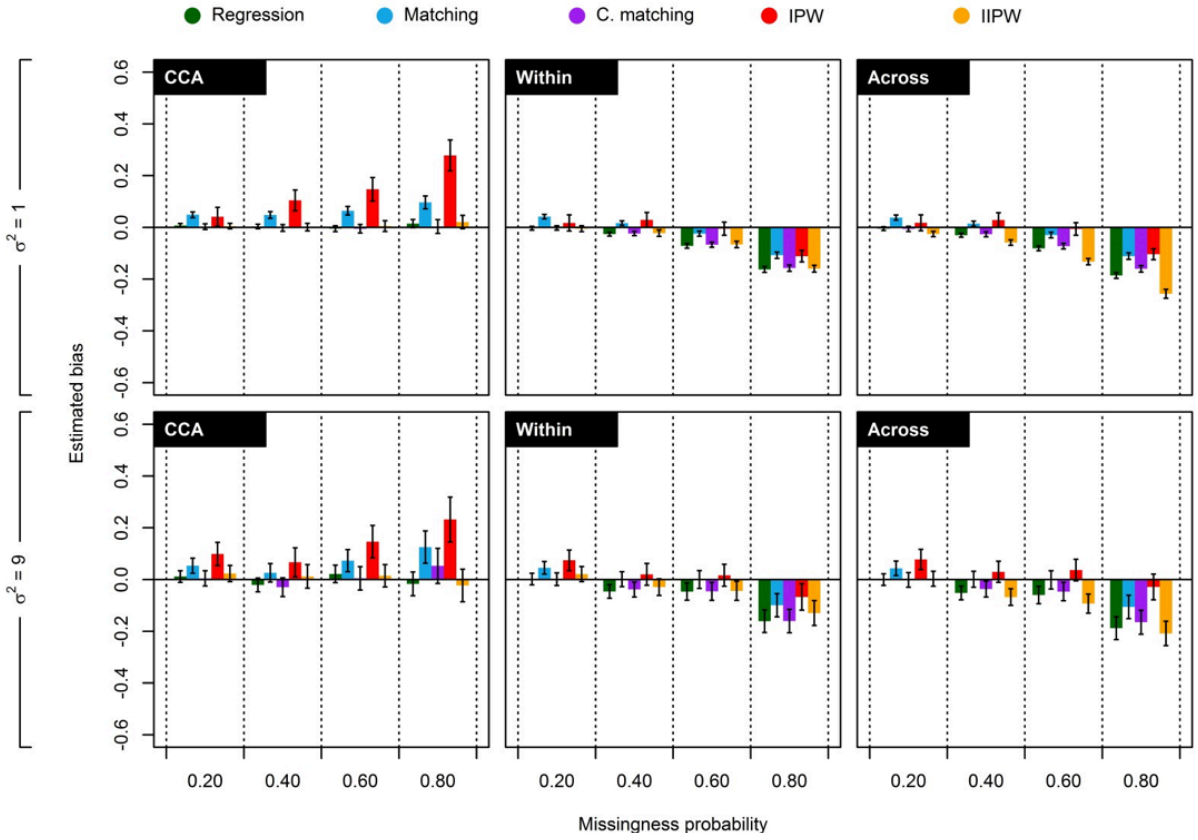
In general, the Within estimators were associated with the smallest empirical variances and MSEs. The simulations also illustrate the implications of using Rubin's rules in estimating the variance. The variances of the Across regression estimators were underestimated and the coverage probability too low (see Supplementary Material). Conversely, when applying the bootstrapping procedure, the estimated variances were on average close to the respective empirical variances. Despite generally adequate coverage probabilities for the Within approach, the variances for calliper matching-, and IIPW-based estimators were on average overestimated.

DISCUSSION

Our primary focus was on examining the relative performance of two approaches to implementing propensity score methods following multiple imputation. Although the Across approach has been applied in practice, our simulations show that, as expected, it fails in settings with missing confounder data, even when the missingness is completely at random and complete case estimators are unbiased.

As stated, untreated subjects with propensity scores that are by random variability underestimated are more likely to be selected as matches than subjects whose propensity score is overestimated. This problem of random variability is inherent to propensity score methods, and is not expected to introduce bias when for example regression adjustment is used. However, its impact was negligible in our simulations, because the calliper matching estimates were highly similar to the regression estimates. The second explanation for the discrepancy in bias between the approaches rests on the resemblance of the Across approach to conditional mean imputation in the context of missing covariate data. This explanation is consistent with our observations that the Across approach

FIGURE 2. Biases of treatment effect estimators for various degrees of missing (MCAR) treatment indicator values and residual variances σ^2.



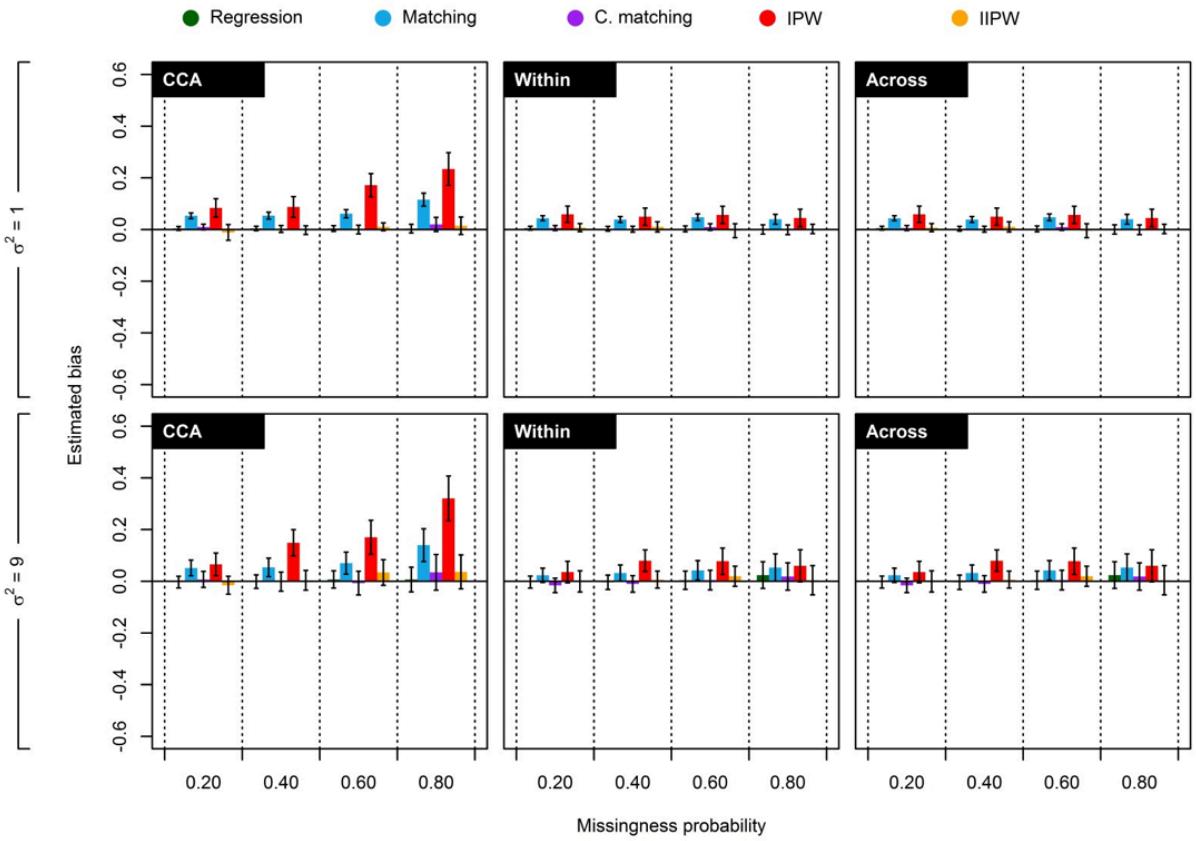
Abbreviations: C. matching, calliper matching; IPW, inverse probability weighting; IIPW, iterative inverse probability weighting; CCA, complete case analysis. Error bars represent 95% confidence intervals for the simulation estimates of bias.

showed more bias for larger missingness probabilities and larger residual variances.

