Propensity Score Analysis with Missing Data: The Comparison of Multiple Imputation Approaches

Eun Sook Kim[1], Jeffrey D. Kromrey[1], Seang-Hwane Joo[1]
Yan Wang[1], Jessica Montgomery[1], Reginald Lee[1]
Patricia Rodriguez de Gil[1], Shetay Ashford[1]
Rheta Lanehart[1], Chunhua Cao[1]
[1]University of South Florida, 4202 E. Fowler Avenue, Tampa, FL 33620

Abstract

The appropriate treatment of missing data under different missing data mechanisms is essential for unbiased estimates and correct statistical inferences in propensity score analysis (PSA). This simulation study investigates the efficacy of multiple imputation approaches to missing data in PSA. Four different approaches are considered in combination of two factors: what to impute (covariates only or PS in concert with covariates) and how to combine multiply imputed data (average treatment effects or average PS). Simulation design factors include sample size (500, 1000), treatment effect magnitude (0, .05, .10, .15), correlation between covariates (0, .50), proportion of missing observations (.20, .40, .60), proportion of missing covariates (.20, .40, .60), the number of covariates (15, 30), and missing data mechanisms (MCAR, MAR, MNAR). The missing data treatments serve as a within group factor. Imputing covariates only, combined with averaging treatment effect estimates across imputations, outperforms other methods under MAR, but none of multiple imputation approaches is apt under MNAR.

Key Words: propensity score analysis, missing data, multiple imputation, simulation

1. Introduction

Missing data are a ubiquitous problem in social science research. As most common statistical procedures assume complete data, failing to address the presence of missing values can lead to a multitude of issues, ranging from decreased power to heightened bias and inaccurate Type I error control. Among the several methods developed to deal with missing data problems, Multiple Imputation (MI) has become one of the most accepted. Provided the assumptions are met, MI has shown excellent qualities (see, for example, reviews by Graham, Cumsille, & Elek-Fisk, 2003; Graham & Hoffer, 2000; Shafer & Olsen, 1998).

However, multiple imputation, when considered in the context of propensity score analysis, requires several decisions on the part of the researcher. Since random assignment is not possible in an observational study, we must account for the process by which individuals select their group (either treatment or control). Estimating and incorporating propensity scores is one method that allows one to control for this selection process. Rosenbaum and Rubin (1983) defined the propensity score as “the conditional probability of assignment to a particular treatment given a vector of observed covariates” (p. 41). An individual’s propensity score can be calculated as follows, where Pr(Z=1) is the probability of assignment to the treatment group and p is the number of covariates in vector x:

$$\text{logit}\,(\mathrm{Z}=1)=\log\left[\frac{\hat{\pi}}{1-\hat{\pi}}\right]=\ \beta_0+\sum_{i=1}^{p}\beta_i x_i$$

Consider a situation where there are missing data on one or more of the covariates included in vector x. These covariates are needed in order to calculate the propensity scores that will be required to account for group selection and make unbiased inferences. It is possible for researchers to estimate propensity scores for those individuals with complete data and multiply impute the remaining propensity scores or conversely, one could opt to impute the missing covariates themselves and use these new datasets to subsequently estimate the propensity scores. Thus, the first decision researchers must make in this context is which missing values to impute, the covariates alone or the propensity score in concert with the covariates.

Regardless of which decision is made the researcher will now have multiple datasets each containing different propensity score values for those cases that included missing data which can now be used to estimate the treatment effect. The typical approach is to use each imputed data set to estimate separate treatment effects which would then be combined. Alternatively, the researcher can opt to average the propensity scores for individuals with more than one value and use that average in estimating the treatment effect (Hill, 2004).

Examining all possible combinations of the two choices presented above leads to four distinct approaches for utilizing multiply imputed data in an observational context: (a) impute only covariates, conduct separate analyses with imputed datasets, and combine treatment effect estimates across imputations – Cov Only (MI), (b) impute only covariates, estimate and average propensity scores across imputations, and estimate a treatment effect using single mean propensity scores – Cov Only (Avg), (c) impute propensity scores along with covariates, conduct separate analyses with imputed datasets, and combine treatment effect estimates across imputations – Cov PS (MI), and (d) impute propensity scores along with covariates, average propensity scores across imputations, and estimate a treatment effect using single mean propensity scores – Cov PS (Avg). This study uses simulation to assess the efficacy of these techniques. Listwise deletion of cases with missing values was also included in the comparison because this method is the common default in software packages.

