A comparison of two methods of estimating propensity scores after multiple imputation

ROBIN MITRA*

School of Mathematics

University of Southampton, Southampton, SO17 1BJ, UK

Tel. (+44) 2380 594550 Email: R.Mitra@soton.ac.uk

JEROME P. REITER

Department of Statistical Science

Duke University, Box 90251, Durham, NC 27708, USA

Tel. (+1) 919 6685227 Email: jerry@stat.duke.edu

* *corresponding author*

Abstract

In many observational studies, analysts estimate treatment effects using propensity scores, e.g., by matching or sub-classifying on the scores. When some values of the covariates are missing, analysts can use multiple imputation to fill in the missing data, estimate propensity scores based on the m completed datasets, and use the propensity scores to estimate treatment effects. We compare two approaches to implement this process. In the first, the analyst estimates the treatment effect using propensity score matching within each completed data set, and averages the m treatment effect estimates. In the second approach, the analyst averages the m propensity scores for each record across the completed datasets, and performs propensity score matching with

these averaged scores to estimate the treatment effect. We compare properties of both methods via simulation studies using artificial and real data. The simulations suggest that the second method has greater potential to produce substantial bias reductions than the first, particularly when the missing values are predictive of treatment assignment.

Keywords: Missing data; Multiple imputation; Observational studies; Propensity score.

1 INTRODUCTION

In many studies of causal effects, analysts can reduce the bias that results from imbalanced covariate distributions, at least for observed covariates, using propensity score matching[1–5]. The propensity score for any subject, $e(\boldsymbol{x}_i)$, is the probability that the subject receives the treatment given its vector of covariates $\boldsymbol{x}_i$; that is, $e(\boldsymbol{x}_i) = P(T_i = 1|\boldsymbol{x}_i)$, where $T_i = 1$ if subject i receives treatment and $T_i = 0$ otherwise. If two units have the same propensity score, then their covariates can be shown to come from the same distribution[1]. Thus, by selecting control units whose propensity scores are similar to the treated units' propensity scores, analysts can create a matched control group whose covariates are similar to the treated group's covariates. Analysts then base inference on the treated and matched control groups, thereby avoiding any bias that results from imbalanced covariate distributions in the two groups, at least for those covariates in $\boldsymbol{x}$. Other approaches to causal inference based on propensity scores include sub-classification[6,7], full matching[8,9] and propensity score weighted-estimation[10].

Propensity scores are typically estimated via regressions of T on functions of $\boldsymbol{x}$[11–14]. When some covariate data are missing, these complete-data methods cannot be easily applied. Several strategies exist for overcoming this complication[6,15–17]. In this article, we focus on the use of multiple imputation[18] to fill in the missing covariate data. In multiple imputation, the analyst repeatedly imputes missing values

by sampling from their predictive distributions (estimated with the observed data) to create $m > 1$ completed datasets. The analyst then performs the complete-data analysis in each imputed dataset and makes inferences by combining the resulting point and variance estimates[19].

After multiply-imputing the missing covariate data, the analyst can estimate the propensity scores in each dataset via complete-data methods, thus obtaining m values of each unit's propensity score. What should the analyst do with these multiple propensity scores? One approach is to match treated and control units within each completed dataset, resulting in m estimates of treatment effects. The analyst then averages these m treatment effect estimates as the multiple imputation point estimate. We call this the Within approach. Another approach is to average each unit's m propensity scores, match treated and control units based on their averaged scores, and estimate the treatment effect from this single set of matched controls. We call this the Across approach. Both of these approaches seem intuitively reasonable strategies: which can we expect to be more effective? To our knowledge, this question has not been thoroughly investigated, except for one simulation study that demonstrated its complexities[20].

In this article, we shed further light on this issue. To do so, we use two types of simulations: a simple setting with artificial data, and a complicated setting with actual data. In both, our goal is to estimate an average treatment effect on those exposed, which we denote as τ. In Section 2, we formally define the Across and Within approaches. In Section 3, we compare properties of point estimates from the two approaches using simulation studies with artificial data. In Section 4, we extend these simulations to show that iterating the Across approach can reduce mean squared errors. In Section 5, we discuss difficulties in using the Across and Within approaches for variance estimation via the usual multiple imputation formulas. In Section 6, we compare the two approaches on genuine data concerning the effect of breast feeding on the child's cognitive development later in life. Finally, in Section 7, we conclude with a summary of our findings.

2 Across and Within approaches

Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ be an $n \times p$ matrix of covariates, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ corresponds to the ith unit's covariates, where $i = 1, \ldots, n$. For each $\boldsymbol{x}_i$, let $\boldsymbol{m}_i = (m_{i1}, \ldots, m_{ip})'$ be a vector of missing data indicators. Here, $m_{ij} = 1$ indicates x_{ij} is missing, and $m_{ij} = 0$ indicates x_{ij} is observed. Let $\boldsymbol{M} = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n)'$ be the $n \times p$ matrix of missing data indicators for $\boldsymbol{X}$. Let $\boldsymbol{X}_{\text{mis}} = \{x_{ij} : m_{ij} = 1\}$ and $\boldsymbol{X}_{\text{obs}} = \{x_{ij} : m_{ij} = 0\}$. For each unit i, the binary treatment indicator is $T_i \in \{0, 1\}$, and the outcome is Y_i. Let $\boldsymbol{T} = (T_1, \ldots, T_n)'$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$. We assume that $\boldsymbol{T}$ and $\boldsymbol{Y}$ are fully observed.

In multiple imputation, values of $\boldsymbol{X}_{\text{mis}}$ are filled in m times with draws from the predictive distribution, $p(\boldsymbol{X}_{\text{mis}}|\boldsymbol{X}_{\text{obs}}, \boldsymbol{T})$, resulting in m completed datasets $\boldsymbol{X}^{(1)}_{\text{com}}, \ldots, \boldsymbol{X}^{(m)}_{\text{com}}$. For each $\boldsymbol{X}^{(k)}_{\text{com}}$, let $e(\boldsymbol{x}^{(k)}_{i,\text{com}})$ be the estimated propensity score for unit i, where $i = 1, \ldots, n$ and $k = 1, \ldots, m$. Each $e(\boldsymbol{x}^{(k)}_{i,\text{com}})$ is estimated using only the data in $\boldsymbol{X}^{(k)}_{\text{com}}$, for example with a logistic regression of $\boldsymbol{T}$ on some function of $\boldsymbol{X}^{(k)}_{\text{com}}$.