Conversely, with complete confounder data, the Across approach is not comparable to conditional mean imputation. Instead, the bias observed in settings with missing treatment indicator values probably is largely attributable to a phenomenon, known as separation or 'perfect prediction', that is associated with regression models for categorical responses. Separation occurs if the responses, here treatment status, can be perfectly separated by a single predictor or a linear combination of predictors. The problem lies with the Normal approximation to the posterior distribution of the parameters of the logistic regression model that is used by the software to predict missing treatment indicator values. When in the presence of separation, logistic regression is applied to the complete cases, modelling the probability of being assigned to treatment as $Pr(T_i = 1 | X_{i1}, X_{i2}, Y_i) = expit(\alpha 0 + \alpha_1 X_i 1 + \alpha_2 X_i 2 + \alpha_3 Y_i)$, then we can find an infinite sequence of parameter specifications with monotonically increasing likelihood converging to unity, such that for at least one parameter α the estimate a tends to infinity [32,33]. Hence, the maximum likelihood estimate does not exist. Nevertheless, given the near-flat nature of the likelihood, typically very large values for the maximum likelihood estimate a and its variance are returned by standard software. If the Normal approximation to the posterior distribution of the parameters is applied, then it is not unlikely that values are drawn such that in the imputation step subjects with incomplete data are assigned to the treatment group whilst the observed data clearly suggests that these subjects should be assigned to the control group [32]. In other words, the Normal approximation to the posterior distribution is poor. One way to prevent these implausible imputations is to add to the dataset a few observations such that separation is no longer present and with such small weights that the impact on the imputation model is limited [34]. mice implements such a data augmentation method to deal with this phenomenon [3,34]. but we suspect that in our simulations the impact of the weights was large enough to produce bias.

Although unbiasedness is arguably more crucial than valid variance estimation, sufficiently large variances, even if they can be estimated validly, may render unbiased estimators of little practical use [29]. In our simulations, the Within approach was superior or comparable to the

FIGURE 3. Biases of treatment effect estimators for various degrees of missing (MCAR) outcomes and residual variances σ^2.



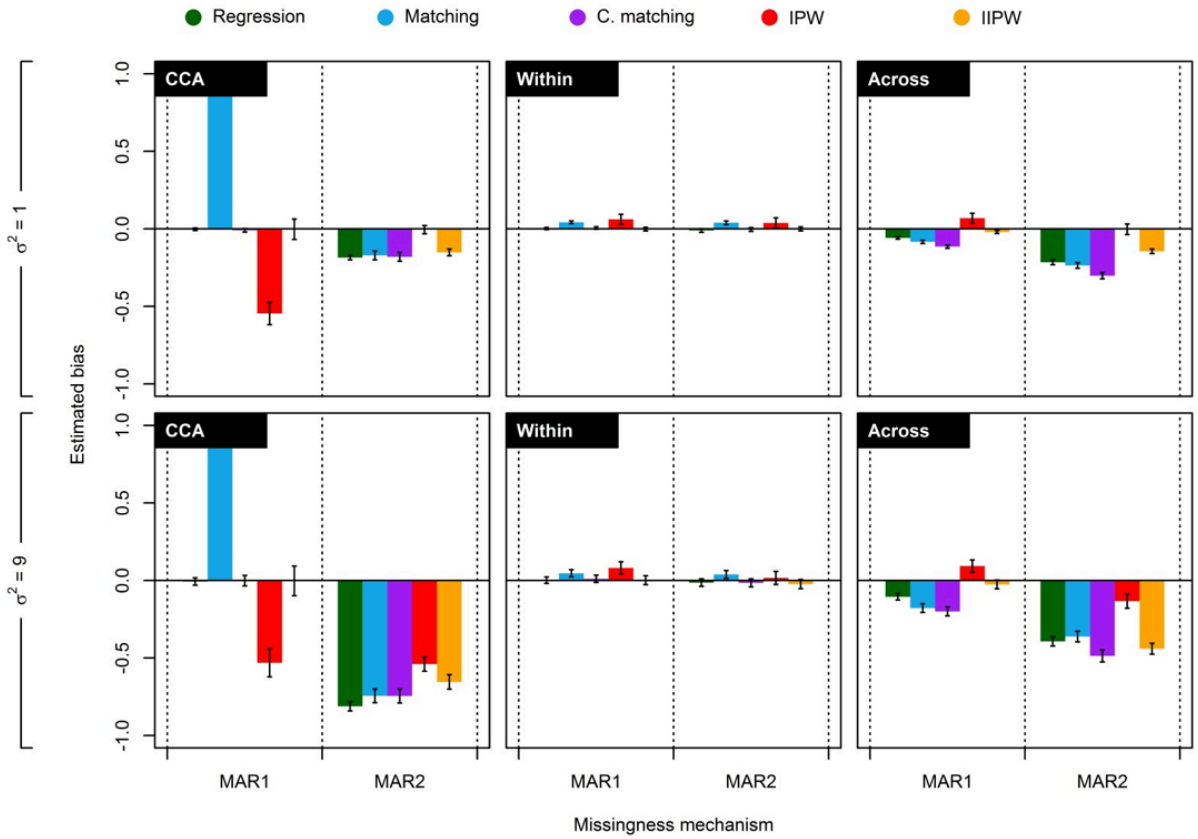
Abbreviations: C. matching, calliper matching; IPW, inverse probability weighting; IIPW, iterative inverse probability weighting; CCA, complete case analysis. Error bars represent 95% confidence intervals for the simulation estimates of bias.

Across in terms of either criterion. Another drawback of the latter approach is the difficulty in making inferences as to the precision of effect estimates. Bootstrapping may provide correct standard errors, but we acknowledge that this approach is computationally intensive. Further, because coverage is affected by the bias in both variance and effect estimation, it is likely to be poor in general for Across estimators. Bootstrapping for the (calliper) matching estimators here yielded slightly overestimated variances and conservative empirical coverage rates. A similar phenomenon was observed by Austin and Small [35]. The bootstrapping procedure defined in Section 3.4 resembles the 'complex bootstrap' of Austin and Small. The rather large discrepancies between the mean estimated variances and the empirical variances for the IIPW estimators are possibly attributable to the unstable nature of inverse probability weighting. Further investigation and development of bootstrapping approaches to variance estimation for the (calliper) matching and (iterative) IPW estimators represent an interesting direction for future research.

Our findings contrast with those of Mitra and Reiter [8]. A crucial difference between the simulations is the inclusion of the outcome in the model used to impute missing covariate values. Failing to include the outcome leads to imputed datasets that do not reflect the association between covariate and outcome that would have been observed had there been no missing values. The consequence of this is that if one adjusts for the imputed covariate values to estimate the treatment effect, the variation in outcomes between the treatment groups that is due to the partially unobserved covariate would in part be attributed to the differences in treatment status.

As with any simulation study, an important limitation of this study is the potentially limited generalisability. The scenarios considered here represent only a small and simplified subset of those likely to be encountered in applied research. Some of the missingness probabilities that were studied are probably unrealistic, and only a single sample size was considered. Practical non-positivity and separation are perhaps less relevant in settings with larger samples and fewer incomplete cases. Furthermore, we considered only two covariates and assumed that the propensity score and imputation models could be correctly specified. Applied researchers do not have the luxury of knowing the data generating and missing data mechanisms and often need to assess and account for

FIGURE 4. Biases of treatment effect estimators under various (MAR) missingness mechanisms and residual variances σ^2.



Abbreviations: C. matching, calliper matching; IPW, inverse probability weighting; IIPW, iterative inverse probability weighting; CCA, complete case analysis. Under mechanism MAR1, the missingness of X2 depends on X1 and T only. Under MAR2, the missingness depends on Y only. Both MAR1 and MAR2 result in ~40% incomplete records. Error bars represent 95% confidence intervals for the simulation estimates of bias.

multiple sources of bias. However, rather than scrutinising methods for these issues in isolation only, it may be interesting to additionally study how they perform in combination. Conducting simulations for specific scenarios of interest may be particularly desirable given the limited generalisability of our results. If these are not possible, we advise researchers not to use the Across approach as the default method, because it appears to offer no advantage over the Within method.