Missing data problems become even more complex when we consider variation in the missing data mechanism. While most statistical analyses assume data are missing at random (MAR), this is often not an assumption that can be tested using the present dataset. In order to determine how robust our methodological choices are under different mechanisms for missingness, the five techniques above were conducted using data that were missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

2. Theoretical Framework

2.1 Rubin’s Missing Data Taxonomy

When applying a missing data treatment, e.g., MI, the process or mechanism that causes missingness might need to be incorporated into the statistical model because incorporating the missing data mechanism leads to efficient data modeling for fitting the data adequately. In other words, the missing data treatment should reflect the underlying uncertainty of the process or mechanism that best explains why data are missing (Rubin, 2005). Rubin's (1976) theoretical framework is the most widely accepted for missing data research.

Let $Y = (Y_1, Y_2, \ldots, Y_p)$ be a random vector variable for p-dimensional multivariate data (y x p matrix), and θ and δ are the parameter of the data and the parameter of the conditional probability of the observed pattern of missing data, respectively. In the presence of missing data, φ, inferences about θ are conditional on δ, the mechanism that causes missingness. That is, given the data at hand, the purpose of the missing data process is to allow making valid inferences about θ but these inferences depend upon or are conditional on δ, the mechanism that generated missing data.

Missing Completely at Random (MCAR). When the probability of a missing value on a variable is independent of the observed or unobserved value for any variable, that is, the probability that y_i is missing is independent of the probability of missingness for any y_{j}, data are missing completely at random or MCAR.

Missing at Random (MAR). When the probability that y_{i} is missing for variable Y is dependent on the data of any other variables but not on the variable Y of interest, data are missing at random or MAR. In addition, MAR requires data on a variable to be missing randomly within subgroups (Roth, 1994).

Missing NOT at Random (MNAR). When the probability of a missing value depends on unobserved data or data that could have been observed, data are missing not at random or MNAR. That is, the value of y_i depends on the value of Y_i. Under this scenario, why data are missing is not ignorable, thus requiring that the missing data mechanism is modeled to make valid inferences about the model parameter.

2.2 Propensity Score Analysis with Observational Data

Randomized trials are the "gold standard procedure" for evaluating the effectiveness of treatment effects; under random assignment to T (t_1 = treatment, t_0 = control), x_i is the vector of baseline covariates (i.e., pretreatment measurements, before treatment is assigned) for which the i observations or units are likely to be similar. Thus, in a randomized experiment, every observation has the same probability of assignment to either t_1 or t_0 independently of x (strong ignorability of assignment assumption) and the average treatment effect is directly estimated as,

$$\tau = E(y_1) - E(y_0),$$

where E means the expectation in the population.

When an experimental study (randomized trial) cannot be conducted because of, for example, ethical issues (e.g., assignment of observations to life-threatening conditions), nonrandomized or observational studies provide the necessary data from which treatment effect inferences can be made. Because in these nonrandomized studies the treatment group (t_1) and the control group (t_0) can differ systematically in the baseline covariates due to selection bias, researchers can apply methods that ensure that selection bias is corrected; otherwise, biased estimates of the treatment effects might result. One of these methods to handle the differences between treatment group and control group in an observational study is propensity score (PS) analysis.

Rosenbaum and Rubin (1983) defined the propensity score as “the conditional probability of assignment to a particular treatment given a vector of observed covariates” x (p. 41). That is, the PS allows the conditional assignment of units to treatment and control groups given x. In practice, it is not possible for any observation or unit to receive both treatments; either y_{i1} or y_{i0} is observed and a unique response y_{it} is expected (stable unit-treatment value assumption; Rubin, 2005). The PS is an important application in observational studies because it adjusts for confounding variables which are a source for potential bias in treatment effect estimates.

2.3 Multiple Imputation

Multiple imputation (MI) creates m imputed data sets for an incomplete p-dimensional multivariate data. That is, missing values are replaced with a set of plausible values that represent a random sample and thus representing the uncertainty about the missing value.

3. Method

This study examines the efficacy of four multiple imputation techniques utilized to address missing data in propensity score analyses. Simulated data are used to empirically assess each method in terms of bias and variability in parameter estimates, Type I error rates, and statistical power.

Data were generated using PROC IML in SAS 9.4 (SAS Institute, 2014) with values for explanatory variables being drawn from normal distributions. Several factors were manipulated: sample size (500, 1000), treatment effect magnitude (0, .05, .10, .15), correlation between covariates (0, .5), proportion of missing observations (.20, .40, .60), proportion of missing covariates (.20, .40, .60), the number of covariates (15, 30), and missing data mechanisms (MCAR, MAR, MNAR). The missing data treatments serve as a within group factor. In addition, the samples were analyzed before missing data were imposed to provide a reference condition for the evaluation of MI effectiveness.

