Multiple Imputation in Longitudinal Studies with Circularity

Elizabeth Petraglia[1]
[1]Westat, 1600 Research Blvd, Rockville, MD 20850

Abstract
A good imputation model leverages relationships in the complete data to make predictions for missing values. However, there is some disagreement about how to handle imputed values in analyses when the imputation is driven by a single strong predictor, and that predictor will be included in the analysis model. One common situation is when subjects are assessed at two time points (y_1 and y_2), but some subjects are missing scores at one or both time points. Other auxiliary data are available for all subjects. The y_2 score for each subject is typically the strongest predictor of the y_1 score in the imputation model, but there is concern about "circularity" if the planned analyses then use the y_1 score to predict the y_2 score. Suggested approaches in the literature include a multiple imputation then deletion (MID) approach, where all missing values are imputed but observations with imputed outcomes are dropped from analyses; or using all observations (AO), including those with imputed outcomes, for analyses following imputation. This paper investigates the conditions under which circularity may be a concern, studies the performance of the MID and AO methods under different settings, and makes recommendations for practice.

Key Words: Chained equations, missing data, nearest neighbor, nonresponse

1. Introduction

Standard practice for building a multiple imputation model is to include all predictors that will be included in any analysis models, as well as any additional variables that are likely to predict missingness. The former satisfies the requirement for congeniality of the analysis and imputation models (Little and Rubin, 2002; Meng, 1994; Allison 2002); the latter set of variables, often referred to as auxiliary variables, should in theory lead to a more powerful imputation model. In an ideal situation, no single variable drives the imputation model. Although some variables may be stronger predictors, removing one variable would not have a major impact on the imputation results.

Imputation for longitudinal studies is a special case (Little and Su, 1989). A common analytic setup is to control for baseline score (y_1), add some other predictors of interest, and use a follow-up score (y_2) for the outcome. The baseline score is typically by far the strongest predictor of follow-up score. If there are missing data at y_1 and y_2, there is then a concern about circularity in the imputation. The concern is more severe if data are missing at both timepoints. If the y_2 score is used to impute the y_1 value when there are few other strong predictors and the y_1 value is used to predict the y_2 value, the analysis may overstate the strength of the y_1- y_2 relationship. Modeling an overly strong relationship between y_1 and y_2 where both have a moderate to large percentage of imputed values may bias estimates for other variables of interest.

1.1 Multiple Imputation, then Deletion (MID)

One recommendation is multiple imputation, then deletion (MID) (von Hippel, 2007). Under MID, all observations are imputed, but before analysis observations with missing outcomes (in this case, y_2 scores) are dropped. Assuming data are missing at random (MAR) and that the analysis and imputation models are identical (contain the exact same set of variables), observations with an imputed outcome do not add any information to the analysis model and can be safely deleted. However, observations with missing outcome but complete predictors can contribute some information to impute missing predictors, so it is useful to retain them for imputation.

MID is robust to a misspecified imputation model for the outcome, although misspecified imputation models for other variables are still a risk. This is the main advantage of MID for longitudinal data: it protects against circularity in the analysis model, since the model relies only on the relationship between the *observed* y_2 and y_1 (where y_1 may include both imputed and observed values). If the percentage of missing data is not too high, the model will be driven primarily by cases with both observed y_1 and y_2.

MID is not currently implemented in any major statistical software package (although a SAS macro is available from von Hippel) and adds an extra step to analysis. Von Hippel also acknowledges, "in terms of efficiency, the advantage of MID is often quite minor," unless the percentage of missing data is high and there are relatively few imputed data sets.

1.2 Issues with MID

In some scenarios, standard multiple imputation (MI) can in fact lead to smaller standard errors than MID. MI using auxiliary variables will always outperform MID asymptotically; that is, with an infinite number of imputations, MI estimates will have smaller standard errors than the corresponding MID estimates in the presence of any auxiliary variables. A simulation study by Sullivan et al. (2015) found that MID can actually induce bias under MAR in the presence of an auxiliary variable related to missingness for the outcome, and provides only a modest decrease in standard errors when the auxiliary variable is not related to missingness in the outcome under a range of conditions. This result is not limited to a very large number of imputations: the simulation study found that when 50 imputations or more were used, standard MI outperformed MID in terms of both bias and standard error under nearly every scenario tested.

Early literature on multiple imputation suggested that 2 to 10 imputations were sufficient, and more imputations were often computationally infeasible. This is the range of values that von Hippel's original simulations cover. However, recent recommendations suggest that the number of imputations should be linked to the percentage of missing data, and more imputations are better if computing time is not an issue (Graham et al., 2007; White et al., 2011). For example, if about 20 percent of cases in the data require imputation, then generate at least 20 imputations.

Given that many modern applications of MI rely on auxiliary variables and use at least 20 imputations, it is unclear whether the potential for bias due to circularity is greater than the potential for bias due to MID under these conditions.