CONCLUSION

In medical research, confounding and missing data are common problems that often occur simultaneously. Complete case analysis, although valid under various circumstances, is discouraged as the default procedure, because it leads to a loss of valuable information and it is typically unknown whether the conditions under which complete case analyses are valid are satisfied. When multiple imputation is to be followed by the implementation of a propensity score method, researchers could apply the Across and Within approaches. The present study highlights a number of aspects of the Across approach that render it suboptimal. Our simulations indicate that the Within approach is superior to the Across approach in terms of both bias and variance in settings with missing confounder data. For incomplete treatment and/or outcome data, the approaches yield similar estimates. We advise researchers not to use the Across approach as the default method, because even in MCAR settings, this may yield biased effect estimates. Finally, when matching or IPW are the propensity methods of choice, we recommend practical non-positivity to be adequately addressed, e.g. by using a narrow calliper or an iterative reweighting algorithm. One should be aware, however, that trimming away or down-weighting observations may direct the focus of inference to a narrower population that may not reflect that of primary interest.

References

1. Rubin DB. Multiple imputation for nonresponse in surveys. vol. 81. John Wiley &Sons; 2004.
2. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychological methods. 2002;7(2):147.

3. Van Buuren S. Flexible imputation of missing data. Boca Raton: CRC Press; 2012.
4. Daniel RM, Kenward MG, Cousens SN, De Stavola BL. Using causal diagrams to guide analysis in missing data problems. Statistical methods in medical research. 2012;21(3):243-56.
5. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. Journal of clinical epidemiology. 2006;59(5):437-e1.
6. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.
7. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research. 2011;46(3):399-424.
8. Mitra R, Reiter JP. A comparison of two methods of estimating propensity scores after multiple imputation. Statistical methods in medical research. 2016;25(1):188-204.
9. Moons KGM, Donders ART, Stijnen T, Harrell F. Using the outcome for imputation of missing predictor values was preferred. Journal of clinical epidemiology. 2006;59(10):1092-101.
10. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
11. Penning de Vries BBL, Groenwold RHH. Comments on propensity score matching following multiple imputation. Statistical methods in medical research. 2016;25(6):3066-8.
12. Leyrat C, Seaman SR, White IR, et al. Propensity core analysis with partially observed covariates: How should multiple imputation be used? Statistical Methods in Medical Research. 2017;0(0):1-17.
13. Neuderth S, Schwarz B, Gerlich C, Schuler M, Markus M, Bethge M. Work-related medical rehabilitation in patients with musculoskeletal disorders: the protocol of a propensity score matched effectiveness study (EVA-WMR, DRKS00009780). BMC Public Health. 2016;16(1):804.
14. Olszewski AJ, Shrestha R, Castillo JJ. Treatment selection and outcomes in early-stage classical Hodgkin lymphoma: Analysis of the National Cancer Data Base. Journal of Clinical Oncology. 2015;33(6):625-33.
15. Gregory EF, Gross SM, Nguyen TQ, Butz AM, Johnson SB. WIC Participation and Breastfeeding at 3 Months Postpartum. Maternal and child health journal. 2016;20(8):1-10.
16. Chiu P, Schafer JM, Oyer PE, et al. Inference of durable mechanical circulatory support and allosensitization on mortality after heart transplantation. The Journal of Heart and Lung Transplantation. 2016;35(6).
17. Brown JB, Leeper CM, Sperry JL, et al. Helicopters and injured kids: Improved survival with scene air medical transport in the pediatric trauma population. Journal of trauma and acute care surgery. 2016;80(5):702-10.
18. Sulkowski JP, Cooper JN, Congeni A, et al. Single-stage versus multi-stage pull-through for Hirschsprung's disease: Practice trends and outcomes in infants. Journal of pediatric surgery. 2014;49(11):1619-25.
19. Sulkowski JP, Cooper JN, Duggan EM, et al. Does timing of neonatal inguinal hernia repair affect outcomes? Journal of pediatric surgery. 2015;50(1):171-6
20. Kutney-Lee A, Melendez-Torres G, McHugh MD, Wall BM. Distinct enough? A national examination of Catholic hospital affiliation and patient perceptions of care. Health care management review. 2014;39(2):134.
21. Ekström N, Svensson AM, Miftaraj M, et al. Cardiovascular safety of glucose-lowering agents as add-on medication to metformin treatment in type 2 diabetes: report from the Swedish National Diabetes Register. Diabetes, Obesity and Metabolism. 2016;18(10):990-8.
22. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2016. Available from: https://www.R-project.org/.
23. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45(3):1-67. Available from: http://www.jstatsoft.org/v45/i03/.
24. Van der Wal WM. Causal modeling in epidemiological practice [PhD thesis]. University of Amsterdam; 2011. Available from: http://hdl.handle.net/11245/1.366252.
25. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. American journal of epidemiology. 2008;168(6):656-664.
26. Efron B, Tibshirani RJ. An introduction to the bootstrap. CRC press; 1994.
27. Barnard J, Rubin DB. Miscellanea. Small-sample degrees of freedom with multiple imputation. Biometrika. 1999;86(4):948-955.
28. Hughes R, Sterne J, Tilling K. Comparison of imputation variance estimators. Statistical methods in medical research. 2014;25(6):2541-2557.
29. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in medicine. 2006;25(24):4279-4292.
30. Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from naive enthusiasm to intuitive understanding. Statistical methods in medical research. 2012;21(3):273-293.
31. Lechner M. A note on the common support problem in applied evaluation studies. Annales d'Économie et de Statistique. 2008;91/92:217-235.
32. Albert A, Anderson J. On the existence of maximum likelihood estimates in logistic regression models. Biometrika. 1984;71(1):1-10.
33. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Statistics in medicine. 2002;21(16):2409-2419.
34. White IR, Daniel R, Royston P. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. Computational statistics & data analysis. 2010;54(10):2267-2275.
35. Austin PC, Small DS. The use of bootstrapping when using

propensity-score matching without replacement: a simulation study. Statistics in medicine. 2014;33(24):4306-4319.

SUPPLEMENTARY MATERIAL

This Supplementary Material has three parts. The first contains a discussion of the Within and Across approaches; the second part reviews the positivity assumption and provides sample R code for the iterative inverse probability weighting estimators for the complete case, Across and Within approaches; and the third part provides the results of the simulation studies on all performance measures.

A Within and Across approaches

Two approaches to implementing propensity score methods following multiple imputation to address missing data are described: the Within and Across approaches [1]. The first step in the analysis procedure is to estimate within each of m imputed datasets a vector of propensity scores $\boldsymbol{e}^{(k)}=(e_1^{(k)},e_2^{(k)},\ldots,e_n^{(k)})$ typically using logistic regression. In the Within approach, any propensity score method is implemented within each completed dataset using $e^{(k)}$, yielding an estimate of the relation between treatment and outcome, $\beta_w^{(k)}$. The Within estimate is defined as the average of $\beta_w^{(1)},\beta_w^{(2)},\beta_w^{(m)}$. In the Across approach, the propensity score method is implemented within each completed dataset, now using $\boldsymbol{e}^A=(e_1^A,e_s^A,\ldots,e_n^A)$, where $e_i^A=\sum_{(k=1)}^{m} e_i^{(k)}/m$. As in the Within approach, the resulting estimates $\beta_A^{(1)},\beta_A^{(2)},\beta_A^{(m)}$ are averaged, yielding the single estimate β_A. This procedure deviates slightly from the Across approach described by Mitra and Reiter [1]. The modified Across approach described here is equivalent to the original procedure when T and Y are fully observed, but can additionally accommodate missings on T and/or Y. Henceforth, we will refer to this modified procedure simply as the Across approach.