4. Results

4.1 Missing Completely At Random (MCAR)

4.1.1. Statistical bias

As illustrated in Figure 1, although the mean biases for Cov Only (Avg), Cov PS (MI), and Cov PS (Avg) were not substantial, great variability existed in the bias distribution especially for the Cov Only (Avg) and Cov PS (Avg) approaches. In addition, the Cov Only (Avg) and Cov PS (Avg) approaches provided underestimation of the treatment effect, while the Cov PS (MI) tended to slightly overestimate the treatment effect. Bias was relatively small for the Cov Only (MI) and listwise deletion as well as the complete data conditions, although for listwise deletion, there were cases in which the treatment effect was severely overestimated. The number of covariates and covariate intercorrelation had significant impact on the bias estimates (Figure 2). With a larger number of covariates and correlated covariates, the statistical bias increased.

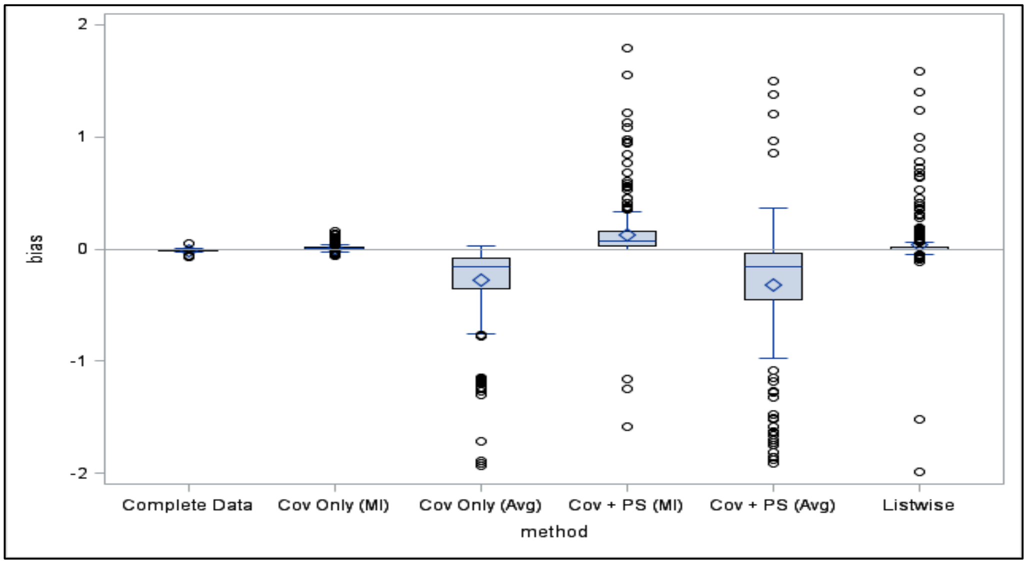

Figure 1: Distributions of bias by method across all MCAR conditions

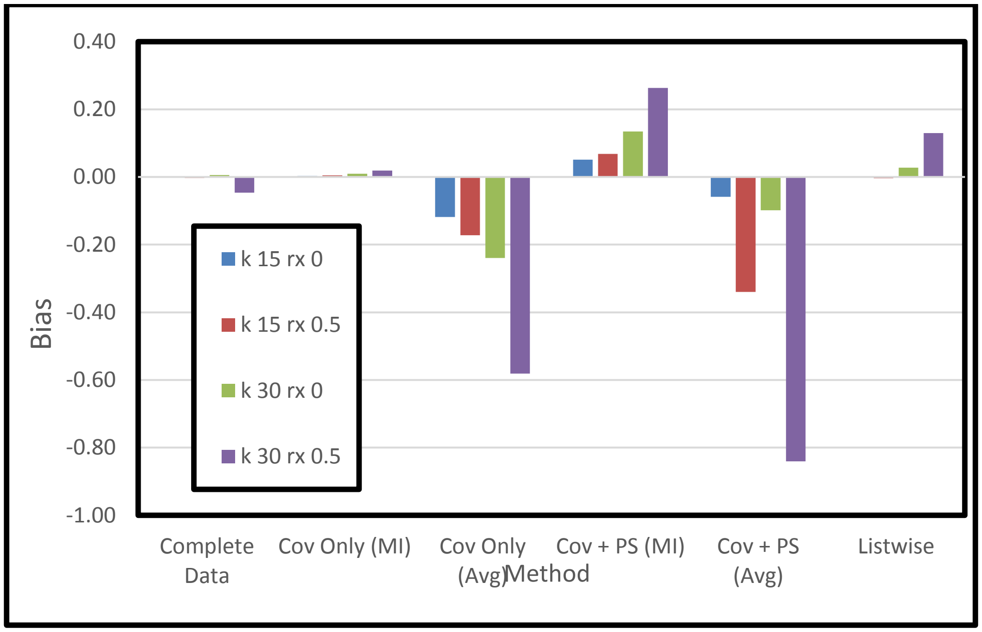

Figure 2: Mean bias by number of covariates and covariate intercorrelation for MCAR

