Letter to the editor

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2016, Vol. 25(6) 3066–3068

DOI: 10.1177/0962280216674296
smm.sagepub.com
SAGE

Comments on propensity score matching following multiple imputation

BBL Penning de Vries and RHH Groenwold

In a recently published simulation study, Mitra and Reiter[1] compared two approaches to implementing propensity score (PS) methods following multiple imputation. Particular emphasis was on PS matching following multiple imputation. In simulation studies, they evaluated two possible approaches, i.e. the so-called Within and Across approaches. In both approaches, PSs are estimated in each of m imputed datasets. In the Within approach, PS matching is performed within each imputed dataset. The resulting m effect estimates are then pooled by averaging. In the Across approach, for each subject the m estimated PSs are averaged first, after which PS matching is performed once, based on each subject's average PS. Apparent from the results was the trend that although both approaches were biased, the Within method was generally more biased than the Across approach, particularly when there was missing confounder data.

We argue that these findings are due to the imputation model and the matching algorithm rather than a genuine difference between the methods. While Mitra and Reiter chose to leave the outcome out of the imputation model, it has been shown that often the outcome should actually be included in the imputation model.[2] To illustrate this, we repeated a selection of Mitra and Reiter's simulations, which represent a setting of a binary treatment, a continuous outcome, and two normally distributed covariates. Here, we focus on the scenarios in which both covariates acted as confounders and both treated and untreated subjects were assigned missing covariate values. Results are presented in Table 1.

In line with Mitra and Reiter, when we applied PS matching while leaving the outcome variable out of the imputation model, the Across approach outperformed the Within approach in terms of bias (Table 1, scenario 1(a)). With the outcome included in the imputation model (scenario 1(b)), the Within estimates still deviate more from the true treatment effect than the Across estimates, but closely approximate the mean estimate based on PS matching before the introduction of missing values (0.053, 95% CI 0.043; 0.064). The bias observed in the absence of missing data is largely due to non-positivity in the tails of the PS distributions of treated and untreated subjects. As a result, treated subjects in the upper tail of the PS distribution are matched to untreated subjects who tend to have lower PS values, thus leading to suboptimal balance in PS between treated and untreated subjects. This balance can, however, be improved, e.g. by using narrow callipers for matching or increasing the sample size n and thus increasing the number of potential matches. With n increased

Julius Center for Health Sciences and Primary Care, The Netherlands

Corresponding author:
BBL Penning de Vries, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, The Netherlands.
Email: b.b.l.penningdevries@umcutrecht.nl

Table 1. Properties of the Across and Within estimators.

	Across			Within		
	Pt. Est.	Variance	MSE	Pt. Est.	Variance	MSE
Scenario 1: Matching						
a	0.282	0.081	0.161	0.557	0.051	0.361
b	−0.012	0.032	0.032	0.060	0.024	0.027
c	−0.048	0.003	0.005	0.025	0.002	0.003
Scenario 2: Regression						
a	0.166	0.039	0.066	0.438	0.032	0.224
b	−0.077	0.020	0.026	0.002	0.018	0.018
Scenario 3: IPTW						
a	0.092	0.002	0.010	0.043	0.001	0.003
b	−0.701	0.805	1.296	0.227	0.616	0.668
c	0.011	0.001	0.002	−0.002	0.001	0.001
d	−0.236	0.664	0.720	−0.022	0.658	0.658
e	0.901	0.009	82.793	0.888	0.009	83.030
f	9.638	2.008	2.139	9.967	2.065	2.066

Pt. Est.: mean effect estimate across 1000 simulations; variance: empirical variance; MSE: mean squared error.
Note: Sample size $n = 1100$, except for scenario 1c where $n = 11{,}000$. In scenario 1, effect estimates were based on PS matching following multiple imputation with outcome left out (a) or included (b, c) in the imputation model. In scenario 2 (a, b), treatment effects were estimated using linear regression, regressing the outcome on treatment, the PS, and both covariates. In 2(a), the outcome was left out of the imputation model, whereas in 2(b) it was included. In scenario 3, effect estimates were based on IPTW, following Mitra and Reiter[1] (a, c, e), or using the traditional weights (see text) (b, d, f). In scenarios 3(a) and (b), the outcome variable was not included in the imputation model, whereas in scenarios 3(c), (d), (e) and (f), the outcome was included in the imputation model. In all scenarios, the true effect of treatment on the outcome was zero, except for scenarios 3(e) and (f), in which it was 10.

by a factor of 10 (Table 1, 1(c)) it becomes apparent that the Within approach is superior to the Across provided the outcome variable is included in the imputation model.

Mitra and Reiter also assessed multiple imputation followed by regression adjustment (i.e. including the confounders as covariates in a linear regression model). In this situation, we observed the same trend across the different imputation models (Table 1, scenarios 2(a) and 2(b)). Again, upon inclusion of the outcome variable in the imputation model, the Within approach yields unbiased estimates, while the Across approach does not.

A third method of controlling for confounding that was studied by Mitra and Reiter was inverse probability weighting. In scenario 3(a), we estimated the true effect using inverse probability of treatment weighting (IPTW) where, following Mitra and Reiter,[1] the weight for any subject equalled 1 if a subject was treated, and PS/(1-PS) if untreated. The treatment effect is then estimated by the difference between the sum of the weighted outcomes in the treatment group and the sum of the weighted outcomes in the control group, divided by the original sample size n. In scenario 3(b), we used the traditional weights discussed by, e.g. Lunceford and Davidian[3] and Robins et al.[4]; i.e. 1/PS if a subject was treated, and 1/(1-PS) otherwise. Note that these are equivalent to those of scenario 3(a) multiplied by PS, meaning that the average of weighted outcomes based on the weights used by Mitra and Reiter is necessarily closer to zero (since PS <1), than if the traditional weights were used. Again, we observed that with the outcome variable included in the imputation model, the Within method is superior to the Across (Table 1, scenarios 3(c) and (d)). Further, in scenarios of a non-null treatment effect (Table 1, 3(e) and (f), true

effect = 10) simulations suggest that the traditional weights are to be preferred – i.e. unless the interest lies in estimating the average effect on the treated,[5] in which case the denominator of the effect estimator should match the effective size of the groups in the pseudopopulation.

In medical research, confounding and missing data are common problems that often occur simultaneously. When multiple imputation is to be followed by PS matching, researchers could apply the Across and the Within approaches that were proposed by Mitra and Reiter. Provided the correct imputation model is applied and there are no other sources of bias (e.g. model misspecification), the Within approach appears to be superior to the Across approach in terms of bias reduction.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: RHHG was funded by the Netherlands Organization for Scientific Research (NWO-Vidi project 917.16.430). The views expressed in this article are those of the authors and not necessarily any funding body.

References

1. Mitra R and Reiter JP. A comparison of two methods of estimating propensity scores after multiple imputation. *Stat Meth Med Res* 2016; **25**: 188–204.
2. van Buuren S. *Flexible imputation of missing data*. Boca Raton, FL: CRC Press, 2012.
3. Lunceford JK and Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med* 2004; **23**: 2937–2960.
4. Robins JM, Hernán MA and Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**: 550–560.
5. Morgan SL and Todd JJ. A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Socio Meth* 2008; **38**: 231–281.