In the Across approach, we estimate the propensity score for each unit, $e^{A,m}(\boldsymbol{x}_i)$, by averaging $e(\boldsymbol{x}^{(k)}_{i,\text{com}})$ over the imputations, so that

$$e^{A,m}(\boldsymbol{x}_i) \quad = \quad \frac{\sum_{k=1}^{m} e(\boldsymbol{x}^{(k)}_{i,\text{com}})}{m}. \tag{1}$$

Let $\boldsymbol{e}^{A,m} = (e^{A,m}(\boldsymbol{x}_1), \ldots, e^{A,m}(\boldsymbol{x}_n))'$. Analysts use $\boldsymbol{e}^{A,m}$ to find a matched control set; in this article we assume analysts use a one-to-one nearest neighbour matching scheme without replacement, although alternative matching schemes such as matching with replacement could also be used. Given the matched set, the analyst estimates τ in the Across approach with

$$\hat{\tau}^{A,m} = \bar{Y}_T - \bar{Y}^{A,m}_{mc}, \tag{2}$$

where $\bar{Y}^{A,m}_{mc}$ is the mean of the matched control units' outcomes selected in the Across approach and $\bar{Y}_T$ is the mean of the treated units' outcomes.

The Within approach uses the propensity scores estimated from each completed dataset, $\boldsymbol{e}(\boldsymbol{X}_{\text{com}}^{(k)}) = (e(\boldsymbol{x}_{1,\text{com}}^{(k)}), \ldots e(\boldsymbol{x}_{n,\text{com}}^{(k)}))'$, to obtain m matched control sets, one for each $\boldsymbol{X}_{\text{com}}^{(k)}$; that is, matching is performed separately in each $\boldsymbol{X}_{\text{com}}^{(k)}$. Let $\bar{Y}_{mc}^{(k)}$ be the average of the outcomes for the matched controls in $\boldsymbol{X}_{\text{com}}^{(k)}$, where $k = 1, \ldots, m$. Let $\hat{\tau}^{W,m,k} = \bar{Y}_T - \bar{Y}_{mc}^{(k)}$. The analyst estimates the treatment effect for the Within approach using

$$\hat{\tau}^{W,m} = \sum_{k=1}^{m} \hat{\tau}^{W,m,k}/m. \tag{3}$$

3 Simulation study of point estimate properties

We now compare the Across and Within approaches using simulations with artificial data. For each simulation run, we generate two covariates $\boldsymbol{X}$ for $n = 1100$ records such that $\boldsymbol{x}_i = (x_{i1}, x_{i2})' \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (10, 10)'$, and $\boldsymbol{\Sigma}$ has variances equal to 5 with correlation 0.5. We generate the response Y so that, for all i,

$$Y_i = x_{i1} + x_{i2} + \epsilon_i, \quad \epsilon_i \sim N(0, 1). \tag{4}$$

Here, without loss of generality for additive treatment effects, $\tau = 0$ for all simulations. We introduce missing data into $\boldsymbol{x}_2$ based on missing at random mechanisms; we leave $\boldsymbol{x}_1$ and $\mathbf{Y}$ fully observed. We consider three mechanisms for assigning treatment, including (i) assignment depends only on $\boldsymbol{x}_1$, (ii) assignment depends only on $\boldsymbol{x}_2$, and (iii) assignment depends equally on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. We assign treatments so that the estimate of τ from the difference in means of the treated and full control groups is severely biased. Results of the simulations are presented in Sections 3.1 to 3.3, and explanations for differences in the performances of the methods are in Section 3.4.

3.1 Simulation 1: treatment assignment depends only on x_1

In this simulation, we assign treatment from Bernoulli distributions where

$$\text{logit}\left(P(T_i = 1|\mathbf{x}_i)\right) = -7.8 + 0.5x_{i1}. \tag{5}$$

Thus, treatment assignment depends only on x_1. In any dataset, this generates approximately 100 treated units and 1000 control units. Figure 1 displays typical covariate patterns that arise from this design.

We consider two mechanisms for introducing missing data in x_2. In the first, we randomly make some control units' x_2 values missing so that

$$\text{logit}\left(P(m_{i2} = 1|T_i = 0, \mathbf{x}_i)\right) = -10.1 + 0.9x_{i1}. \tag{6}$$

In this way, units with larger x_1 values, which are the units most likely to be selected as matches, are more likely to be missing their x_2 values. Approximately 30% of control units' values of x_2 are missing. In the second, we use the same missing data patterns for the control units and also introduce missing values into 30% of the treated units' x_2 through a missing completely at random (MCAR) mechanism. We use the MCAR mechanism because the treated units already tend to have large values of x_1.

We impute missing $\mathbf{x}_2$ from a normal linear regression of $\mathbf{x}_2$ on $(\mathbf{x}_1, T)$ with main effects only, using the appropriate Bayesian posterior predictive distribution with flat prior distributions. We do not control for Y in the imputations. This is done to remain consistent with the philosophy of propensity score matching: manipulation of covariates and the creation of a matched control set is done without consideration of the outcome values[15]. In this way, causal inferences based on the propensity scores are not affected by assumptions about the outcome variable. We note, however, that it can be advantageous to include the outcome variable in imputation models[21,22]. We note that we observed similar results when imputing missing $\mathbf{x}_2$ for treated and control units with separate models.

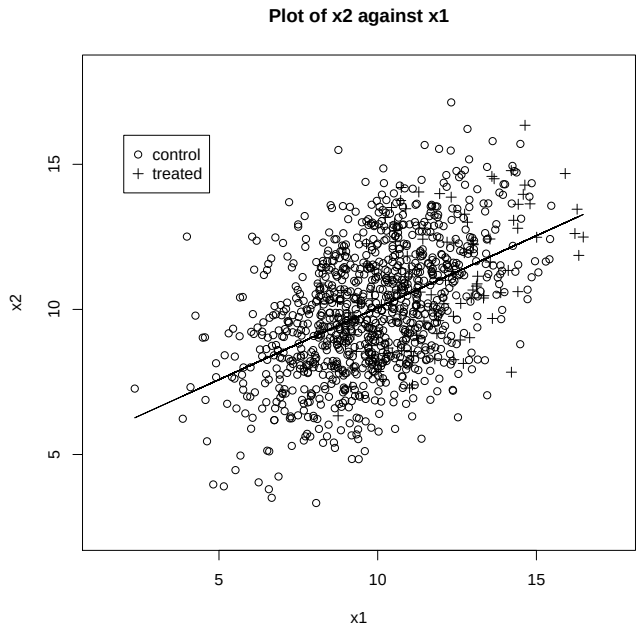

Figure 1: Plot of the covariate distribution in the simulation design where treatment assignment depends on $\boldsymbol{x}_1$ together with the fitted regression line based on a normal linear model for $\boldsymbol{x}_2$.