4.1.2 RMSE

Substantial variability in RMSE is evident for all missing data treatment approaches as well as for the samples before missingness was imposed. In comparing these distributions, the MI approach with covariates only produced only slightly larger RMSE values than those obtained with complete data, while the other three approaches to MI yielded notably larger RMSE values. The magnitude of RMSE varied as a function of the number of covariates and the correlation between the covariates.

4.1.3 Confidence interval coverage

The overall distributions of 95% confidence interval (CI) coverage by missing data methods under MCAR are presented in Figure 3. The MI approach with covariates only surpassed the other MI methods in terms of CI coverage. For this outperforming method, there is no noticeable difference from the CI coverage under the complete data conditions. The listwise deletion and Cov PS (MI) also showed reasonable CI coverage around 95% except outlying cases. However, when the treatment effect was estimated with the average propensity scores, the CI coverage was notably below .95 with large variability.

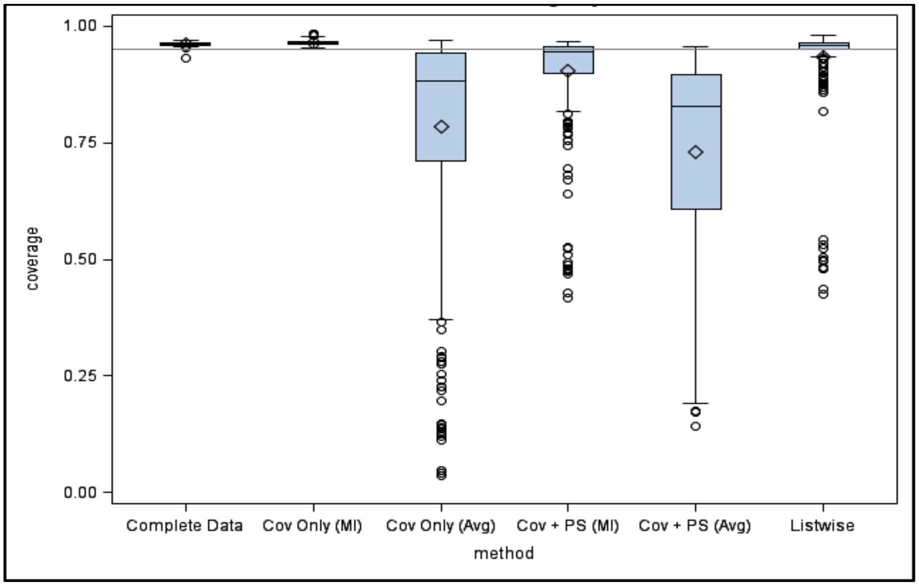

Figure 3: Distributions of CI coverage by Method across all MCAR conditions

4.1.4 Confidence interval width

The overall distributions of confidence interval width were not substantially different across missing data methods and very similar to that of the complete data. The major design factors related to the variability of CI width include the number of covariates, covariate intercorrelation, and the interaction between them.

4.1.5 Type I error control

The MI approach with covariates only, listwise deletion, and complete data evidenced Type I error rates that were adequately controlled, while the other three approaches yielded notably larger Type I error estimates (see Figure 4).

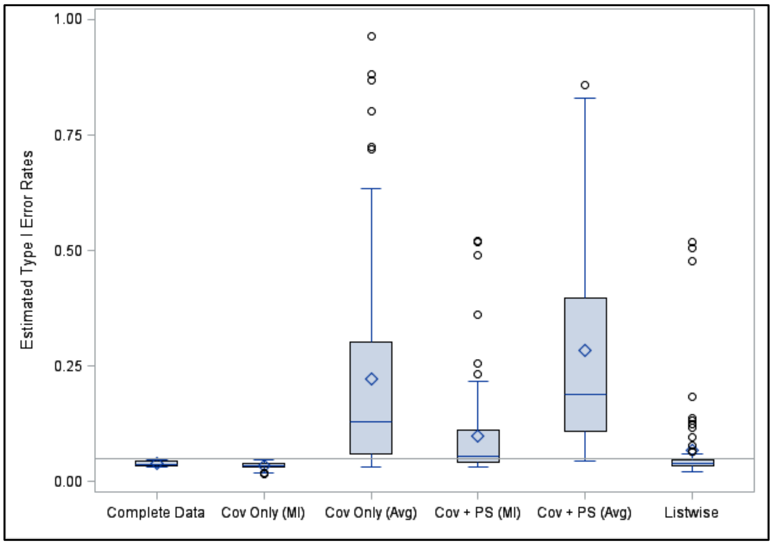

Figure 4: Distributions of Type I error rates by Method across all MCAR conditions