1.3 Motivation

The simulation study described in Sections 2 and 3 is designed to assess under what conditions MID may be a better choice for practice. Section 2 describes the simulation study setup. Four experimental settings are varied (imputation model correctly specified or misspecified, analysis model correctly specified or misspecified, size of treatment effect, and imputation method—available case, single imputation, or multiple imputation). Each combination of settings is tested with and without deletion of imputed outcomes. In Section 3, the bias in estimating a treatment effect, as well as the width of the associated confidence intervals, is compared at each setting to assess whether including all observations or deleting imputed outcomes performs better. This paper is intended to provide recommendations for practice: using standard statistical software under a range of plausible scenarios, when is MID a better choice for longitudinal studies?

2. Simulation Study

2.1 Simulation Setup and Experimental Settings

This simulation study considers a very simple case: outcomes are measured at baseline and at a single follow-up (y_1 and y_2), with one auxiliary variable x and one treatment effect T. The goal is to fit the linear analysis model

$$y_2 = \alpha + \gamma y_1 + \beta T$$

and estimate the treatment effect, β.

Let $P_i(z)$ denote the i^{th} percentile of z. Then for all simulation runs, define[1]

$$y_1 \sim N(50, 3)$$

$$T \sim Bernoulli(0.5)$$

$$x = \begin{cases} 1, x' < P_{10}(x') \\ 2, P_{10}(x') \leq x' < P_{50}(x') \\ 3, P_{50}(x') \leq x' < P_{90}(x') \\ 4, P_{90}(x') \leq x' \end{cases}, x' \sim N(y_1, 20)$$

Since the relationship between y_1 and y_2 is one of the experimental settings and varies between simulation runs, it is defined below. Essentially x is a categorized variable, constructed so that it is correlated with y_1: observations with low values of y_1 are more likely to also have low values of x. The missingness for y_1 depends only on x; 40 percent of y_1 values are set to missing, with probability proportional to ${}^{1}\!/_{1+x}$.

The missingness for y_2 depends on the difference $y_2 - y_1$ in all simulation runs. Preliminary work showed that if missingness is related to y_1, y_2, or x alone, there is very little difference between estimates and standard errors for MI vs. MID. Practically, it makes sense that in a longitudinal study one may expect to see dropout related to change over time. In a clinical trial, for example, patients who see negative results may be more likely to leave the trial. In educational studies, children and youth with falling test scores may be more likely to change schools or drop out entirely. The missingness is defined as:

[1] Throughout this paper, the normal distribution is given as $N(\mu, \sigma^2)$

$$P(y_2\ mis) = \begin{cases} 0.8,\ y_2\text{-}y_1 < 0 \\ 0.5,\ 0 < y_2\text{-}y_1 < median(y_2\text{-}y_1) \\ 0.25\ otherwise \end{cases}$$

This results in between 40 to 50 percent missingness on y_2.

Four experimental settings are varied:

1. Imputation model is correctly specified (includes x) or is incorrectly specified (excludes x).
2. The relationship between y_1 and y_2 is:
   a. Linear: $y_2 \sim N(y_1 + \beta T,\ 25)$
   b. Nonlinear: $y_2 \sim N(0.5y_1^2 - 50y_1 + 1300 + \beta T,\ 25)$
3. True treatment effect: large (β = 5), small (β= 1), or none (β=0)
4. Imputation method
   a. None (available case analysis)
   b. Single imputation (k-nearest-neighbors via `kNN` function in `VIM` `R` package (Kowarik and Templ, 2016))
   c. Multiple imputation (MICE via `mice` `R` package (van Buuren and Groothuis-Oudshoorn, 2011), m=40 imputations).

For the single and multiple imputation methods, standard imputation is also tested vs. imputation with deletion. Under multiple imputation, this is a test of MI vs. MID; single imputation is included as a test to see if the recommendations are similar across single and multiple imputation methods. Available case analysis is included as a control condition. Methods that consistently perform worse than available case analysis should not be considered.

The imputation methods and implementations were selected because they are widely available, often recommended, and simple to use with default settings (see, for example, Penone et al., 2014; Kalaycioglu et al., 2016; or White et al., 2011). Other imputation methods could be considered as well. Single imputation methods like hot-deck, predictive mean matching, last observation carried forward, etc., were purposely not tested because these methods are not appropriate for high percentages of missing data. More complex imputation methods, such as Bayesian imputation or pattern-mixture models, are beyond the scope of this paper, as are tree-based imputation methods that can require extensive tuning.

The experimental grid contains data generated under 12 scenarios (2 imputation models x 2 true models x 3 true parameter values). Each scenario is run 1,000 times; in each simulation run, a dataset of size n=500 is independently generated. The data from each run are passed through five imputation frameworks: available case (no imputation), single imputation (with or without deletion), and multiple imputation (with or without deletion). The complete data (i.e., before generating missingness) are also retained for each run.

2.2 Results

As a broad metric, we first compare estimates from the fitted linear models after each imputation method to the corresponding estimates from models fit to the complete data. (To simplify discussion, available case analysis is referred to throughout this section as an imputation method, although that is not strictly true.) For each of the 1,000 simulation runs, the relative bias (relbias) for each estimated coefficient was calculated as:

$$relbias = \frac{(\widehat{imputed}\ data - \widehat{complete}\ data)}{\widehat{complete}\ data}$$

Relative bias is used here rather than absolute bias because the size of the true estimate varies depending on the simulation settings, and relative bias is a standardized value that can be compared across settings. The comparison here is to the estimate from the complete data set rather than the true parameter value, as a test of how well each imputation method recovers the information in the complete data.

