Received: 7 August 2018 | Revised: 6 December 2019 | Accepted: 13 December 2019

DOI: 10.1002/sim.8468

RESEARCH ARTICLE

Statistics in Medicine WILEY

A fair comparison of tree-based and parametric methods in multiple imputation by chained equations

Emily Slade[1,2] | **Melissa G. Naylor**[1]

[1]Pfizer Worldwide Research and Development, Cambridge, Massachusetts

[2]Department of Biostatistics, University of Kentucky, Lexington, Kentucky

Correspondence
Emily Slade, Department of Biostatistics, University of Kentucky, 725 Rose Street, Lexington, KY 40536.
Email: emily.slade@uky.edu

Funding information
Northern California Institute for Research and Education; Foundation for the National Institutes of Health; Canadian Institutes of Health Research; Transition Therapeutics; Takeda Pharmaceutical Company; Servier; Piramal Imaging; Pfizer Inc.; Novartis Pharmaceuticals Corporation; Neurotrack Technologies; NeuroRx Research; Meso Scale Diagnostics, LLC.; Merck & Co., Inc.; Lundbeck; Lumosity; Johnson & Johnson Pharmaceutical Research & Development LLC.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; IXICO Ltd.; GE Healthcare; Fujirebio US; Genentech, Inc.; F. Hoffmann-La Roche Ltd; Euroimmun US, Inc; Eli Lilly and Company; Elan Pharmaceuticals, Inc.; Cogstate; Eisai Inc.; CereSpir, Inc.; Bristol-Myers Squibb Company; Biogen; BioClinica, Inc.; Araclon Biotech; Alzheimer's Drug Discovery Foundation; Alzheimer's Association; AbbVie; National Institute of Biomedical Imaging and Bioengineering; National Institute on Aging; Department of Defense, W81XWH-12-2-0012; National Institutes of Health, U01 AG024904; Alzheimer's Disease Neuroimaging Initiative

Multiple imputation by chained equations (MICE) has emerged as a leading strategy for imputing missing epidemiological data due to its ease of implementation and ability to maintain unbiased effect estimates and valid inference. Within the MICE algorithm, imputation can be performed using a variety of parametric or nonparametric methods. Literature has suggested that nonparametric tree-based imputation methods outperform parametric methods in terms of bias and coverage when there are interactions or other nonlinear effects among the variables. However, these studies fail to provide a fair comparison as they do not follow the well-established recommendation that any effects in the final analysis model (including interactions) should be included in the parametric imputation model. We show via simulation that properly incorporating interactions in the parametric imputation model leads to much better performance. In fact, correctly specified parametric imputation and tree-based random forest imputation perform similarly when estimating the interaction effect. Parametric imputation leads to slightly higher coverage for the interaction effect, but it has wider confidence intervals than random forest imputation and requires correct specification of the imputation model. Epidemiologists should take care in specifying MICE imputation models, and this paper assists in that task by providing a fair comparison of parametric and tree-based imputation in MICE.

KEYWORDS
imputation, interaction, missing data, regression tree

ABBREVIATIONS: ADAS-Cog 13, Alzheimer's Disease Assessment Scale - Cognition 13-item; ADNI, Alzheimer's Disease Neuroimaging Initiative; CART, classification and regression trees; CI, confidence interval; HV, hippocampal volume; JAV, just another variable; MAE, mean absolute error; MAR, missing at random; MI, multiple imputation; MICE, multiple imputation by chained equations; PMM, predictive mean matching; RF, random forests.