4.1.6 Statistical power

Because statistical power should only be considered after adequate Type I error control has been established, power was estimated only for complete data, listwise deletion, and MI with covariates only. The complete data condition evidenced the greatest power, but the MI treatment provided power that was nearly as large on average. The listwise deletion approach provided notably lower power.

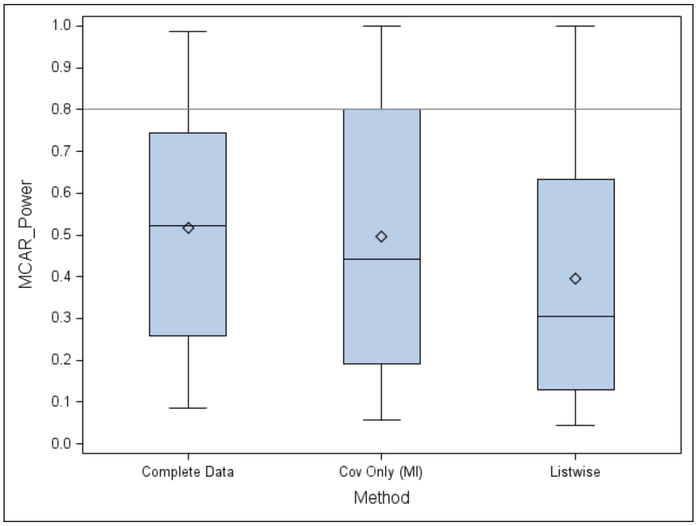

Figure 5: Distributions of power by Method across MCAR conditions

4.2 Missing At Random (MAR)

4.2.1 Statistical bias

The results under MAR were generally comparable to those under MCAR in terms of statistical bias. With complete data, bias was small in all conditions. Among the five approaches for missing data treatment, the MI with covariates only and Listwise deletion produced relatively small bias values.

4.2.2 RMSE/CI coverage/ CI width

The results of RMSE, CI coverage, and CI width under MAR were very similar to those under MCAR and are not repeated here.

4.2.3 Type I Error Control

The overall distributions of Type I error estimates for the MAR conditions are presented in Figure 6. In comparing these distributions, the MI approach with covariates only controlled Type I error rates nearly as well as the complete data conditions. Listwise deletion controlled Type I errors adequately, although it had slightly larger mean Type I error rate.

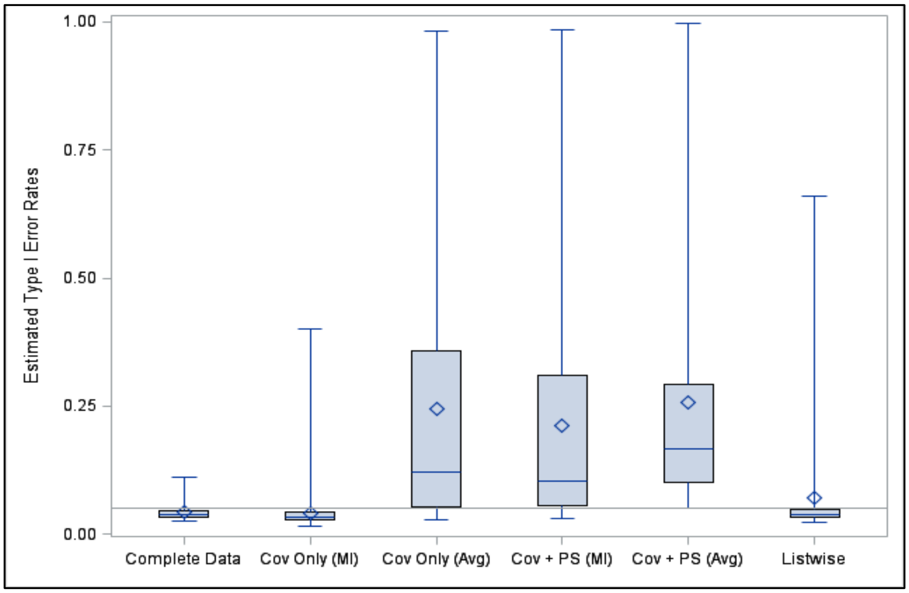

Figure 6: Overall distributions of Type I error rates by Method across all MAR conditions

The number of covariates and proportion of missing covariates (Figure 7) had significant impact on the Type I error control of most approaches. In general, Type I error rates increased with more missing data and with more covariates. Across levels of these factors, however, the use of MI with the covariates only provided the best Type I error control and listwise deletion of cases provided adequate control unless the number of covariates was large.