Figures 1a and 1b summarize the relative bias by parameter and imputation method. The plots are divided by whether the true relationship is linear or nonlinear, since the true parameter values for the intercept and baseline effect varied between the two models.

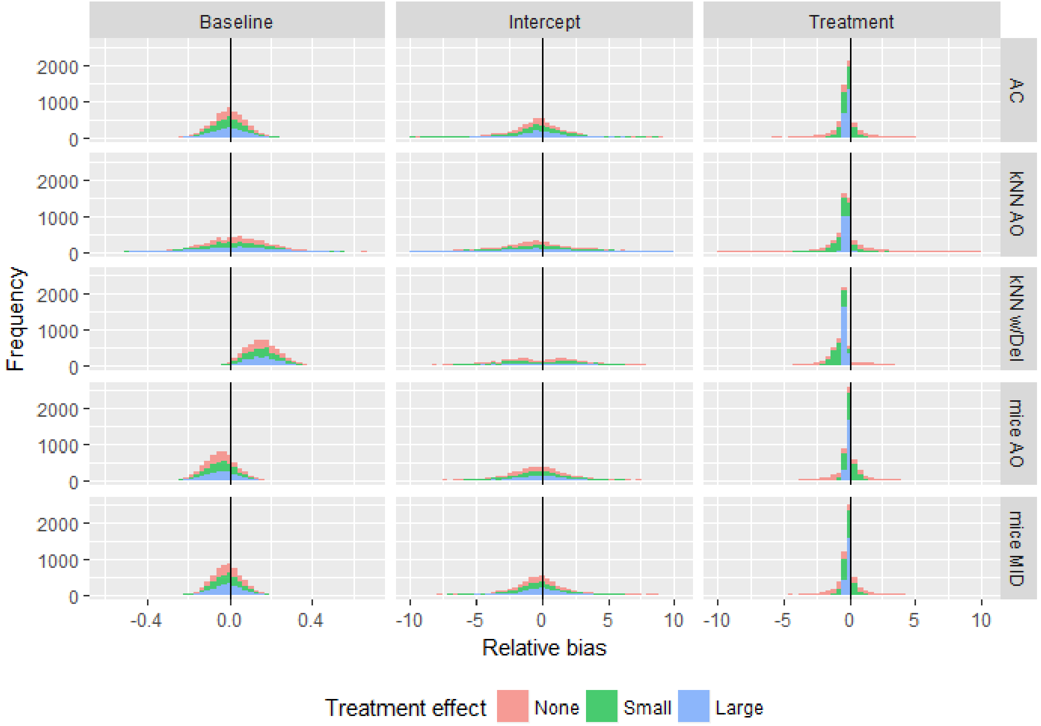

Figure 1a: Histograms of relative bias of estimates, by imputation method, when true model is linear. (Note: For readability, histograms exclude outliers outside of (-10, 10), which represents about 5% of all observations.)

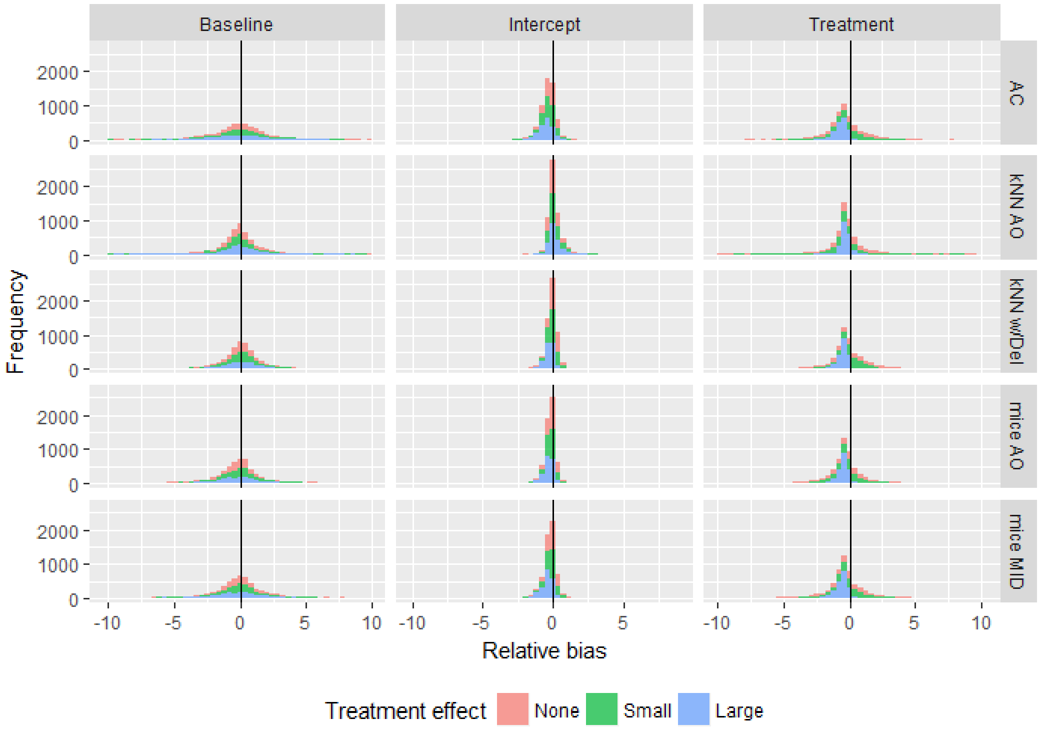

Figure 1b: Histograms of relative bias of estimates, by imputation method, when true model is nonlinear. (Note: For readability, histograms exclude outliers outside of (-10, 10), which represents about 5% of all observations.)