After multiple imputation of $\mathbf{x}_2$, we estimate the propensity scores $e(\boldsymbol{x}_{i,\text{com}}^{(k)})$ for each unit i in each of $k = 1, \ldots, m$ completed datasets using a logistic regression of $\mathbf{T}$ on $(\mathbf{x}_1, \mathbf{x}_2)$. We then compute $\hat{\tau}^{A,m}$ and $\hat{\tau}^{W,m}$ as in Section 2. We repeat this process 1000 times, each time using new values of $(\mathbf{X}, \mathbf{T}, \mathbf{Y}, \mathbf{M})$. We can then also empirically estimate the true variance of the estimators $\hat{\tau}^{A,m}$ and $\hat{\tau}^{W,m}$ by taking the sample variance of the 1000 estimates.

Table 1 summarizes the point estimates and variances of $\hat{\tau}^{A,m}$ and $\hat{\tau}^{W,m}$ across the 1000 simulations for different values of m. Both the Across and Within approaches result in estimates of τ close to zero. The bias in $\hat{\tau}^{A,m}$ tends to be slightly smaller than that of $\hat{\tau}^{W,m}$, but its variance is slightly larger. The variance of $\hat{\tau}^{W,m}$ appears to decrease as m increases; the variances show no such pattern for $\hat{\tau}^{A,m}$. The Within approach dominates on mean squared error, at least for these values of m.

	Across			Within		
m	Pt. Est.	Variance	MSE	Pt. Est.	Variance	MSE
Only control units missing $\mathbf{x}_2$						
5	0.055	0.077	0.080	0.080	0.050	0.056
10	0.057	0.083	0.086	0.078	0.046	0.052
15	0.065	0.075	0.079	0.079	0.044	0.050
20	0.065	0.077	0.081	0.079	0.043	0.050
50	0.058	0.081	0.085	0.077	0.042	0.048
Treatment and control units missing $\mathbf{x}_2$						
5	0.030	0.080	0.081	0.072	0.053	0.058
10	0.032	0.083	0.084	0.072	0.049	0.055
15	0.031	0.080	0.081	0.074	0.048	0.054
20	0.035	0.078	0.080	0.075	0.046	0.052
50	0.029	0.081	0.081	0.074	0.045	0.050

Table 1: Properties of treatment effect estimates from the Across and Within approaches in the simulation where treatment assignment depends only on $\mathbf{x}_1$. The average treatment effect estimate before introduction of missing data is 0.0738. When only control units are missing $\mathbf{x}_2$, the average treatment effect estimates based on only the complete cases is 1.118. When both treated and control units' are missing $\mathbf{x}_2$, the average treatment effect estimate based on only the complete cases is 0.8961.

3.2 Simulation 2: treatment assignment depends only on $\boldsymbol{x}_2$

In this simulation, we assign treatment from Bernoulli distributions where

$$\text{logit}\left(P(T_i = 1|\mathbf{x}_i)\right) = -7.8 + 0.5x_{i2}. \tag{7}$$

As before, this generates approximately 100 treated units and 1000 control units, but now the treatment assignment depends only on $\boldsymbol{x}_2$. Figure 2 displays a typical covariate distribution for this design. We introduce missing values in x_2 using the same two scenarios as in Section 3.1, and impute missing values from a normal linear regression as before. We run the simulation 1000 times, each time using new values of $(\mathbf{X}, \mathbf{T}, \mathbf{Y}, \mathbf{M})$.

Table 2 summarizes the results for different m. Here, $\hat{\tau}^{A,m}$ has substantially smaller bias than $\hat{\tau}^{W,m}$. When both treated and control units are missing $\boldsymbol{x}_2$, the bias in $\hat{\tau}^{A,m}$ tends to decrease as m increases; this is not the case for $\hat{\tau}^{W,m}$. The variance

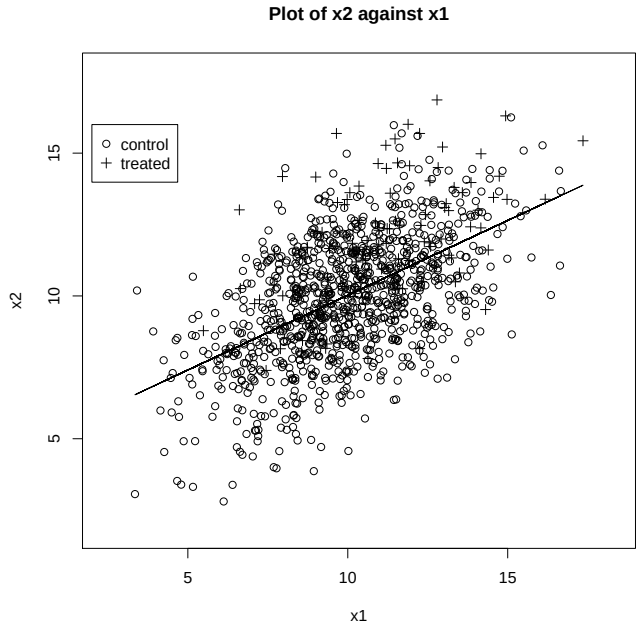

Figure 2: Plot of the covariate distribution in the simulation design where treatment assignment depends on $\boldsymbol{x}_2$ together with the fitted regression line assuming a normal linear model for $\boldsymbol{x}_2$.

of $\hat{\tau}^{W,m}$ continues to be lower than that of $\hat{\tau}^{A,m}$ and to decrease with m, whereas the variance of $\hat{\tau}^{A,m}$ does not decrease with m. In this scenario, the Across approach dominates on mean squared error.

3.3 Simulation 3: treatment assignment depends equally on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$

In this simulation, we assign treatment from Bernoulli distributions where

$$\text{logit}\left(P(T_i = 1|\mathbf{x}_i)\right) = -7.8 + 0.255x_{i1} + 0.255x_{i2}. \tag{8}$$

This generates approximately 100 treated units with treatment assignment depending equally on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Figure 3 displays a typical covariate distribution for this design. We introduce missing values in x_2 values using the same two scenarios as in Section 3.1, and impute missing values from a normal linear regression as before. We run the

	Across			Within		
m	Pt. Est.	Variance	MSE	Pt. Est.	Variance	MSE
Only control units missing $\mathbf{x}_2$						
5	0.565	0.084	0.403	0.825	0.045	0.725
10	0.532	0.088	0.371	0.826	0.041	0.723
15	0.541	0.092	0.385	0.826	0.039	0.721
20	0.538	0.090	0.380	0.826	0.038	0.721
50	0.548	0.100	0.400	0.826	0.036	0.718
Treatment and control units missing $\mathbf{x}_2$						
5	0.311	0.097	0.194	0.840	0.054	0.760
10	0.221	0.094	0.143	0.842	0.045	0.754
15	0.182	0.088	0.121	0.844	0.043	0.755
20	0.174	0.093	0.123	0.845	0.042	0.756
50	0.156	0.096	0.120	0.845	0.039	0.753

Table 2: Properties of treatment effect estimates from the Across and Within approaches in the simulation where treatment assignment depends on $\mathbf{x}_2$. The average treatment effect estimate before introduction of missing data is 0.0614. When only control units are missing $\mathbf{x}_2$, the average treatment effect estimate based on only the complete cases is 0.7653. When both treated and control units' are missing $\mathbf{x}_2$, the average treatment effect estimate based on only the complete cases is 0.5580.

simulation 1000 times, each time using new values of $(\mathbf{X}, \mathbf{T}, \mathbf{Y}, \mathbf{M})$.