1 | INTRODUCTION

Dealing with missing data is a necessary reality faced by researchers analyzing epidemiological data. One solution, multiple imputation (MI), involves filling in the missing data with plausible values multiple times, carrying out the desired analysis on each filled-in dataset, and combining the results using Rubin's rules.[1] MI has become popular for a number of reasons. First, it gives researchers the ability to perform the imputation procedure in advance so future analysts can easily carry out their complete-data analysis of choice on the multiply-imputed datasets. Also, when performed correctly, MI produces unbiased estimates and valid inference for data that are missing at random (MAR) (i.e., in scenarios when the probability of data being missing does not depend on the missing data given the observed data).[2,3]

Multiple imputation by chained equations (MICE), also called fully conditional specification, is commonly used to implement MI.[4,5] At each step, imputed values for one variable are drawn from a predictive model conditional on all other variables. This process cycles through the imputation of each variable until imputations converge. The primary advantage of MICE is that it is not necessary to specify a joint distribution of all variables. By default, software to implement MICE includes each variable as a linear predictor in the imputation model with no interactions or nonlinearities considered. However, it is widely known that the imputation model must be "compatible" with, that is, at least as complex as, the final analysis model.[3,6-9] Therefore, any interactions or non-linear relationships that are estimated in the final analysis model must be accounted for in the imputation model. Failure to specify a compatible imputation model can result in biased parameter estimates and invalid inference.[6]

In some cases, however, it may not be desirable or even possible to include interactions in a parametric imputation model due to a large number of predictors, small sample size, or highly correlated predictors. In these cases, recursive partitioning (tree-based) methods are commonly utilized within the MICE algorithm. Tree-based methods such as classification and regression trees (CART) and random forests (RF) are nonparametric and have been used to model large, complex data in clinical medicine, genetics, and more.[10] The primary advantages of tree-based methods are their ability to capture complex relationships such as interactions and other nonlinearities as well as their nonparametric nature, that is, they do not require the user to specify an imputation model.[10-12] At each step in the MICE algorithm, values are imputed for a given variable using a tree built with all other variables as predictors.[11] A primary drawback of tree-based methods is the difficulty of interpreting their results, but this is inconsequential for imputation as interest lies only in preserving complex data structure to make unbiased parameter estimates and valid inference.

Many have demonstrated the success of tree-based methods for MI.[11-13] However, these studies have failed to compare tree-based MICE to MICE using a correctly specified parametric imputation model. In some studies,[13] the number of predictors is too large to include all possible interactions in a parametric imputation model, but in others,[12] the selected parametric imputation model is simply not as complex as the final analysis model. Doove et al.[12] selected their parametric imputation model based on the default settings for MICE in R, leading to an imputation model that includes only main effects of each variable while the analysis model also includes an interaction term. Doove et al.[12] demonstrated poor performance of parametric MICE in this case, but this result is expected as it has been well-documented that biased estimates and low coverage will result from a final analysis model that is more complex than the imputation model.[2,3,6,7]

This paper makes a fair and novel comparison of tree-based imputation models to parametric imputation models in MICE by specifying a compatible parametric imputation model as would be done in practice. We compare performance of the MICE imputation methods via simulation, and we present an application of these methods to data relating hippocampal volume (HV) and age to cognitive function in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort.

2 | METHODS

We compared the performance of parametric and tree-based imputation methods via simulation using two data generation models. For each combination of data generation model and imputation method, the following steps were performed: data generation, removal of observations based on a MAR mechanism, imputation, regression analysis, and calculation of bias, coverage, and confidence interval (CI) width for each coefficient. In general, a desirable imputation method leads to low bias, coverage of at least 95%, and narrow CIs.

2.1 | Data generation

The data generation models represent moderately sized studies containing a continuous outcome and four covariates, similar to that in Doove et al.[12] Scenario 1 represents this setting when there is a true underlying interaction between two of the covariates, as shown in Equation (1) below.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{int} X_1 X_2 + \varepsilon \quad (1)$$

Scenario 2 represents this setting when there is no true underlying interaction, as shown in Equation (2) below.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \quad (2)$$

For both scenarios, we simulated data for $n = 200$ subjects. $X_1, \ldots, X_4$ are generated by a multivariate normal distribution with $E[X_i] = 0$, $\text{Var}(X_i) = 1$, $\text{Cov}(X_i, X_j) = 0.5$ for $i,j \in \{1,\ldots,4\}$, $i \neq j$, and $\varepsilon \sim N(0,1)$. In Scenario 1, we set $\beta_0 = 0$ and $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_{int} = 0.3$, and in Scenario 2, we set $\beta_0 = 0$ and $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.317$, so that in each scenario 50% of the variation in y is explained by the predictors.