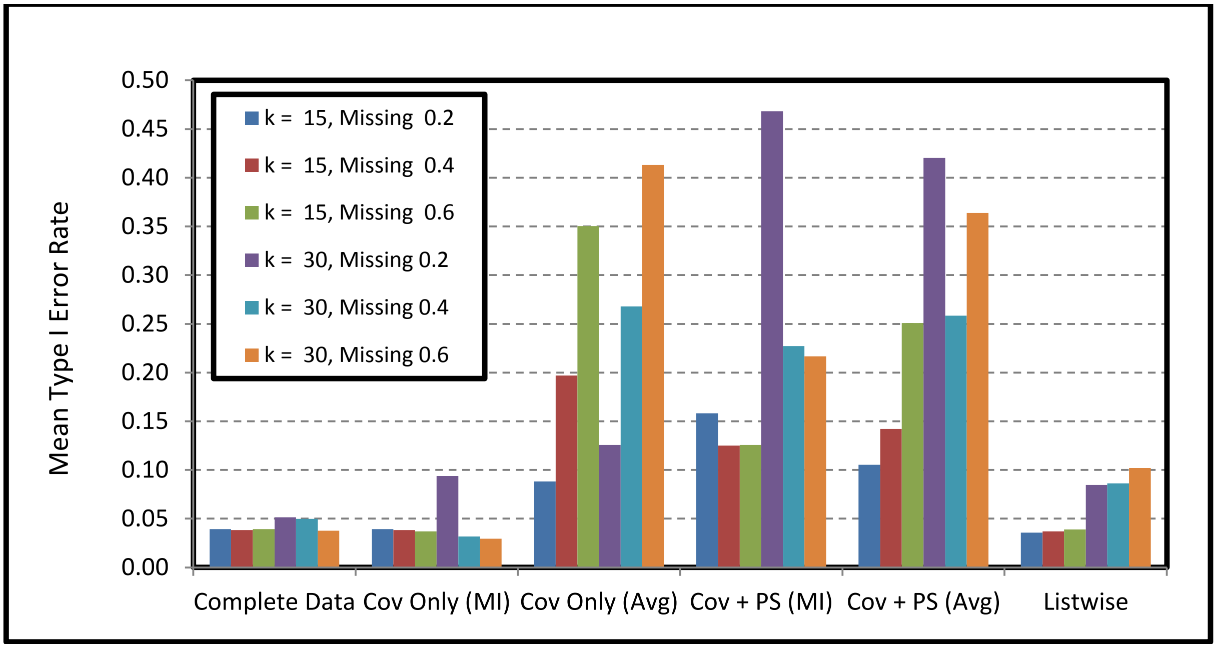

Figure 7: Mean Type I error rate by number of covariates and covariate intercorrelation under MAR

4.2.4 Statistical power

The overall distributions of statistical power for the three methods that provided the best Type I error control for the MAR conditions are evaluated. Power varied across methods but as expected, the complete data method had the higher overall mean power. The power for the MI approach with the covariates provided only slightly lower power, but the power of the listwise deletion approach was notably lower.

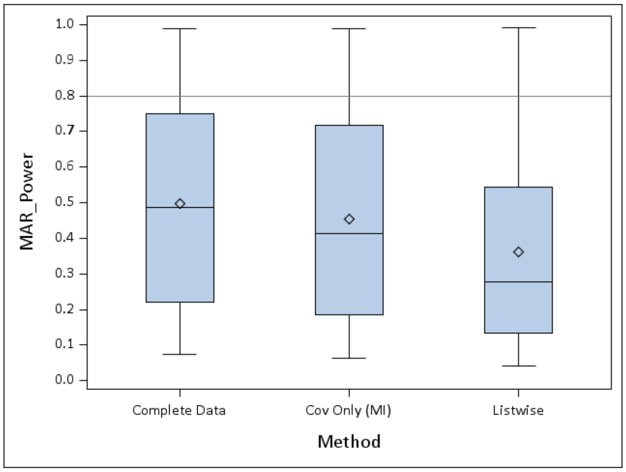

Figure 8: Distributions of power estimates across MAR conditions

4.3 Missing Not At Random (MNAR)

4.3.1 Statistical bias

As illustrated in Figure 9, all of the imputation methods evidenced substantial positive bias when the data were MNAR, with the greatest bias seen when both the covariates and the propensity score were imputed. In contrast, the listwise deletion approach showed an average bias near zero, although the range of bias values (both overestimating and underestimating the effect) was substantial.