Table 3 summarizes the results for different m. Here, $\hat{\tau}^{A,m}$ again has consistently smaller bias than $\hat{\tau}^{W,m}$. The differences between the two point estimators are smaller than observed in Table 2, yet larger than those observed in Table 1. The bias in $\hat{\tau}^{A,m}$ decreases as m increases, whereas the bias in $\hat{\tau}^{W,m}$ does not depend on m. As before, the variance of $\hat{\tau}^{W,m}$ is smaller than the variance of $\hat{\tau}^{A,m}$, and it appears to decrease with m. In this simulation, the Across approach dominates on mean squared error.

3.4 Reasons for differences in point estimates

In terms of bias reduction, both approaches perform similarly when treatment assignment is conditionally independent of the variables with missing data, whereas the Across approach offers greater reductions when assignment is conditionally dependent on the variables with missing data. This suggests that the importance of

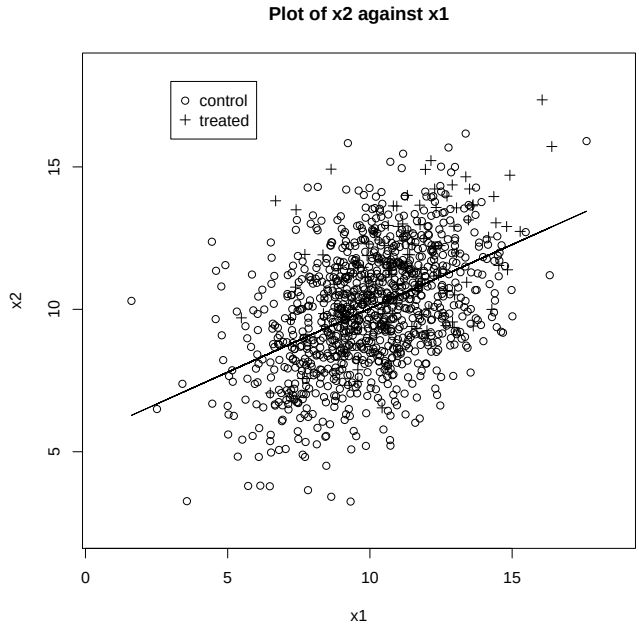

Figure 3: Plot of the covariate distribution in the simulation design where treatment assignment depends equally on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ with the fitted regression line assuming a normal linear model for $\boldsymbol{x}_2$.

variables with missing data in treatment assignment is key to these differences, as we now explain more fully.

When treatment assignment depends only on $\boldsymbol{x}_1$, which is fully observed, the true (not estimated) propensity scores and matched control sets are identical for the Across and Within approaches. However, since matching is based on estimated propensity scores, the coefficient of $\boldsymbol{x}_2$ in the logistic regression is non-zero, so that $\boldsymbol{x}_2$ does play a typically minor role in the matching. Nonetheless, values of $\boldsymbol{x}_1$ remain central for matching even for estimated propensity scores. Hence, the Across and Within estimated propensity scores are generally similar in this scenario, which explains the similar bias reductions.

Interestingly in this scenario, the Across treatment effect actually is slightly closer to $\tau = 0$ than the treatment effect before introducing missing data. For control records with missing x_2, the Across method effectively averages over the distribution of $\boldsymbol{x}_2$ to compute propensity scores, resulting in estimates for those records that effectively

	Across			Within		
m	Pt. Est.	Variance	MSE	Pt. Est.	Variance	MSE
Only control units missing $\mathbf{x}_2$						
5	0.370	0.059	0.196	0.548	0.042	0.343
10	0.338	0.056	0.170	0.551	0.039	0.343
15	0.323	0.053	0.158	0.550	0.038	0.341
20	0.319	0.056	0.158	0.549	0.038	0.339
50	0.316	0.056	0.156	0.550	0.036	0.338
Treatment and control units missing $\mathbf{x}_2$						
5	0.275	0.080	0.155	0.550	0.046	0.349
10	0.236	0.077	0.133	0.551	0.042	0.345
15	0.209	0.079	0.122	0.553	0.040	0.346
20	0.204	0.079	0.120	0.553	0.039	0.345
50	0.196	0.081	0.120	0.551	0.038	0.342

Table 3: Treatment effect estimates from the Across and Within approaches in the simulation design where treatment assignment depends equally on $\mathbf{x}_1$ and $\mathbf{x}_2$. The average treatment effect estimates based on the covariates before introduction of missing data is 0.0467. The average treatment effect estimates based on the complete cases when missing data is only introduced into control units' $\mathbf{x}_2$ value is 0.8917. The average treatment effect estimates based on the complete cases when missing data is introduced into both treatment and control units' $\mathbf{x}_2$ value is 0.6973.

condition only on $\boldsymbol{x}_1$. This closer approximation of the true propensity score model for records with missing $\boldsymbol{x}_2$ explains why the Across estimates have slightly lower bias than the estimates before introducing missing data. We note that this averaging is not a feature of the Within approach.

When treatment assignment depends on $\boldsymbol{x}_2$, as in Simulations 2 and 3, biases increase for both approaches, mainly because now we match based on imputed rather than actual $\boldsymbol{x}_2$. Within any completed dataset, the Within method results in very close balance on $\boldsymbol{x}_1$ and on the completed version of $\boldsymbol{x}_2$ in the treated and matched control set. However, balance on completed $\boldsymbol{x}_2$ does not imply balance on actual $\boldsymbol{x}_2$. In fact, in the Within method, typically the imputed values of $\boldsymbol{x}_2$ for records in the matched control sets were larger than those records' true $\mathbf{x}_2$ values; thus, these records' true values of $\boldsymbol{x}_2$ were smaller than the true values of $\boldsymbol{x}_2$ for the treated records. In the Across method, the distributions of true $\boldsymbol{x}_2$ and completed $\boldsymbol{x}_2$ were

similar; however, both sets of values were typically smaller than the true values of $\boldsymbol{x}_2$ for the treated records. The differences in true $\boldsymbol{x}_2$ values for the treated and matched control sets were smaller in the Across method than in the Within method.