In what follows, we first give a heuristic explanation for the bias due to the Across approach in a simple setting with missing covariate data and then offer more technical arguments. Consider for simplicity the case with only a single continuous covariate X, and suppose that the treatment-outcome effect can be parameterised through the linear regression model

$$E(Y|T,X)=\beta_0+\beta_1 T+\beta_2 X$$

where β_1 is the parameter of interest. Further, assume that the relationship between the probability of being assigned to treatment $(T=1)$ given X can be modelled by a logistic function

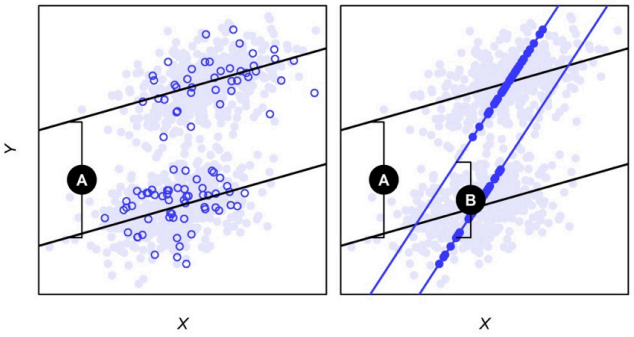
$$\Pr(T=1|X)=\frac{\exp(\alpha_0+\alpha_1 X)}{1+\exp(\alpha_0+\alpha_1 X)}$$

In this case, we may rewrite the treatment-outcome model in terms of a linear transformation $logit\ e(X)=\alpha_0+\alpha_1 X$ of the covariate values, the logit of the propensity score $e(X)$,

$$E(Y|T,X)=(\beta_0-\beta_1\alpha_1^{-1}\alpha_0)+\beta_1 T+\beta_1\alpha_1^{-1}\text{logit}\{e(X)\}$$

The ordinary least squares estimator of the true treatment effect is unbiased if the regressors of the linear model are T and X, or T and β_1. A similar observation can be made in the case of multiple covariates [2]. However, if we impute missing X values using conditional mean imputation, and regress Y on T and the (linearly transformed) imputed covariates to estimate the treatment effect, then, as illustrated in Figure 1, the estimator will be biased—provided that the conditional variance of Y given T and X is greater than zero and treatment assignment depends on X.

FIGURE 1. A single random sample (left) with missing completely at random (MCAR) covariate data imputed using conditional mean imputation (right). A valid treatment effect (A) is obtained by the regression of Y on T and X (the analysis model) applied to the complete cases. Applying the analysis model to the subset with covariate values imputed through conditional mean imputation yields a biased treatment effect estimate (B).



Likewise, in the case of missing (e.g. MCAR) X values, averaging (transformed) imputed values across many multiply imputed datasets (i.e. effectively conditional mean imputation) also renders the effect estimator biased. The crux of the matter lies in that the default imputation model is the linear regression with X as the dependent variable and T and Y as the independent variables, whereas the analysis model regards Y as the dependent variable. Switching dependent and independent variables results in best fit equations that are not in general equivalent (unless orthogonal regression is used). Bias can therefore also be expected for the Across approach, because in the context of missing covariate data it is comparable to conditional mean imputation, except that taking the logit of the average propensity score is not the same as taking the average of the logit of propensity scores.

We shall now describe the asymptotic behaviour of the Across approach in a simple setting with a binary confounder

X, binary treatment variable T, and binary outcome Y; missingness is MCAR and univariate, only affecting X.

Let ε be a uniformly distributed random variable over the interval (0, 1). Define Y to be equal to $\varepsilon < f(T,X)$ 1 if for some deterministic mapping f to the interval (0, 1), and let it be 0 otherwise. For any subject with realisations u of ε and x of X, let $Y_0 = I(u<f(0,x))$ and $Y_1 = I(u<f(1,x))$ being the indicator function) denote the outcomes if treatment were set, possibly contrary to fact, to 0 and 1, respectively. Suppose further that ε is independent of T given X, so that there is no unmeasured confounding, i.e., $(Y_0, Y_1) \perp T|X$. We also assume consistency throughout, i.e., that for any treated subject is observed and for any untreated subject is observed. In addition, it is assumed that positivity (i.e., $0<\Pr(T=1|X=x)<1$ for $x=0, 1$) holds.

The marginal odds ratio OR for the causal effect of T on Y among the treated (ATT) is

$$\mathrm{OR} = \frac{E[Y_1|T=1]/(1-E[Y_1|T=1])}{E[Y_0|T=1]/(1-E[Y_0|T=1])} \quad (1)$$

By the consistency assumption, $E[Y_1|T=1]=E[Y|T=1]$. Since $f(T,X)=Pr(Y=1|T,X)$,

$$\begin{aligned} E[Y_0|T=1] &= E_X\{E[Y_0|T=1,X]|T=1\} \\ &= E_X\{E[\varepsilon < f(0,X)|T=1,X]|T=1\} \\ &= E_X\{\Pr(Y=1|T=0,X)\,|T=1\} \end{aligned}$$

Hence, we may rewrite (1) as follows:

$$\mathrm{OR} = \frac{E[Y|T=1]/(1-E[Y|T=1])}{E_X\{Pr(Y=1|T=0,X)\,|T=1\}/(1-E_X\{Pr(Y=1|T=0,X)\,|T=1\})} \quad (2)$$

Now consider the inverse probability weighting estimator

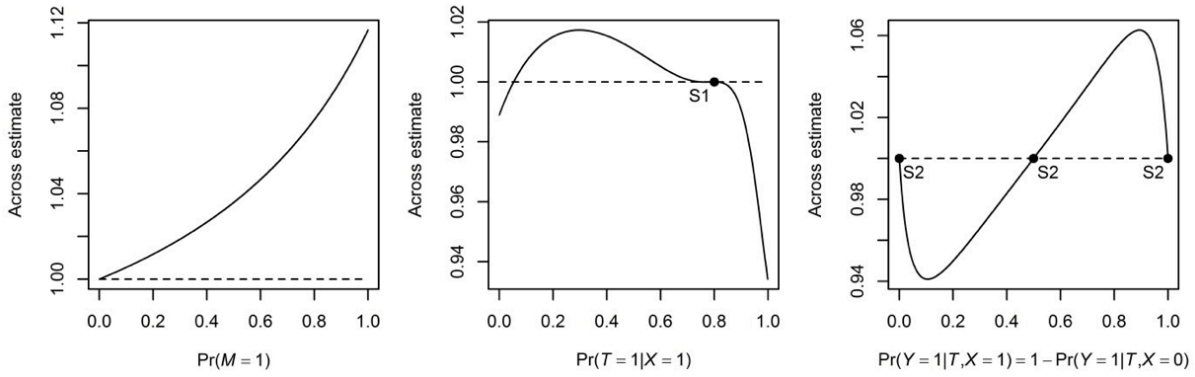
$$\mathrm{OR} = \frac{E[WY|T=1]/(1-E[WY|T=1])}{E[WY|T=0]/(1-E[WY|T=0])} \quad (3)$$

with weights W defined as follows. First, let $W^*=1$ if $T=1$ and $W^*=Pr(T=1|X)/Pr(T=0|X)$ if $T=0$. Then, let $W=W^*/E[W^*|T=1]$ if $T=1$ and $W=W^*/E[W^*|T=0]$ if $T=0$. Clearly, the numerator of (2) coincides with that of (3). Also, we have

$$\begin{aligned} E[WY|T=0] &= \frac{1}{E[W^*|T=0]} E\left[\frac{\Pr(T=1|X)}{\Pr(T=0|X)} Y|T=0\right] \\ &= \frac{1}{E[W^*|T=0]} E_X\left\{\frac{\Pr(T=1|X)}{\Pr(T=0|X)} E[Y|T=0,X]|T=0\right\} \\ &= \frac{1}{E[W^*|T=0]}\left\{\frac{\Pr(T=1|X=0)}{\Pr(T=0|X=0)} E[Y|T=0,X=0]\,\middle|\,T=0\right. \\ &\quad \left. + \frac{\Pr(T=1|X=1)}{\Pr(T=0|X=1)} E[Y|T=0,X=1]\,\middle|\,T=0\right\} \\ &= \frac{1}{E[W^*|T=0]} \frac{\Pr(T=1)}{\Pr(T=0)} E_X\{\Pr(Y=1|T=0,X)|T=1\} \end{aligned} \quad (4)$$