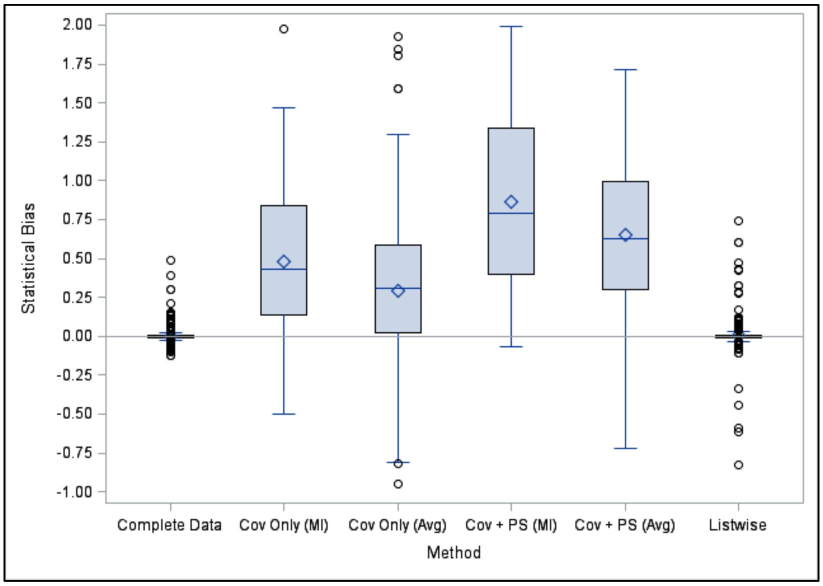

Figure 9: Distributions of bias by method across all MNAR conditions

Both the proportion of missing covariates and the correlation between covariates were substantially related to the resulting bias in the estimate of the treatment effect realized by the imputation approaches (Figure 10). As expected, the bias tended to increase with greater proportions of missing data and with correlated covariates.

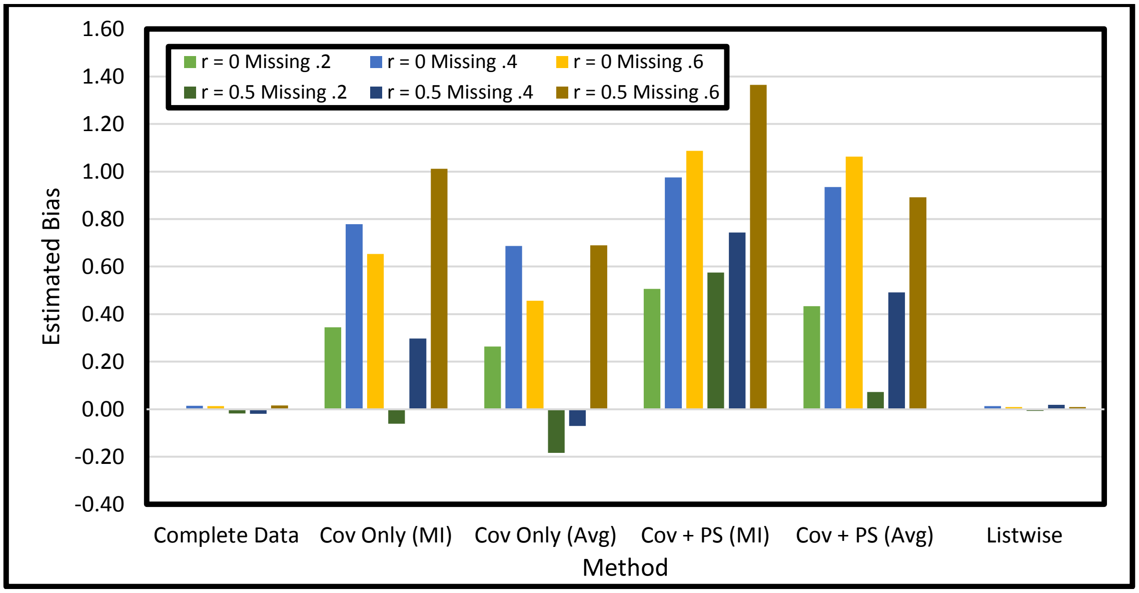

Figure 10: Mean bias by proportion of missing covariates and covariate intercorrelation

for MNAR

4.3.2 RMSE

The results suggested that RMSE estimates for listwise deletion were comparable to those with complete data. MI with covariates only showed larger RMSE, but this approach provided smaller RMSE values than the other imputation strategies.