The Across method selects matched controls with missing $\boldsymbol{x}_2$ less frequently than the Within method does, thus mitigating the problems from inaccurate balance on true $\boldsymbol{x}_2$ in missing cases. For example, in Simulation 2, imputed values for $\boldsymbol{x}_2$ appear in the matched control sets typically around 12% more often in the Within method than in the Across method. Thus, it appears that the advantage of the Across method derives from lesser reliance on (inaccurate) imputed values.

In all simulations, the bias in the Within method does not change (beyond simulation error) as m increases. This is because, given the observed data, each treatment effect estimate is independent and identically distributed. For the Across method, however, the bias appears generally to decrease with m, with smaller reductions for larger m. The estimated propensity scores in the Across method approach their complete-data values as m increases, so that the matching is done on estimated scores that are increasingly closer to fixed values.

With regard to trends in variances, in these simulations the variance for the Across method appears not to decrease with m, whereas the variance for the Within method does. The Across method results in only one propensity score (averaged across imputations) for each unit. Increasing m improves precision of the propensity score estimates for cases with missing x_2, but this improvement tends to reduce bias rather than variance. On the other hand, the Within method averages treatment effects over independently generated imputations; hence, as with all means, the variance decreases as m increases.

3.5 Additional Simulations

To investigate if the differences in the Across and Within methods are artefacts of matching without replacement, we repeated the simulations using matching with re-

placement. The results, presented in Appendix 1 of the online supplement, indicate similar bias and variance profiles as matching without replacement for these scenarios. We also considered three other typical uses of propensity scores for estimating treatment effects that were suggested by reviewers: inverse weighting of the propensity score, regression including the propensity score as a covariate, and subclassification of the propensity score. The results are presented in Appendix 2 of the online supplement. Other than inverse weighting, the general trends in Sections 3.1 to 3.3 for bias persist for these other estimation approaches. For inverse weighting, the Within method tends to result in greater bias reductions than the Across method in all scenarios. We also note that, for these additional methods and the values of m considered here, the variance of $\hat{\tau}^{A,m}$ can be smaller than the variance of $\hat{\tau}^{W,m}$.

4 Augmenting the Across method

The Across method can be iterated $r > 1$ times on any dataset, i.e., independently generate m completed datasets r times. One then can estimate τ using

$$\widehat{\tau}^{Amr} = \sum_{l=1}^{r} \widehat{\tau}^{A,m,(l)}/r, \tag{9}$$

where $\widehat{\tau}^{A,m,(l)}$ is the treatment effect estimate from iteration l. In fact, the Within approach can be viewed as an augmented Across approach with $m = 1$ and r equal to the number of imputed datasets, so that $\widehat{\tau}^{W,r} = \widehat{\tau}^{A,1,r}$. Hence, the decision about using Across or Within approaches can be viewed as selecting (m, r). In this section, we illustrate the potential for bias and variance reduction when setting $r > 1$ by repeating the simulations from Section 3 with combinations of (m, r) such that $m \times r = 100$. We generate 1000 simulations, each time using new values of $(\mathbf{X}, \mathbf{T}, \mathbf{Y}, \mathbf{M})$.

Table 4 displays simulated expected values, variances, and MSEs of $\widehat{\tau}^{Amr}$ for the simulation design from Section 3.1. The table also includes expected values of two multiple imputation variance estimators; we defer discussion of these until Section 5.

m	r	Point estimate	Variance	MSE	T^{pool}	T^{pair}
Only control units missing $\mathbf{x}_2$						
1	100	0.077	0.042	0.048	0.332	0.143
2	50	0.064	0.043	0.048	0.330	0.140
5	20	0.060	0.046	0.050	0.329	0.139
10	10	0.057	0.051	0.054	0.330	0.139
20	5	0.052	0.051	0.054	0.331	0.140
50	2	0.057	0.063	0.066	0.338	0.147
100	1	0.060	0.080	0.084	-	-
Treatment and control units missing $\mathbf{x}_2$						
1	100	0.075	0.045	0.051	0.335	0.148
2	50	0.051	0.046	0.049	0.333	0.144
5	20	0.038	0.048	0.050	0.331	0.141
10	10	0.030	0.052	0.053	0.331	0.141
20	5	0.024	0.053	0.053	0.334	0.144
50	2	0.028	0.067	0.068	0.338	0.147
100	1	0.016	0.081	0.081	-	-

Table 4: Treatment effect estimates for different allocations of m and r in the simulation design where treatment assignment depends on $\boldsymbol{x}_1$. Here, T^{pool} and T^{pair} are the two multiple imputation variance estimators described in Section 5.

The effect of increasing m on bias reduction is modest. This is because the bias is already small when $m = 1$, as evident in Table 1. As r increases, as expected the variance of $\widehat{\tau}^{Amr}$ is reduced. The smallest mean squared errors are obtained for ($m = 2, r = 50$), although there is little difference in MSEs up to $m = 5$. The preference for smaller m reflect the greater gains in precision from increasing r compared to the reductions in bias from increasing m.

Table 5 displays results from the simulation design of Section 3.2. Increasing m has a noticeable effect on decreasing bias up until around $m = 20$, after which reductions are modest. The variance continues to decrease as r increases. In these simulations, setting $m \geq 10$ results in the smallest MSEs. The preference for larger m reflects the greater reductions in bias from increasing m compared to the gains in precisions from increasing r. Results from the simulation design of Section 3.3, displayed in Table 6, show similar patterns.

Given that different selections of (m, r) can result in different properties, how

m	r	Point estimate	Variance	MSE	T^{pool}	T^{pair}
Only control units missing $\mathbf{x}_2$						
1	100	0.825	0.035	0.716	0.332	0.199
2	50	0.666	0.039	0.482	0.332	0.183
5	20	0.560	0.046	0.360	0.329	0.167
10	10	0.538	0.057	0.346	0.327	0.162
20	5	0.540	0.067	0.358	0.328	0.161
50	2	0.546	0.087	0.385	0.333	0.165
100	1	0.547	0.105	0.403	-	-
Treatment and control units missing $\mathbf{x}_2$						
1	100	0.843	0.038	0.748	0.339	0.224
2	50	0.553	0.040	0.346	0.331	0.202
5	20	0.311	0.046	0.143	0.323	0.184
10	10	0.221	0.054	0.103	0.319	0.177
20	5	0.174	0.064	0.095	0.320	0.178
50	2	0.155	0.083	0.107	0.326	0.184
100	1	0.141	0.100	0.119	-	-

Table 5: Treatment effect estimates for different allocations of m and r in the simulation design where treatment assignment depends on $\boldsymbol{x}_2$. Here, T^{pool} and T^{pair} are the two multiple imputation variance estimators described in Section 5.

should one determine them? Ideally, one generates large m and large r, so as to gain benefits in bias reduction and precision. However, it may be impractical to make the total number of imputations (mr) large, so that one must select an allocation. In practice, of course, one does not have the luxury of repeated sampling from known population models to aid decision-making. However, taken together, these simulation results suggest a heuristic for selecting (m, r). When the missing data are not important predictors of treatment assignment—which can be assessed from multiple imputation inference—use a small m and large r. When the missing data are important predictors of treatment assignment, use a large m and small r. We illustrate this heuristic in Section 6.