The biases for the baseline effect estimates are generally small and centered around zero, with the exception of kNN with deletion under the linear condition. The relative bias for the baseline effect estimate is slightly lower under mice with deletion, on average, than the relative bias for mice when using all observations, suggesting that MID may offer some protection against circularity and the associated bias in estimating the effect of the baseline score. Biases for the intercept term are more variable than those for the baseline under the linear condition but less variable under the nonlinear condition. However, the treatment effect is almost uniformly underestimated on average; the single imputation method (kNN) performs the worst.

The remaining discussion focuses on the treatment effect, since that is the parameter of most interest. **Table 1** shows that when the analysis model is correct (the true model is linear), all observations (AO) and MID perform similarly, on average, in most of the scenarios tested. The table gives a summary of the differences between the estimate and the true value, so that closer to zero is better for all parameter settings. The average and median biases indicate that using AO is better, on average, under most conditions (smaller average bias). The exception seems to be when $\beta = 1$ and the imputation model is misspecified. This is not a particularly useful result since *a priori* one would not know the size of the treatment effect, nor whether the imputation model is correct. Available case analysis actually performs quite well comparatively, which is not entirely surprising when the missing data mechanism is not MAR; imputation methods assuming MAR are not capturing the full missingness mechanism.

Table 1: Summary of Bias From Linear Relationship Runs (Estimate – True Value), 1,000 Simulation Runs Per Setting

	Correct imputation model				*Misspecified imputation model*			
				$\beta = 0$				
	Min	*Median*	*Mean*	*Max*	*Min*	*Median*	*Mean*	*Max*
Available case	-1.78	0.00	0.00	1.44	-6.98	0.49	0.51	9.07
kNN, All obs	-2.20	-0.06	-0.06	1.88	-5.89	-1.33	-1.34	3.09
kNN, Deletion	-1.84	-0.54	-0.53	0.76	-5.91	1.38	1.42	8.06
mice, All obs	-1.18	0.30	0.29	1.67	-9.46	0.20	0.29	7.83
mice, MID	-1.25	0.14	0.13	1.38	-10.46	0.21	0.33	9.53
				$\beta = 1$				
	Min	*Median*	*Mean*	*Max*	*Min*	*Median*	*Mean*	*Max*
Available case	-1.69	-0.23	-0.23	1.37	-7.17	-0.43	-0.42	7.06
kNN, All obs	-2.57	-0.39	-0.40	1.63	-6.08	-1.45	-1.44	2.74
kNN, Deletion	-2.09	-0.75	-0.75	0.52	-5.39	0.50	0.40	5.04
mice, All obs	-1.21	0.03	0.02	1.61	-6.71	-0.73	-0.68	7.32
mice, MID	-1.39	-0.10	-0.12	1.28	-7.45	-0.58	-0.57	7.47
				$\beta = 5$				
	Min	*Median*	*Mean*	*Max*	*Min*	*Median*	*Mean*	*Max*
Available case	-2.29	-0.81	-0.81	0.83	-8.80	-3.03	-3.00	3.73
kNN, All obs	-2.75	-0.99	-0.96	1.47	-7.61	-1.89	-1.88	2.10
kNN, Deletion	-2.86	-1.31	-1.32	0.31	-7.42	-2.38	-2.42	2.57
mice, All obs	-1.85	-0.59	-0.58	1.07	-11.29	-2.94	-2.82	4.37
mice, MID	-2.05	-0.71	-0.71	0.67	-10.91	-3.34	-3.23	4.53

Figure 2 shows the results for the linear, $\beta = 5$ scenarios under the correct imputation model. First, a subsample of 100 simulation runs was selected so that the graph does not appear too cluttered. The final estimate from each simulation run is shown as a colored dot, with its 95% confidence interval displayed as a light grey horizontal line. The average estimate from each imputation method is marked with a darker grey vertical line. The "correct" (complete data) estimates are shown as green stars, with the average complete data estimate marked by a black vertical line. The ideal imputation plot would show the green stars virtually covering the colored dots, and the grey and black vertical lines overlapping. In Figure 1 it appears that estimates from all imputation methods are biased low, on average, since the grey vertical line is uniformly lower than the black line. The least biased method (shortest distance between the grey and black lines) seems to be mice with AO, although there is a great deal of variability in all plots. Note also the difference in the 95% confidence interval lengths: kNN AO has very short, almost certainly underestimated 95% confidence intervals, while mice AO has much wider intervals.