It is easily verified that $E[W^*|T=0]=Pr(T=1)/Pr(T=0)$;

FIGURE 2. Asymptotic results (solid lines) of the Across inverse probability weighting estimator of the odds ratio for the marginal causal treatment-outcome effect in a hypothetical setting with a binary confounder, binary treatment variable and binary outcome (default parameter values given in text). Dashed lines indicate the true OR of 1. S1 and S2 represent sufficient conditions for asymptotic unbiasedness; see text.



$$
\begin{aligned}
E[W^*|T=0] &= E\left[\frac{\Pr(T=1|X)}{\Pr(T=0|X)}|T=0\right] \\
&= E\left[\frac{\Pr(T=1|X)}{\Pr(T=0|X)}|T=0, X=0\right]\Pr(X=0|T=0) \\
&+ E\left[\frac{\Pr(T=1|X)}{\Pr(T=0|X)}|T=0, X=1\right]\Pr(X=1|T=0) \\
&= \frac{\Pr(T=1|X=0)}{\Pr(T=0|X=0)}\Pr(X=0|T=0) \\
&+ \frac{\Pr(T=1|X=1)}{\Pr(T=0|X=1)}\Pr(X=1|T=0) \\
&= \frac{\Pr(T=1|X=0)\Pr(X=0)}{\Pr(T=0)} + \frac{\Pr(T=1|X=1)\Pr(X=1)}{Pr(T=0)} \\
&= \frac{\Pr(T=1)}{\Pr(T=0)}
\end{aligned}
$$

Thus, the right-hand side of (4) becomes $E_X\{Pr(Y=1|T=0,X)|T=1\}$, and, therefore, (3) is equal to the right-hand side of (2).

Now suppose that X is unobserved with constant non-zero probability. Clearly, since the missingness is completely at random, an application of the above inverse probability weighting estimator to the complete records would result in correct inference as to the OR. However, the missing data mechanism is typically unknown to the researcher, so that one might opt for multiple imputation and proceed under the less restrictive ignorability assumption. Since all variables under consideration are binary, misspecification may easily be averted by selecting a fully saturated imputation model, e.g., an additive logistic regression of X on T and Y, allowing for interaction between T and Y, so that $Pr(X=1|T,Y)$ is consistently estimated from the observed data. Following imputation, the propensity score model parameterised by $Pr(T=1|X)$ can also be consistently estimated from each imputed sample.

With the number of imputations m and sample size tending to infinity, the Across approach produces estimated propensity scores that equal their true equivalents for subjects with observed covariate values; however, for an incomplete record with realisations t and y of T and Y, the estimated propensity score equals

$$\sum_x \Pr(T=1|X=x)\Pr(X=x|T=t,Y=y) \quad (5)$$

The Across propensity score estimate (5) is not generally equal to the respective true propensity score. Sufficient conditions for equality (and therefore asymptotic unbiasedness of the Across approach) are (S1) the independence between T and X and (S2) the deterministic relation between X and Y given T.

We may compute the quantity estimated by the Across approach in the context of inverse probability weighting, with weights defined with the ATT in mind, as follows. As above, the weights are defined as a function of the (estimated) propensity scores; specifically, redefine W^* such that

$$W^* = \begin{cases} 1 & \text{if } T=1 \\ \Pr(T=1|X)/\Pr(T=0|X) & \text{if } T=0 \wedge M=0 \\ A/(1-A) & \text{if } T=0 \wedge M=1 \end{cases}$$

with A denotinig $\sum_x Pr(T=1|X=x)\ Pr(X=x|T=0,Y)$ and M being the indicator variable that takes the value of 1 if X is missing and 0 otherwise. Next, observe that $E[WY|T=1]=E[Y|T=1],$, as before, and

$$
\begin{aligned}
E[WY|T=0] &= \frac{E[W^*Y|T=0]}{E[W^*|T=0]} \\
&= \frac{E_{MXY}\{E[W^*Y|T=0,M,X,Y]|T=0\}}{E_{MXY}\{E[W^*|T=0,M,X,Y]|T=0\}}
\end{aligned}
\quad (6)
$$

The numerator of (6) may be evaluated using

$$
\begin{aligned}
&E_{MXY}\{E[W^*Y|T=0,M,X,Y]|T=0\} \\
&= \sum_m \sum_y \sum_x \{yE[W^*|T=0,M=m,X=x,Y=y] \\
&\times \Pr(M=m)\Pr(Y=y|X=x,T=0)\Pr(X=x|T=0)\}
\end{aligned}
$$

Similarly, the denominator may be evaluated using

$$
\begin{aligned}
&E_{MXY}\{E[W^*|T=0,M,X,Y]|T=0\} \\
&= \sum_m \sum_y \sum_x \{E[W^*|T=0,M=m,X=x,Y=y] \\
&\times \Pr(M=m)\Pr(Y=y|X=x,T=0)\Pr(X=x|T=0)\}
\end{aligned}
$$

Substituting $E[WY|T=1]$ and E[WY|T=0] in (3) with (6) and E[WY|T=1], respectively, yields the quantity asymptotically estimated by the Across approach.

To fix ideas, suppose that X, T, Y, and M are distributed such that

$$
\begin{aligned}
\Pr(X=1) &= 0.5 \\
\Pr(T=1|X=x) &= \begin{cases} 0.2 & \text{if } x=0 \\ 0.8 & \text{if } x=1 \end{cases}
\end{aligned}
$$

$$
\Pr(Y=1|T=t,X=x) = \begin{cases} 0.2 & if\ t=0 \wedge x=0 \\ 0.4 & if\ t=0 \wedge x=1 \\ 0.2 & if\ t=1 \wedge x=0 \\ 0.4 & if\ t=1 \wedge x=1 \end{cases}
$$

$$\Pr(M=1|Y=y,T=t,X=x) = \Pr(M=1) = 0.25$$

The discrepancies between the true OR of 1 (consistently estimated by the Within approach) and the asymptotic results of the Across approach for the above distribution and one-way deviations are depicted in Figure 2.

B Practical non-positivity and iterative inverse probability weighting

Given the propensity score, treated and untreated subjects tend to be comparable or 'exchangeable' with respect to measured covariates. In 'counterfactual outcomes parlance, the distribution of potential outcomes is expected to be the same for subjects with the same propensity score. Estimators typically compare outcomes between groups of subjects with 'exchangeable' observations. However, these comparisons become problematic if a treatment group has observations that are not 'exchangeable' with any of those in the other group. This problem represents a violation of the practical positivity assumption, i.e. that for each level of confounders there are records of both treated and untreated individuals in the dataset. Regression adjustment responds to this problem by extrapolating over covariate/propensity score regions of non-positivity through modelling the outcome, which may be appropriate under homogeneity assumptions, yet less attractive when these are not tenable. Indeed, one might question the ability of the model to accurately predict outside the region of positivity. Semi- or nonparametric methods, such as matching and IPW, do not involve explicit modelling of the outcome with respect to the PS, but are sensitive to practical non-positivity.

The consequences in terms of bias due to practical non-positivity are perhaps most intuitive in the setting where propensity score matching is used to estimate the ATT. When the upper tail of the propensity score distribution of the treated group shares no support with the propensity score distribution of the untreated group, it may be that for those subjects in this propensity score region the most suitable (closest in terms of the propensity score) matches are those with a substantially lower propensity score. As a result, unless corrections for non-positivity are made, treatment is still associated with the covariates in the matched set, potentially leading to bias.