For very large m, the role of r is essentially irrelevant. This is because, by the law of large numbers, the propensity scores approach fixed values as m increases. Hence, for very large m the matched controls obtained will be essentially the same in all sets of m datasets (assuming unique propensity scores), so that $\widehat{\tau}^{A,m,(l)}$ is fixed for any l.

m	r	Point estimate	Variance	MSE	T^{pool}	T^{pair}
Only control units missing $\mathbf{x}_2$						
1	100	0.549	0.036	0.337	0.316	0.107
2	50	0.447	0.033	0.233	0.319	0.093
5	20	0.370	0.034	0.170	0.320	0.082
10	10	0.340	0.036	0.152	0.321	0.078
20	5	0.325	0.038	0.143	0.321	0.077
50	2	0.313	0.046	0.144	0.326	0.080
100	1	0.310	0.055	0.151	-	-
Treatment and control units missing $\mathbf{x}_2$						
1	100	0.551	0.038	0.342	0.320	0.129
2	50	0.396	0.039	0.196	0.320	0.114
5	20	0.275	0.047	0.123	0.318	0.101
10	10	0.228	0.053	0.105	0.316	0.095
20	5	0.201	0.058	0.098	0.317	0.093
50	2	0.195	0.068	0.106	0.321	0.096
100	1	0.189	0.076	0.112	-	-

Table 6: Treatment effect estimates for different allocations of m and r in the simulation design where treatment assignment depends equally on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Here, T^{pool} and T^{pair} are the two multiple imputation variance estimators described in Section 5.

5 Multiple imputation variance estimators

Multiple imputation frameworks are appealing for handling missing data in part because they facilitate accounting for uncertainty due to the presence of missing values. Multiple imputation variance estimators comprise two terms: a complete-data variance ($\bar{u}$ in the notation of Rubin[18]) and a between-imputation variance (b in the notation of Rubin[18]). Unfortunately, it is difficult to estimate components of the multiple imputation variance formula reliably for both the Across and Within methods, as we now document.

Regarding the complete-data variance component, we do not believe that there is one commonly accepted variance estimator for propensity score matching, even when $\mathbf{X}$ has no missing values. Some analysts use conservative two-sample variance estimators[23]; some use matched pairs variance estimators[24,25]; and, others embed treatment effect estimation in regression models based on the matched data[26–28].

Regarding the between-imputation variance estimators, we first consider the Across approach. Let $b_\infty^{(Am)}$ be the between-imputation variance for the Across approach based on m imputed datasets, defined as

$$b_\infty^{(Am)} = \lim_{r\to\infty} b_r^{(Am)} = \lim_{r\to\infty} \sum_{l=1}^{r} (\widehat{\tau}^{A,m,(l)} - \widehat{\tau}^{Amr})^2/(r-1), \tag{10}$$

Obviously when $r = 1$, as in the standard Across approach, it is not possible to construct a method of moments estimator of $b_\infty^{(Am)}$. The augmented Across approach appears to alleviate this problem since $r > 1$. However, $b_r^{(Am)}$ is not guaranteed to exceed zero. For example, as $m \to \infty$, the Across propensity scores converge to fixed points, so that one set of matched controls is used for all iterations r and $b_r^{(A\infty)} = 0$. This is the case for any missing data pattern, not just the one observed. Hence, while $b_r^{(Am)}/r$ can estimate the variability due to imputations given the observed missing data pattern, $b_r^{(Am)}$ itself cannot serve as a valid estimate of the variability over repeated realizations of the missing data pattern.

To investigate the performances of multiple imputation variance estimators for the augmented Across approach, we use the simulations of Section 4. We estimate the complete data variance $\bar{u}$ with the two-sample pooled variance estimator or the matched pairs variance estimator; we denote the resulting multiple imputation variance estimates as T^{pool} and T^{pair}, respectively. Tables 4 - 6 display the averages of T^{pool} and T^{pair} over the 1000 simulated datasets. Since we cannot construct a variance estimator when $r = 1$, we report these results only for $r > 1$. As evident in the tables, both T^{pool} and T^{pair} greatly over-estimate the true variances, although T^{pair} has smaller bias than T^{pool}.

For the Within approach, let $b_\infty^{(W)}$ be the between-imputation variance for the Within approach, defined as

$$b_\infty^{(W)} = \lim_{m\to\infty} b_m^{(W)} = \lim_{m\to\infty} \sum_{l=1}^{m} (\hat{\tau}^{W,m,l} - \widehat{\tau}^{W,m})^2/(m-1), \tag{11}$$

Analysts can estimate $b_{\infty}^{(W)}$ following the usual multiple imputation strategy: take the unbiased estimator $b_m^{(W)}$, i.e., the sample variance of $\bar{Y}_{mc}^{(k)}$. Here it is clear that $b_m^{(W)}/m$ estimates variability due to imputations given the observed missing data pattern. We also suspect that it is a reasonable estimator of the variability over repeated realizations of the missing data mechanism, although to our knowledge this has not been proved mathematically to be a randomization-valid variance estimator. We note, however, that some researchers employ the multiple imputation variance estimator for the Within case[17,29]. Tables 4 - 6 for the Within case ($m = 1, r = 100$) suggest that this variance estimator can have positive bias.

Clearly, developing accurate multiple imputation variance estimators for both the Across and Within approaches is a key area for future research.

6 Empirical Comparison Using Genuine Data

We now apply the Across and Within approaches on data intended to inform analysis of the effects of breast feeding on child's later cognitive development. The data are a subset of the U.S. National Longitudinal Survey of Youth, commonly referred to as the NLSY79. We have analyzed these data previously to illustrate latent class, general location mixture models for multiple imputation of missing covariates. We refer readers to our article[25] on these techniques for more description of the data. Our purpose here is to compare the Across and Within methods; we do not claim that the analyses here represent valid causal inferences.

The response variable is the Peabody individual assessment test math score (PIATM) administered to children at 5 or 6 years of age. The treatment variable is breast feeding duration, which is measured in weeks. We dichotomize this variable into a control condition, < 24 weeks, and a treatment condition, ≥ 24 weeks. The 24 week cutoff corresponds to the number that has been given by the American Academy of Pediatrics[30] and the World Health Organization as a minimum standard for breast feeding duration.