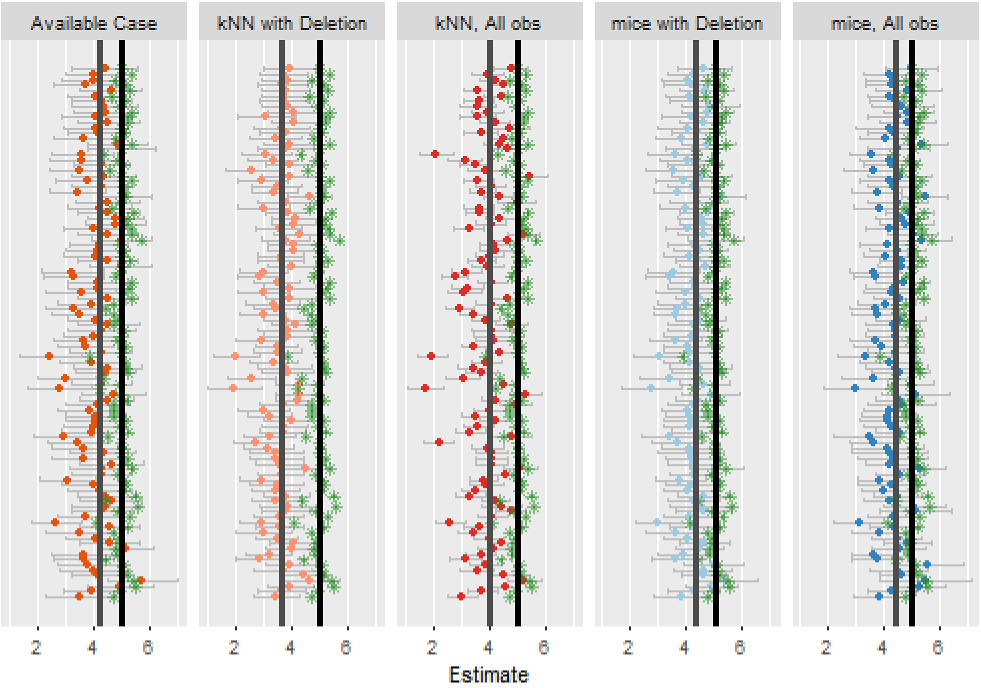

Figure 2: Treatment effect estimates and confidence intervals by imputation type, linear relationship, $\beta = 5$, correct imputation model. Grey vertical line marks average estimate for imputation type; green stars indicate complete data estimates, and black vertical line marks average complete data estimate.

The difference in confidence interval lengths is more striking when the true relationship between y_1 and y_2 is nonlinear—that is, the analysis model is incorrect. **Figure 3** displays treatment effect estimates for each simulation run in the nonlinear case when the imputation model is misspecified and the treatment effect is large ($\beta = 5$). The confidence intervals for mice AO are extremely wide, while the confidence intervals for MID are much narrower. The average bias still appears to be slightly smaller under AO vs. deletion. **Table 2** summarizes the full results for the simulation runs when the analysis model is misspecified. Under most conditions AO performs somewhat better than or only slightly worse than deletion, on average, and the range of observed biases is often narrower as well. The exception is when $\beta = 1$, where both single and multiple imputation perform better on average with deletion.

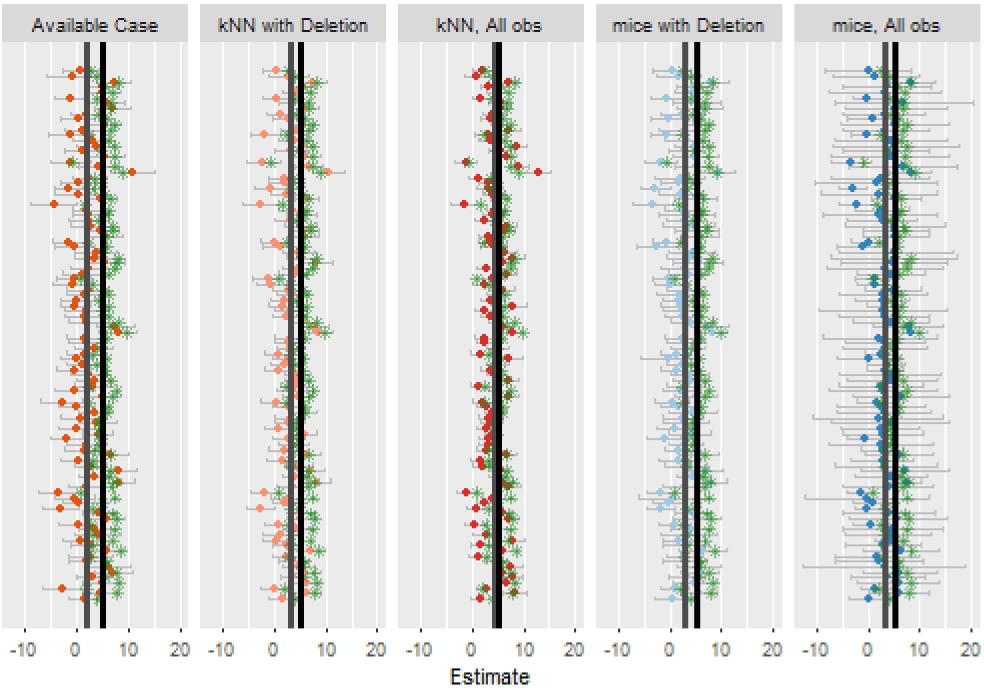

Figure 3: Treatment effect estimates and confidence intervals by imputation type, nonlinear relationship, $\beta = 5$, misspecified imputation model. Grey vertical line marks average estimate for imputation type; green stars indicate complete data estimates, and black line marks average complete data estimate.