4.3.3 Confidence interval coverage

The average confidence interval coverage across all conditions was unacceptably low with considerable variability when multiple imputation was implemented under MNAR regardless of imputation models as shown in Figure 11. On the other hand, for listwise deletion the 95% CI coverage was about 95%, which was very comparable to that of the complete data conditions except for slightly larger variability.

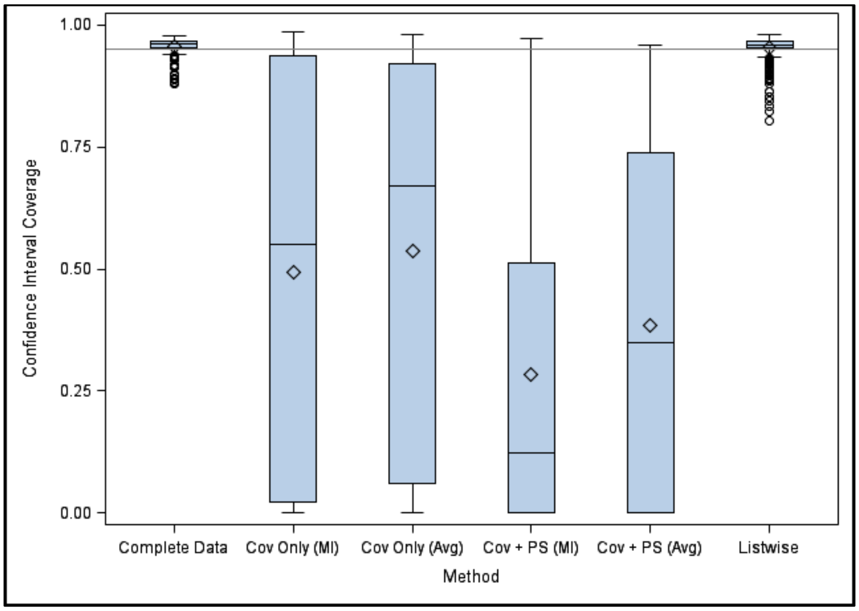

Figure 11: Distributions of CI coverage by Method across all MNAR Conditions

4.3.4 Confidence interval width

Similar to the results of MCAR and MAR, generally there was no salient difference in the CI width across missing data methods including the complete data conditions. However, the dispersion of the CI width across conditions was notably larger for the listwise deletion possibly due to the loss of observations and subsequently larger standard errors.

4.3.5 Type I error control and power

None of the imputation methods provided adequate Type I error control in the majority of conditions. However, the listwise deletion approach controlled Type I error probabilities nearly as well as the complete data conditions. It should be noted that (as expected) the statistical power of listwise deletion was notably smaller than that obtained for the complete data conditions and the power differential became greater as the effect size increased.

5. Conclusions

The results of this study will be of practical significance to researchers working with observational data. In such situations missing data are a frequent occurrence and the results of this study can inform practice by determining which methods perform best. The analysis of differential effectiveness by the simulation design factors will help to assess the sensitivity of results under research design choices and missing data mechanisms.

Overall, the results of this study indicate the importance of selecting a missing data treatment with care. For imputation, the use of MI with the covariates only, followed by a separate estimate of the treatment effect within each imputed data set, is clearly the preferable strategy. This method provided the smallest bias and the best CI coverage for MCAR and MAR missing data mechanisms. With the MNAR conditions, however, none of the imputation methods were effective. For these conditions, the listwise deletion approach provided notably better estimates than any of the imputation methods.

References

Graham, J. W., Cumsille, P. E. & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.), *Research Methods in Psychology* (pp. 87-114). New York: Wiley.

Graham, J. W., & Hofer, S. M. (2000). Multiple imputation in multivariate research. In T. D. Little, K. U. Schanabel, & J. Baumert (Eds.). *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 201-218). *Mahwah, NJ:* Lawrence Erlbaum Associates Publishers.

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

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*(4), 545-571.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. Personnel Psychology, 47, 537-560).

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

Rubin, D. B. (2005). Causal inference using potential outcomes. Journal of the American Statistical Association, 100(469), 322-331.