We use the same fourteen covariates as in our previous analyses[25]. These include child's race (Hispanic, black or other), mother's race (Hispanic, black, Asian, white, Hawaiian/Pacific Islander/American Indian, or other), child's sex, two variables indicating whether the spouse or grandparents were present at birth, the number of weeks the child was born premature (zero weeks, one to four weeks, and five or more weeks with cut points determined from guidelines of the March of Dimes), the number of weeks that the mother worked in the year prior to giving birth (not worked at all, worked between 1 and 47 weeks, worked 48-51 weeks, and worked all 52 weeks), number of years between 1979 and the mother's age at the child's birth, mother's intelligence as measured by an armed forces qualification test, mother's highest educational attainment, child's birth weight, the number of weeks that the child spent in hospital, the number of weeks that the mother spent in hospital, and family income. We apply Box-Cox transformations[31] to several variables to facilitate imputation modeling. Three covariates were completely observed in the study, and nine covariates had missing data rates of less than 10%. The two covariates with the largest rates of missing data were family income (22.4%) and the number of weeks that the mother worked in the year prior to giving birth (23.1%).

For this empirical comparison, we begin by creating a fully observed sample: we discard all units with missing values in any covariates, breast feeding duration, or PIATM score. We include youths only if they are first born and are singleton births. The resulting data comprise 1306 youths, of whom 216 are treated. The difference between the sample average PIATM for the 216 treated records and 1090 controls is 5.65.

Among these 1306 cases, several covariates are clearly imbalanced in the treated and control groups. As examples, Figure 4 summarizes the distributions of mother's intelligence score and education for treated and control units, and Table 7 displays the proportion of treated and control units in each level of child's race. Treated units tend to have higher mother's intelligence scores, more mother's years of education and lower proportions of Hispanics and blacks. After matching on estimated propensity

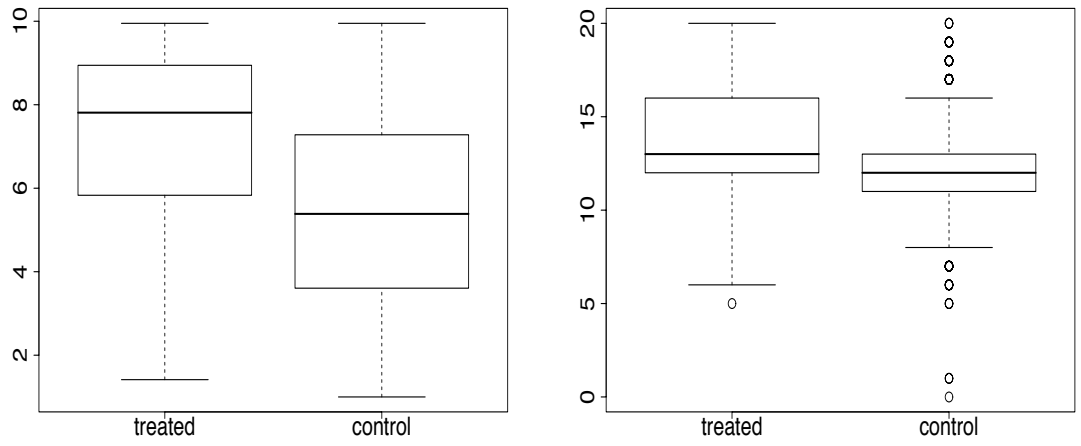

Figure 4: Box plots of mother's intelligence score and mother's years of education respectively for treated and control units before matching.

race	treated	control
Hispanic	0.138	0.190
black	0.111	0.284
other	0.751	0.525

Table 7: Distribution of child's race.

scores, which we computed with a logistic regression of treatment on main effects of all covariates, the estimated treatment effect is 2.32.

To compare the Across and Within methods, we introduce missing values by randomly sampling with replacement from the missing covariate patterns present in the original data. This results in 717 units with fully observed covariates; the remainder have some missing data. For imputation, we use the data augmentation algorithm based on the general location model[32], which is a convenient modeling strategy to handle missing values in mixed categorical and continuous data. We run the model for both treated and control records simultaneously, including an indicator for treatment effect in the imputation model. We observe similar results when imputing the missing values for treated and control units separately. We run the data augmentation algorithm for 200000 iterations after discarding an initial 1000 as burn-in.

m	r	Avg. $\widehat{\tau}^{A,m,r}$	Var($\widehat{\tau}^{A,m,r}$)
1	10000	1.61	0.39
5	5000	1.50	0.37
50	2000	1.46	0.36
100	200	1.48	0.37
500	100	1.45	0.26
1000	20	1.41	0.28
2000	10	1.35	0.15
5000	5	1.53	0.23
10000	1	1.60	NA

Table 8: Treatment effect estimates for different (m, r) combinations. The complete-data treatment effect equals 2.32.

Autocorrelation diagnostics indicate that parameters are approximately uncorrelated after twenty iterations of the algorithm, so that we have potentially 10000 completed datasets to work with for the Across and Within approaches.

We repeat the process of generating missing data patterns and 10000 completed datasets ten times. In each case, we compute $\widehat{\tau}^{A,m,r}$ for various combinations of (m, r) such that $mr = 10000$. Table 8 summarizes the averages and variances of $\widehat{\tau}^{A,m,r}$ across the ten replications. Generally, there is not much difference among the average point estimates across the different combinations; in fact, the average $\widehat{\tau}^{A,m,r}$ are within simulation errors of one another. We note that all of the treatment effect estimates are lower than the complete-data estimate of 2.32.

Why are the results similar for different values of (m, r)? In the simulation studies, the Across and Within methods yield similar results when the missing values are not strongly associated with treatment assignment. This is largely the case for these 1306 records: the variables with the highest fractions of missing data are not that strongly associated with assignment, as indicated by the propensity score regression on the 1306 cases (see Appendix 3 in the online supplement for the results).

We also ran a simulation with a modest value of m, which is often the case in practice. Specifically, we created 100 incomplete data sets by repeatedly drawing from the missing data patterns. For each of these data sets, we ran the data augmentation algorithm to generate 30 multiply imputed datasets. We compute both $\widehat{\tau}^{A,30,1}$ and

$\widehat{\tau}^{A,1,30}$, i.e., the Across and Within methods used in Section 3. Figure 5 displays boxplots of the treatment effect estimates from both methods. Both methods yield similar point estimates on average, but the Within method has smaller variability. These results accord with the findings from the simulations in Section 3.1.