Table 2: Summary of Bias From Nonlinear Relationship Runs (Estimate – True Value), 1,000 Simulation Runs Per Setting

	Correct imputation model				*Misspecified imputation model*			
					$\beta = 0$			
	Min	*Median*	*Mean*	*Max*	*Min*	*Median*	*Mean*	*Max*
Available case	-6.98	0.49	0.51	9.07	-7.98	0.55	0.50	7.81
kNN, All obs	-5.89	-1.33	-1.34	3.09	-6.02	-0.74	-0.58	4.76
kNN, Deletion	-5.91	1.38	1.42	8.06	-4.99	1.89	1.86	7.14
mice, All obs	-9.46	0.20	0.29	7.83	-7.43	0.98	0.97	7.87
mice, MID	-10.46	0.21	0.33	9.53	-9.36	1.15	1.11	7.56
					$\beta = 1$			
	Min	*Median*	*Mean*	*Max*	*Min*	*Median*	*Mean*	*Max*
Available case	-7.17	-0.43	-0.42	7.06	-9.36	-0.64	-0.64	7.28
kNN, All obs	-6.08	-1.45	-1.44	2.74	-5.18	-0.77	-0.65	4.65
kNN, Deletion	-5.39	0.50	0.40	5.04	-4.86	0.52	0.56	5.91
mice, All obs	-6.71	-0.73	-0.68	7.32	-6.43	-0.24	-0.20	5.78
mice, MID	-7.45	-0.58	-0.57	7.47	-6.83	0.01	0.01	6.33
					$\beta = 5$			
	Min	*Median*	*Mean*	*Max*	*Min*	*Median*	*Mean*	*Max*
Available case	-8.80	-3.03	-3.00	3.73	-9.93	-3.09	-3.06	5.31
kNN, All obs	-7.61	-1.89	-1.88	2.10	-5.29	-0.98	-0.86	4.91
kNN, Deletion	-7.42	-2.38	-2.42	2.57	-7.35	-2.11	-2.06	3.27
mice, All obs	-11.29	-2.94	-2.82	4.37	-10.18	-1.83	-1.92	3.56
mice, MID	-10.91	-3.34	-3.23	4.53	-11.16	-2.37	-2.41	3.20

The much wider confidence intervals under AO in the nonlinear case are evident in **Figure 4**. Each bar shows the range of 95% confidence interval half-widths, minimum to maximum, and the mean half-width is marked with a dot. Under the linear condition, the half-widths do not vary dramatically across imputation methods, regardless of the other settings. However, under the nonlinear condition, it is clear that available case analysis and single imputation produce confidence intervals that are too narrow. More importantly, mice AO results in confidence interval half-widths that are much larger, on average, than those generated by MID. When the imputation model is also misspecified, the maximum half-width observed for MID is equal to or less than the *average* half-width observed for mice AO.

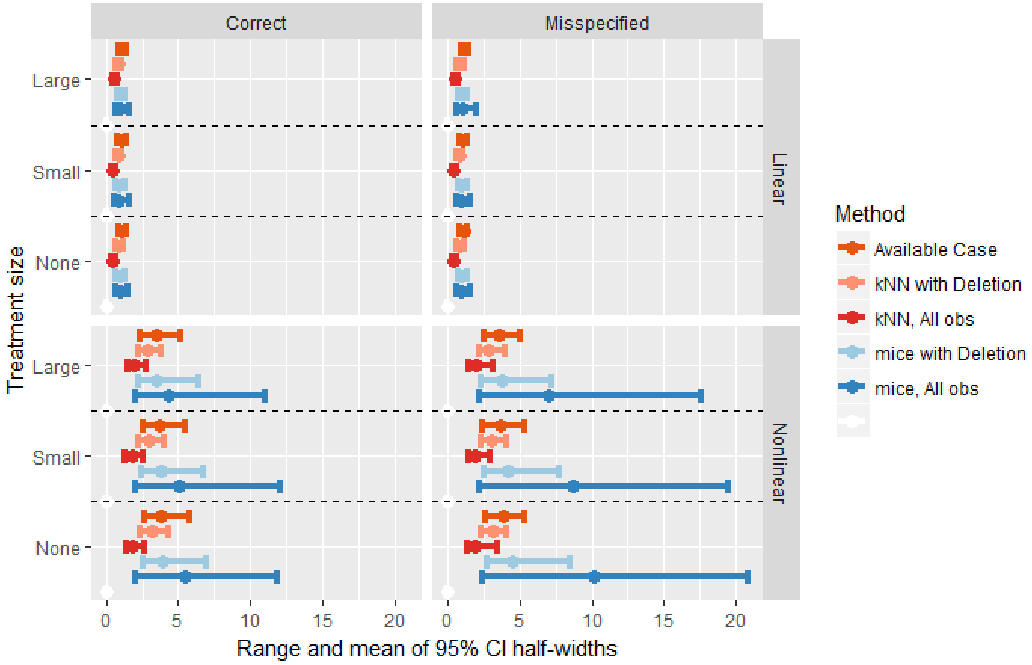

Figure 4: Range and mean of 95% confidence interval half-widths, by setting and imputation method.