IPW adjusts for confounding by weighting observations such that the association between treatment and confounders is removed. Suppose that at some (possibly multidimensional) covariate level, there are only treated subjects in the sample. With mere categorical confounders, this would preclude the fitting of a propensity score model. With continuous covariates, random zeros are inevitable because of the infinite number of confounder levels. In such settings, parametric models can be used to obtain estimated propensity scores by 'borrowing' information from individuals with similar covariate values to those that have zero probability of occurring. However, although weights may be defined at every level of the covariate, reweighting observations that have zero frequency of occurring is impossible. As such, if zero frequencies of being assigned to treatment (but not to the control group) tend to occur in, say, the lower or upper end of the covariate distribution, treatment will still be associated with the covariate in the pseudopopulation, so that the IPW estimator may be biased by residual confounding.

To improve covariate balance in the presence of practical non-positivity, Van der Wal proposed an algorithm in which the dataset is iteratively reweighted [3]. The idea underpinning this algorithm is as follows. By fitting a propensity model on the weighted dataset, new weights can be estimated that (partially) adjust for the residual confounding. Multiplying these weights with the original yields weights that, when applied to the dataset, correct for confounding more than the original weights. As the covariate balance improves, the probability of being assigned to treatment becomes less dependent on the covariate values, and so the variance of the log-transformed new weights reduces. The above process is therefore repeated until the variance drops below a convergence threshold.

The iterative inverse probability weighting (IIPW) algorithm was defined in the context of complete data. With multiply imputed data, one can apply IIPW within each imputed dataset, in a way consistent with the Within approach, until the algorithm converges within each dataset or until a maximum number of iterations is reached. Alternatively, at each iteration one can average the estimated propensity scores across the imputed datasets, as per the Across approach, before reweighting the imputed datasets. Sample R code for these algorithms is given below.

```
truncate <- function(x,left=0.05,right=0.95){
                Q <- quantile(x,probs=c(left,right))
                x[x>Q[2]] <- Q[2]; x[x<Q[1]] <- Q[1]
                return(x)
}

IIPTW <- function(
                # Iterative Inverse Probability of Treatment Weighting
                # Returns a list with weights and number of iterations
                data, # a dataframe containing columns T, X1, X2 and Y
                formula =`T~X1+X2',
                T =`T'
```

```
                left =0,
                right =1,
                cstop =1e-4,
                maxit =100,
                estimand =`ATE'
                ){
                warning <- FALSE
                it <- 1
                n <- nrow(data)
                w <- rep(1,n)
                for(i in 1:maxit){
                        psnew <- predict(glm(formula=as.formula(formula),
                                family=binomial,data=data,weights=w),type=`response')
                        if(estimand==`ATE'){wnew <- ifelse(data[,T]==1,1/psnew,1/(1-psnew))}
                        if(estimand==`ATT'){wnew <- ifelse(data[,T]==1,1,psnew/(1-psnew))}
                        if(estimand==`ATU'){wnew <- ifelse(data[,T]==1,(1-psnew)/psnew,1)}
                        if(i>=2 && var(log(wnew))<=cstop){it <- i; break}
                        if(i==maxit){it <- i; warning <- TRUE}
                        w <- truncate(w*wnew,left,right)
                        w <- w/mean(w)
                }
                if(warning==TRUE){warning(`Algorithm did not converge');it<-NA}
                return(list(w=data$w,it=it))
}

IIPTW.A <- function(
                # Iterative Inverse Probability of Treatment Weighting
                # consistent with the Across approach
                # Returns a list with weights for each imputed dataset and the number
                # of iterations
                data, # a list of multiply imputed datasets
                formula =`T~X1+X2',
                T =`T',
                left =0,
                right =1,
                cstop =1e-4,
                maxit =100,
                estimand =`ATE'
                ){
                warning <- FALSE
                it <- 1
                m <- length(data)
                n <- nrow(data[[1]])
                for(i in 1:m){data[[i]]$w <- rep(1,n)}
                psnew <- matrix(nrow=n,ncol=m)
                for(i in 1:maxit){
                        for(u in 1:m){
                                psnew[,u] <- predict(glm(formula=as.formula(formula),
                                                family=binomial,data=data[[u]],
                                                weights=data[[u]]$w),type=`response')
                        }
                        psnewA <- apply(psnew,1,mean)
                        for(u in 1:m){
                        if(estimand==`ATE'){data[[u]]$wnew <-
                                ifelse(data[[u]][,T]==1,1/psnewA,1/(1-psnewA))}
                        if(estimand==`ATT'){data[[u]]$wnew <-
```

```
                                    ifelse(data[[u]][,T]==1,1,psnewA/(1-psnewA))}
                        if(estimand==`ATU'){data[[u]]$wnew <-
                                    ifelse(data[[u]][,T]==1,(1-psnewA)/psnewA,1)}
                        }
                        if(i>=2 && prod(unlist(lapply(data,FUN=function(x)
                            {var(log(x$wnew))<=cstop})))==1){it <- i; break}
                        if(i==maxit){it <- i; warning <- TRUE}
                                    for(u in 1:m){
                        data[[u]]$w <- truncate(data[[u]]$w*data[[u]]$wnew,
                                    left=left,right=right)
                        data[[u]]$w <- data[[u]]$w/mean(data[[u]]$w)
                        }
            }
            if(warning==TRUE){warning(`Algorithm did not converge.');it<-NA}
            return(list(w=lapply(data,function(x)x$w),it=it))
}

IIPTW.W <- function(
            # Iterative Inverse Probability of Treatment Weighting
            # consistent with the Within approach
            # Returns a list with weights for each imputed dataset and the number # of iterations
            data, # a list of multiply imputed datasets
            formula =`T~X1+X2',
            T =`T',
            left =0,
            right =1,
            cstop =1e-4,
            maxit =100,
            estimand =`ATE'
            ){
            out <- lapply(data,FUN=IIPTW,formula=formula,T=T,
                    left=left,right=right,cstop=cstop,maxit=maxit,
                    estimand=estimand)
            w <- lapply(out,function(x)x$w)
            itmax <- max(unlist(lapply(out,function(x)x$it)))
            return(list(w=w,it=itmax))
}
```

C Results

TABLE 1. Performance of treatment effect estimators for various degrees p of missing (MCAR) covariate data and residual variances σ^2.