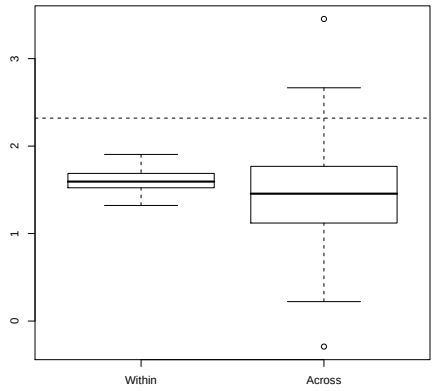

Figure 5: Box plots of the Within and Across treatment effect estimates in the simulation involving NLSY79 data and $mr = 30$. The dotted line represents the treatment effect based on the 1306 complete records.

7 Concluding remarks

In the simulations studied here, the Across approach had the potential for greater bias reduction than the Within approach when treatment assignment depended on the missing covariates. However, the Within approach resulted in smaller variances than the Across approach. Of course, as with any simulation study, these results may have limited generalizability. For some response surfaces, covariate distributions, treatment assignments, or missing data patterns, it may be that one approach always dominates the other. Alternatively, in other settings the two approaches may always give the same answer, for example if data were missing only for control units in a region of covariate space far away from that of the treated units (these units never would be selected as matches). Furthermore, the choice of imputation model also affects treatment effect estimates[25], as might the choice of whether or not to condition on

the response in the imputation models[20]. Thus, we recommend that analysts run simulation studies akin to the one done on the complete cases in the breast-feeding simulation study to get a rough guide of the relative potentials of each procedure for bias reduction. When such studies are not possible, we suggest the augmented Across approach as a default, since it showed the potential for greater bias reductions in the artificial-data simulations. To choose m and r, we suggest following the heuristic described in Sections 4 and 6. Further investigation and development of approaches to select m and r represents an interesting direction for future research.

Acknowledgments

The authors would like to thank Professor Jennifer Hill who provided us with the data from the breast feeding study analyzed in Section 6. This research was supported by the National Science Foundation [NSF-ITR-0427889].

References

[1] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.

[2] Rosenbaum PR, Rubin DB. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. The American Statistician. 1985;39(1):33–38.

[3] D'Agostino Jr RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Statistics in Medicine. 1998;17:2265–2281.

[4] Park GS, Wong WK, Oh M, Khanna D, Gold RH, Sharp JT, et al. Classifying Radiographic Progression Status in Early Rheumatoid Arthritis Patients Using

Propensity Scores to Adjust for Baseline Differences. Statistical Methods in Medical Research. 2007;16(1):13–29.

[5] Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: From naive enthusiasm to intuitive understanding; 2011. Statistical Methods in Medical Research (online early).

[6] Rosenbaum PR, Rubin DB. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. Journal of the American Statistical Association. 1984;79:516–524.

[7] Hullsiek KH, Louis TA. Propensity Score Modeling Strategies for the Causal Analysis of Observational Data. Biostatistics (Oxford). 2002;3(2):179–193.

[8] Rosenbaum PR. A characterization of optimal designs for observational studies. Journal of the Royal Statistical Society, Series B-Methodological. 1991;53:597–610.

[9] Stuart EA, Green KM. Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. Developmental Psychology. 2008;44(2):395–406.

[10] Lunceford JK, Davidian M. Stratification and Weighting Via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study. Statistics in Medicine. 2004;23(19):2937–2960.

[11] Hanley JA, Dendukuri N. Efficient Sampling Approaches to Address Confounding in Database Studies. Statistical Methods in Medical Research. 2009;18(1):81–105.

[12] Woo MJ, Reiter JP, Karr AF. Estimation of Propensity Scores Using Generalized Additive Models. Statistics in Medicine. 2008;27(19):3805–3816.

[13] Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. Journal of Clinical Epidemiology. 2010;63:826–833.

[14] 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. 2008;17:546–555.

[15] D'Agostino Jr RB, Rubin DB. Estimating and using propensity scores with partially missing data. Journal of the American Statistical Association. 2000;95(451):749–759.

[16] Haviland A, Nagin DS, Rosenbaum PR. Combining Propensity Score Matching and Group-Based Trajectory Analysis in an Observational Study. Psychological Methods. 2007;12(3):247–267.

[17] Qu Y, Lipkovich I. Propensity Score Estimation with Missing Values Using a Multiple Imputation Missingness Pattern (MIMP) Approach. Statistics in Medicine. 2009;28(9):1402–1414.

[18] Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley; 1987.

[19] Reiter JP, Raghunathan TE. The multiple adaptations of multiple imputation. Journal of the American Statistical Association. 2007;102:1462–1471.

[20] Hill J. Reducing bias in treatment effect estimation in observational studies suffering from missing data. Columbia University Institute for Social and Economic Research and Policy (ISERP). 2004;Working paper 04-01.

[21] Little RJA. Regression with Missing X's: A Review. Journal of the American Statistical Association. 1992;87:1227–1237.

[22] Moons KGM, Donders RART, Stijnen T, Harrell Jr FE. Using the outcome for imputation of missing predictor values was preferred. Journal of Clinical Epidemiology. 2006;59(10):1092–1101.

[23] Lechner M. Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification. Journal of Business and Economic Statistics. 1999;17(1):74–90.

[24] Austin PC. Type I Error Rates, Coverage of Confidence Intervals, and Variance Estimation in Propensity-score Matched Analyses. The International Journal of Biostatistics. 2009;5(1).

[25] Mitra R, Reiter JP. Estimating propensity scores with missing covariate data using general location mixture models. Statistics in Medicine. 2011;30(6):627–641.

[26] Hill J, Reiter JP. Interval estimation for treatment effects using propensity score matching. Statistics in Medicine. 2006;25:2230–2256.

[27] Rubin DB, Thomas N. Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates. Journal of the American Statistical Association. 2000;95(450):pp. 573–585.

[28] Hill JL, Reiter JP, Zanutto EL. A comparison of experimental and observational data analyses. In: Gelman A, Meng XL, editors. Applied Bayesian Modeling and Causal Inference from an Incomplete-Data Perspective. Wiley; 2004. .

[29] Alecke B, Mitze T, Reinkowski J, Unitedt G. Does Firm Size make a Difference? Analysing the Effectiveness of R&D Subsidies in East Germany. German Economic Review. 2011;.

[30] Chantry CJ, Howard CR, Auinger P. Full Breastfeeding Duration and Associated Decrease in Respiratory Tract Infection in US Children. Pediatrics. 2006;117(2):425–432.

[31] Box GEP, Cox DR. An Analysis of Transformations. Journal of the Royal Statistical Society Series B (Methodological). 1964;26(2):211–252.

[32] Schafer JL. Analysis of Incomplete Multivariate Data. London: Chapman & Hall; 1997.