If AO and MID perform similarly, with AO perhaps having slightly less bias on average but much wider confidence intervals under some conditions, the natural question to ask is whether the wider confidence intervals reflect the true variability of the estimates. While it seems attractive to choose a method with much smaller standard errors at the risk of a small amount of bias, falsely low standard errors can lead to nominal *p*-values that are much too low and nominal 95% confidence intervals with true coverage rates that are much lower.

In **Figure 5**, the percentage of 95% confidence intervals that cover the true parameter value ($\beta = 0, 1$ or 5, as appropriate) is shown for each method. Ideally, for a 95% confidence interval we would like to see 95% coverage. Across all conditions kNN, the single imputation method, performs poorly, with coverage rates less than 50 percent in several scenarios. This is not surprising, since single imputation methods do not account for uncertainty due to imputation and are known to underestimate variance. When the true parameter value is large, mice AO does not always achieve the desired 95% coverage, but it consistently performs better than MID.

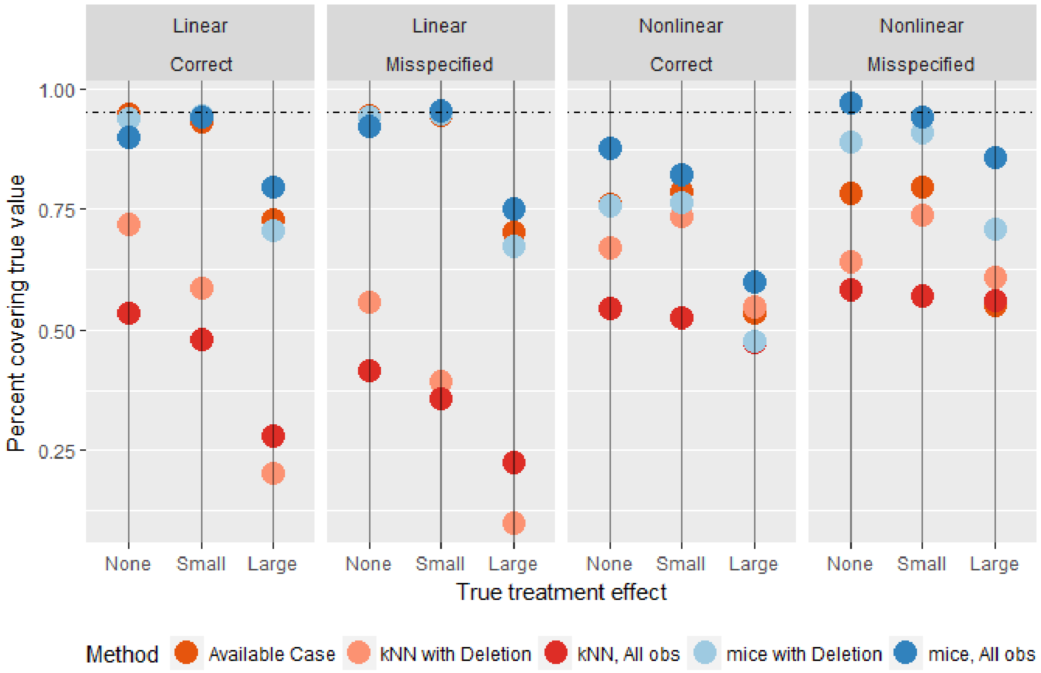

Figure 5: Coverage of nominal 95% confidence intervals, by imputation method.

A related question is whether one would make the correct conclusion about statistical significance under each imputation method. **Figure 6** assumes that the analysis uses a bright line at *p* = 0.05 to determine statistical significance and plots the percentage of simulation runs that make the correct decision about the presence or absence of the treatment effect. (This metric disregards the American Statistical Association's best practice guidelines about the use of *p*-values but makes the decision rule much easier to program!). Again, with a *p*-value cutoff of 0.05 one would expect to see that the correct decision is made about 95% of the time when there is no treatment effect (β = 0). In the presence of a treatment effect, 80 percent power to detect the effect is a commonly used cutoff. In general, higher (closer to 1) is better across all plots.