			CCA			Within			Across		
σ^2	*p*	PS method	Bias	Var.	MSE	Bias	Var.	MSE	Bias	Var.	MSE
1	20	Regression	-0.002	0.016	0.016	-0.003	0.015	0.015	-0.063	0.016	0.020
		Matching	0.037	0.033	0.034	0.037	0.021	0.022	-0.028	0.029	0.030
		C. matching	-0.003	0.032	0.032	-0.001	0.019	0.019	-0.069	0.028	0.033
		IPW	0.091	0.322	0.330	0.067	0.263	0.268	0.077	0.262	0.268
		IIPW	0.001	0.031	0.031	0.008	0.043	0.043	-0.025	0.172	0.173
	40	Regression	-0.003	0.020	0.020	-0.002	0.017	0.017	-0.128	0.022	0.039
		Matching	0.050	0.044	0.046	0.039	0.020	0.021	-0.086	0.032	0.040
		C. matching	-0.005	0.046	0.046	-0.002	0.019	0.019	-0.134	0.032	0.050
		IPW	0.048	0.461	0.463	0.022	0.288	0.288	0.039	0.288	0.289
		IIPW	0.006	0.088	0.088	0.005	0.039	0.039	-0.081	0.029	0.036
	60	Regression	0.011	0.029	0.030	0.010	0.021	0.021	-0.199	0.042	0.082
		Matching	0.078	0.071	0.077	0.051	0.024	0.026	-0.150	0.048	0.071
		C. matching	0.015	0.075	0.075	0.012	0.024	0.024	-0.203	0.054	0.096
		IPW	0.127	0.566	0.582	0.061	0.257	0.261	0.088	0.254	0.262
		IIPW	0.017	0.064	0.064	0.017	0.036	0.036	-0.132	0.038	0.056
	80	Regression	0.008	0.060	0.060	0.002	0.042	0.042	-0.335	0.128	0.240
		Matching	0.112	0.172	0.185	0.055	0.040	0.043	-0.268	0.117	0.189
		C. matching	0.004	0.198	0.198	0.006	0.045	0.045	-0.340	0.143	0.258
		IPW	0.281	0.884	0.963	0.079	0.264	0.271	0.127	0.254	0.270
		IIPW	0.014	0.107	0.108	0.018	0.053	0.054	-0.217	0.089	0.136
9	20	Regression	0.013	0.134	0.134	-0.002	0.112	0.112	-0.106	0.116	0.127
		Matching	0.063	0.242	0.246	0.029	0.142	0.142	-0.071	0.209	0.214
		C. matching	0.021	0.251	0.251	-0.008	0.139	0.139	-0.116	0.217	0.231
		IPW	0.082	0.539	0.545	0.057	0.405	0.409	0.075	0.410	0.416
		IIPW	0.021	0.284	0.285	0.012	0.207	0.207	-0.056	0.220	0.223
	40	Regression	-0.003	0.174	0.174	0.009	0.115	0.115	-0.222	0.139	0.188
		Matching	0.047	0.314	0.316	0.051	0.140	0.142	-0.188	0.202	0.237
		C. matching	0.001	0.339	0.339	0.011	0.143	0.143	-0.249	0.217	0.279
		IPW	0.085	0.747	0.754	0.077	0.430	0.436	0.109	0.433	0.445
		IIPW	-0.012	0.346	0.346	0.010	0.224	0.224	-0.148	0.240	0.261
	60	Regression	0.001	0.287	0.287	0.009	0.130	0.130	-0.363	0.214	0.346
		Matching	0.073	0.489	0.494	0.050	0.146	0.149	-0.321	0.278	0.380
		C. matching	0.012	0.548	0.548	0.012	0.153	0.153	-0.386	0.315	0.464
		IPW	0.111	1.172	1.184	0.088	0.457	0.465	0.135	0.458	0.476
		IIPW	0.013	0.550	0.550	0.031	0.283	0.284	-0.246	0.413	0.474
	80	Regression	0.027	0.589	0.589	0.015	0.180	0.180	-0.593	0.490	0.842
		Matching	0.142	1.076	1.096	0.075	0.189	0.194	-0.507	0.483	0.740
		C. matching	0.018	1.225	1.225	0.024	0.198	0.199	-0.625	0.612	1.002
		IPW	0.214	1.848	1.894	0.091	0.459	0.467	0.170	0.435	0.464
		IIPW	0.005	0.997	0.997	0.029	0.290	0.291	-0.503	0.483	0.735

Abbreviations: CCA, complete case analysis; p, missingness probability (%); PS method, propensity score method; Var., empirical variance; MSE, empirical mean squared error; C. matching, calliper matching; IPW, inverse probability weighting; IIPW, iterative inverse probability weighting.

TABLE 2. Performance of treatment effect estimators for various degrees p of missing (MCAR) treatment indicator values and residual variances σ^2.

			CCA			Within			Across		
σ^2	*p*	PS method	Bias	Var.	MSE	Bias	Var.	MSE	Bias	Var.	MSE
1	20	Regression	0.008	0.015	0.015	-0.003	0.014	0.014	-0.005	0.014	0.014
		Matching	0.049	0.032	0.035	0.042	0.020	0.022	0.038	0.027	0.028
		C. matching	0.003	0.032	0.032	-0.002	0.018	0.018	-0.006	0.026	0.026
		IPW	0.041	0.347	0.348	0.017	0.253	0.254	0.018	0.247	0.247
		IIPW	0.005	0.029	0.029	-0.005	0.033	0.033	-0.026	0.027	0.027
	40	Regression	0.004	0.019	0.019	-0.026	0.015	0.016	-0.031	0.015	0.016
		Matching	0.048	0.044	0.046	0.017	0.019	0.019	0.014	0.024	0.024
		C. matching	-0.003	0.044	0.044	-0.024	0.019	0.020	-0.027	0.024	0.025
		IPW	0.104	0.423	0.433	0.029	0.211	0.212	0.029	0.203	0.204
		IIPW	0.001	0.054	0.054	-0.022	0.039	0.039	-0.059	0.028	0.032
	60	Regression	-0.005	0.033	0.033	-0.071	0.021	0.027	-0.081	0.022	0.028
		Matching	0.064	0.071	0.075	-0.024	0.026	0.027	-0.030	0.032	0.033
		C. matching	-0.005	0.070	0.070	-0.067	0.025	0.030	-0.072	0.031	0.036
		IPW	0.147	0.545	0.567	-0.005	0.165	0.165	-0.006	0.154	0.154
		IIPW	0.005	0.112	0.112	-0.066	0.041	0.045	-0.132	0.040	0.057
	80	Regression	0.015	0.065	0.065	-0.163	0.033	0.060	-0.186	0.034	0.069
		Matching	0.096	0.160	0.169	-0.108	0.040	0.051	-0.111	0.042	0.054
		C. matching	0.003	0.187	0.187	-0.157	0.039	0.063	-0.160	0.042	0.068
		IPW	0.278	0.927	1.004	-0.111	0.128	0.141	-0.104	0.113	0.124
		IIPW	0.020	0.170	0.171	-0.160	0.045	0.070	-0.257	0.076	0.142
9	20	Regression	0.011	0.131	0.131	0.002	0.131	0.131	-0.001	0.131	0.131
		Matching	0.053	0.222	0.225	0.045	0.155	0.157	0.043	0.204	0.206
		C. matching	0.004	0.228	0.228	0.001	0.156	0.156	-0.002	0.209	0.209
		IPW	0.099	0.528	0.538	0.074	0.409	0.415	0.077	0.402	0.408
		IIPW	0.023	0.257	0.258	0.021	0.223	0.223	0.003	0.215	0.215
	40	Regression	-0.021	0.183	0.184	-0.046	0.185	0.187	-0.052	0.185	0.188
		Matching	0.026	0.327	0.328	0.001	0.217	0.217	0.001	0.245	0.245
		C. matching	-0.030	0.341	0.342	-0.039	0.215	0.216	-0.037	0.250	0.251
		IPW	0.067	0.815	0.820	0.020	0.457	0.458	0.030	0.433	0.434
		IIPW	0.012	0.541	0.541	-0.030	0.273	0.273	-0.068	0.271	0.276
	60	Regression	0.022	0.297	0.297	-0.047	0.290	0.292	-0.060	0.290	0.294
		Matching	0.072	0.472	0.478	-0.000	0.309	0.309	-0.002	0.324	0.324
		C. matching	0.004	0.532	0.532	-0.046	0.314	0.316	-0.047	0.331	0.333
		IPW	0.146	1.021	1.042	0.016	0.470	0.471	0.037	0.449	0.450
		IIPW	0.015	0.485	0.485	-0.044	0.356	0.358	-0.093	0.352	0.360
	80	Regression	-0.017	0.549	0.549	-0.161	0.501	0.527	-0.188	0.504	0.540
		Matching	0.125	1.011	1.026	-0.100	0.527	0.536	-0.106	0.538	0.549
		C. matching	0.052	1.189	1.191	-0.160	0.535	0.561	-0.165	0.553	0.581
		IPW	0.232	1.954	2.008	-0.068	0.683	0.688	-0.029	0.645	0.646
		IIPW	-0.023	1.029	1.029	-0.130	0.600	0.617	-0.209	0.584	0.627

Abbreviations: CCA, complete case analysis; p, missingness probability (%); PS method, propensity score method; Var., empirical variance; MSE, empirical mean squared error; C. matching, calliper matching; IPW, inverse probability weighting; IIPW, iterative inverse probability weighting.

TABLE 3. Performance of treatment effect estimators for various degrees p of missing (MCAR) outcomes and residual variances σ^2.