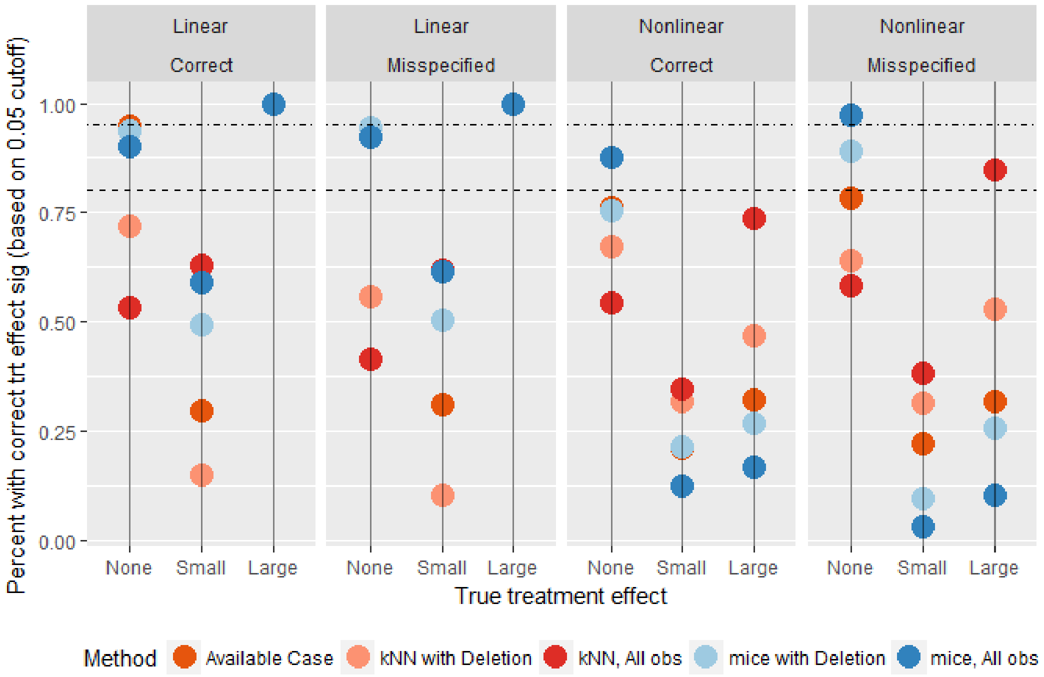

Figure 6: Percentage of simulation runs with the correct inference on the treatment effect, using a cutoff of p = 0.05, by imputation method.

When the analysis model is correct (linear), all methods correctly detect the large treatment effect. The single imputation method performs poorly, however, when the treatment effect is small or there is no treatment effect. For multiple imputation, MID and AO perform similarly when there is no treatment effect, and AO is slightly better in the presence of a small treatment effect. When the analysis model is incorrect (the true relationship is nonlinear), the large standard errors for the multiple imputation estimates mean that it is very difficult to achieve statistical significance, even in the presence of a large treatment effect. MID performs somewhat better than AO but still only detects the treatment effect 25 percent of the time, and consistently underperforms even available case analysis. Conversely, AO is more accurate when there is no treatment effect, while MID falsely concluded that there was a treatment effect in up to 25 percent of simulation runs.

3. Conclusion and Recommendations

The limited simulation study performed on longitudinal data suggests that MID does lead to narrower confidence intervals but at the risk of serious undercoverage of nominal 95% confidence intervals, especially when the analysis model is misspecified. Any bias potentially due to circularity when using all observations was found to be either similar to, or smaller than, the bias induced by MID. This suggests that concern about circularity alone is not sufficient to justify the use of MID. Deletion of imputed outcomes generally improved power but decreased coverage of the true parameter estimates when a treatment effect was present.

Any analysis method promising smaller standard errors is always attractive; for some simulation runs where both the analysis and imputation models were misspecified, the half widths of the 95% confidence intervals generated under MID were less than half the size of those generated using AO. This simulation work suggests that those narrower confidence intervals may often understate the true variability in the MID estimates, however.

A conservative recommendation is to use all observations for analysis and treat large confidence intervals as a signal that either the imputation or analysis models, or both, may be misspecified. If the model(s) remain misspecified, inference based on AO will likely be conservative; in most situations, conservative inference is preferable to reporting nominal coverage rates that are misleadingly high. More research is needed, however, especially with more complex imputation models.

Acknowledgements

The author would like to thank Robert Baskin, Sharon Lohr, Andrea Piesse, and many other statisticians at Westat for helpful comments that improved this research.

References

Allison, P. (2002). *Missing data*. Thousand Oaks, CA: Sage Publications, Inc.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science,* 8(3), 206-213.

Kalaycioglu, O., Copas, A., King, M., and Omar, R. Z. (2016). A comparison of multiple-imputation methods for handling missing data in repeated measurements observational studies. *Journal of the Royal Statistical Society, Series A*, 179(3), 683-706.

Kowarik, A., and Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software*, 74(7), 1-16. doi:10.18637/jss.v074.i07

Little, J. A., and Rubin, D. B. (2002). Statistical analysis with missing data: Second Edition. New York: John Wiley & Sons.

Little, J. A., and Su, H-L. (1989). Item nonresponse in panel surveys. In *Panel Surveys* (D. Kasprzyk, G.J. Duncan, G. Kalton, and M.P. Singh, eds.). Wiley: New York, pp. 400-425.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science,* 9, 538-558.

Penone, C., Davidson, A. D., Shoemaker, K. T., Di Marco, M., Rondinini, C., Brooks, T. M., Young, B. E., Graham, C. H., and Costa, G. C. (2014). Imputation of missing data in life-history trait datasets: which approach performs the best? Methods in Ecology and Evolution, 5, 961-970. doi:10.1111/2041-210X.12232

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Sullivan, T. R., Salter, A. B., Ryan, P., Lee, K. J. (2015). Bias and precision of the "multiple imputation, then deletion" method for dealing with missing outcome data. *American Journal of Epidemiology*, 182(6), 528-534.

van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1-67. URL http://www.jstatsoft.org/v45/i03/

von Hippel, P. T. Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology,* 37(1), 83-117.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine,* 30, 377-399