			CCA			Within			Across		
σ^2	*p*	PS method	Bias	Var.	MSE	Bias	Var.	MSE	Bias	Var.	MSE
1	20	Regression	0.004	0.014	0.014	0.005	0.015	0.015	0.005	0.015	0.015
		Matching	0.053	0.033	0.036	0.044	0.024	0.026	0.044	0.024	0.026
		C. matching	0.010	0.033	0.033	0.006	0.024	0.024	0.006	0.024	0.024
		IPW	0.084	0.324	0.331	0.059	0.261	0.265	0.059	0.261	0.265
		IIPW	-0.011	0.244	0.244	0.007	0.064	0.064	0.007	0.064	0.064
	40	Regression	0.004	0.020	0.020	0.003	0.022	0.022	0.003	0.022	0.022
		Matching	0.054	0.049	0.052	0.039	0.034	0.035	0.039	0.034	0.035
		C. matching	0.002	0.046	0.046	0.001	0.032	0.032	0.001	0.032	0.032
		IPW	0.087	0.421	0.428	0.049	0.290	0.292	0.049	0.290	0.292
		IIPW	-0.002	0.074	0.074	0.011	0.101	0.101	0.011	0.101	0.101
	60	Regression	0.004	0.031	0.031	0.003	0.035	0.035	0.003	0.035	0.035
		Matching	0.061	0.068	0.071	0.047	0.043	0.045	0.047	0.043	0.045
		C. matching	0.000	0.073	0.073	0.010	0.043	0.043	0.010	0.043	0.043
		IPW	0.171	0.524	0.553	0.056	0.291	0.294	0.056	0.291	0.294
		IIPW	0.011	0.060	0.061	-0.004	0.189	0.189	-0.004	0.189	0.189
	80	Regression	0.003	0.073	0.073	0.000	0.082	0.082	0.000	0.082	0.082
		Matching	0.116	0.166	0.179	0.040	0.092	0.094	0.040	0.092	0.094
		C. matching	0.020	0.193	0.193	-0.001	0.090	0.090	-0.001	0.090	0.090
		IPW	0.234	1.029	1.084	0.045	0.301	0.303	0.045	0.301	0.303
		IIPW	0.015	0.300	0.301	0.002	0.087	0.087	0.002	0.087	0.087
9	20	Regression	-0.004	0.135	0.135	-0.003	0.139	0.139	-0.003	0.139	0.139
		Matching	0.051	0.245	0.248	0.023	0.198	0.199	0.023	0.198	0.199
		C. matching	0.007	0.254	0.254	-0.016	0.199	0.200	-0.016	0.199	0.200
		IPW	0.065	0.499	0.503	0.035	0.450	0.451	0.035	0.450	0.451
		IIPW	-0.015	0.311	0.311	-0.001	0.434	0.434	-0.001	0.434	0.434
	40	Regression	-0.002	0.181	0.181	-0.004	0.199	0.199	-0.004	0.199	0.199
		Matching	0.054	0.333	0.336	0.031	0.255	0.256	0.031	0.255	0.256
		C. matching	-0.002	0.358	0.358	-0.011	0.265	0.265	-0.011	0.265	0.265
		IPW	0.149	0.665	0.687	0.079	0.456	0.463	0.079	0.456	0.463
		IIPW	0.003	0.380	0.380	0.006	0.283	0.283	0.006	0.283	0.283
	60	Regression	0.007	0.293	0.293	0.004	0.326	0.326	0.004	0.326	0.326
		Matching	0.070	0.476	0.481	0.042	0.373	0.375	0.042	0.373	0.375
		C. matching	-0.007	0.548	0.549	0.004	0.381	0.381	0.004	0.381	0.381
		IPW	0.170	1.119	1.148	0.077	0.675	0.681	0.077	0.675	0.681
		IIPW	0.034	0.637	0.638	0.019	0.393	0.393	0.019	0.393	0.393
	80	Regression	0.007	0.597	0.597	0.024	0.698	0.699	0.024	0.698	0.699
		Matching	0.139	1.055	1.074	0.053	0.723	0.726	0.053	0.723	0.726
		C. matching	0.034	1.244	1.245	0.018	0.727	0.728	0.018	0.727	0.728
		IPW	0.320	1.938	2.041	0.060	1.000	1.004	0.060	1.000	1.004
		IIPW	0.036	1.115	1.116	0.004	0.824	0.824	0.004	0.824	0.824

Abbreviations: CCA, complete case analysis; p, missingness probability (%); PS method, propensity score method; Var., empirical variance; MSE, empirical mean squared error; C. matching, calliper matching; IPW, inverse probability weighting; IIPW, iterative inverse probability weighting.

TABLE 4. Performance of treatment effect estimators for various (MAR) missingness mechanisms and residual variances σ^2.

			CCA			Within			Across		
σ^2	*p*	PS method	Bias	Var.	MSE	Bias	Var.	MSE	Bias	Var.	MSE
1	MAR1	Regression	-0.003	0.016	0.016	0.002	0.014	0.014	-0.059	0.016	0.020
		Matching	1.054	0.075	1.187	0.042	0.018	0.020	-0.084	0.029	0.036
		C. matching	-0.010	0.034	0.034	0.005	0.016	0.016	-0.115	0.028	0.041
		IPW	-0.546	1.373	1.671	0.061	0.271	0.275	0.068	0.269	0.274
		IIPW	-0.003	1.128	1.128	-0.002	0.036	0.036	-0.021	0.024	0.025
	MAR2	Regression	-0.186	0.055	0.090	-0.011	0.036	0.036	-0.216	0.065	0.111
		Matching	-0.172	0.205	0.235	0.039	0.036	0.038	-0.236	0.088	0.144
		C. matching	-0.181	0.210	0.243	-0.004	0.039	0.039	-0.303	0.105	0.197
		IPW	-0.005	0.187	0.187	0.038	0.283	0.284	-0.003	0.307	0.307
		IIPW	-0.153	0.121	0.144	-0.001	0.045	0.045	-0.145	0.057	0.078
9	MAR1	Regression	-0.007	0.145	0.145	0.001	0.112	0.112	-0.105	0.117	0.128
		Matching	1.044	0.257	1.347	0.046	0.138	0.140	-0.178	0.203	0.235
		C. matching	-0.001	0.297	0.297	0.011	0.141	0.141	-0.201	0.214	0.254
		IPW	-0.532	2.143	2.425	0.081	0.409	0.415	0.093	0.407	0.416
		IIPW	-0.003	2.358	2.358	0.001	0.210	0.210	-0.026	0.208	0.209
	MAR2	Regression	-0.812	0.225	0.885	-0.014	0.147	0.147	-0.394	0.234	0.389
		Matching	-0.745	0.506	1.061	0.038	0.166	0.167	-0.362	0.305	0.436
		C. matching	-0.746	0.516	1.072	-0.016	0.170	0.170	-0.487	0.388	0.625
		IPW	-0.540	0.542	0.833	0.016	0.472	0.472	-0.134	0.544	0.562
		IIPW	-0.654	0.570	0.998	-0.024	0.234	0.234	-0.440	0.318	0.512

Abbreviations: CCA, complete case analysis; MDM, missing data mechanism; PS method, propensity score method; Var., empirical variance; MSE, empirical mean squared error; CP, empirical coverage probability, VR, ratio of mean estimated variance to empirical variance; C. matching, calliper matching; IPW, inverse probability weighting; IIPW, iterative inverse probability weighting. Under mechanism MAR1, the missingness of X_2 depends on X_1 and T only. Under MAR2, the missingness depends on Y only. Both MAR1 and MAR2 result in ~40% incomplete records.

References

1. Mitra R, Reiter JP. A comparison of two methods of estimating propensity scores after multiple imputation. Statistical methods in medical research. 2016;25(1):188-204.
2. Wan, F, Mitra, N. An evaluation of bias in propensity score-adjusted non-linear regression models. Statistical methods in medical research. 2016;0(0);1-17.
3. Van der Wal WM. Causal modeling in epidemiological practice [PhD thesis]. University of Amsterdam; 2011. Available from: http://hdl.handle.net/11245/1.366252.