2.2 | Removal of observations based on a MAR mechanism

We induced missingness via a MAR mechanism such that the total proportion of missing data is approximately 50%. In each scenario, missingness in the outcome and covariates is based on the fully observed X_4. Subjects with $X_4 \geq 0$ have 35% missingness in other variables, and those with $X_4 < 0$ have 65% missingness in other variables, on average.

2.3 | Imputation of missing data

Throughout the analysis, we performed MICE using the *mice* package in R version 3.1.0.[4,14] As implemented in Doove et al.,[12] we created 10 imputations for each simulated dataset. Within the MICE algorithm, we applied four imputation methods: (1) predictive mean matching using default *mice* settings (PMM-Naive), (2) predictive mean matching (PMM) with an interaction term in the imputation model (PMM-Int), (3) CART, and (4) RF.

2.3.1 | PMM-Naive

To impute for a given subject, PMM first uses a parametric model to identify subjects with similar predictive means, then it samples one observed value from this group of similar subjects.[1] As a parametric method, PMM is generally preferred over standard regression because it creates imputations from the observed data itself, which maintains data structure such as skewness and avoids problems such as imputing impossible values.[15] The *mice* function uses PMM as the default parametric imputation method for continuous variables. The method we call "PMM-Naive" consists of the default implementation of PMM in *mice*, which includes only main effects in the imputation model. PMM-Naive corresponds to standard PMM presented in Doove et al.[12]

2.3.2 | PMM-Int

The only difference between PMM-Int and PMM-Naive is the inclusion of an interaction term in the imputation model for PMM-Int. To account for an interaction term in the imputation model, some use passive imputation methods,[4-6] while others recommend the just-another-variable method.[5,7,16] Using the *mice* function, there are several ways to specify both the passive and just-another-variable models. A brief comparison of all passive and just-another-variable implementations we considered is presented in Supplementary Material A in Appendix S1. For PMM-Int, we included the interaction term as just another variable and used the default predictor matrix in the *mice* function since this method performed better than or equivalent to all other methods considered.

2.3.3 | CART

CART is a tree-based imputation method that does not require specification of an imputation model. In short, at each step the CART algorithm identifies a binary decision rule based on one predictor variable that partitions the data into two nodes by minimizing variance of the outcome within each node.[17] The tree is grown by continuing this splitting recursively until reaching a stopping point determined by the tuning parameters.[17] Predictions (or in this case, imputations) are made from the regression tree by identifying the terminal node to which a new subject belongs and sampling from the outcomes in that node.[17] For a more detailed description of the CART algorithm in MICE, refer to Burgette and Reiter.[11] We used the *rpart* package to implement CART in *mice* with all default tree-based tuning parameters for CART in the *mice* function including a complexity parameter of 10^{-4} and a minimum of five observations in any terminal node.[4,18]

2.3.4 | RF

The RF imputation method creates multiple regression trees where imputations are a random draw from what would be the imputed value from each tree.[19] To implement RF, variation is introduced in the trees by using bootstrap samples of the original data combined with random input selection to fit each tree.[19,20] Random input selection restricts the possible predictors on which to split each node to a random subset of all possible predictors.[19,20] For a more detailed description of the RF algorithm in MICE, refer to Doove et al.[12] We used the *randomForest* package to implement RF in *mice*.[4,20] We created 10 trees in each random forest, which is the default in the *mice* function.[4,13] By default, the number of predictors considered for splitting at each node is $p/3$ rounded down to the nearest integer, where p is the number of predictors.[4]

2.4 | Regression analysis

For each of the 10 imputed datasets, we fit a correctly specified final analysis model, and the results are combined using Rubin's rules.[1] For each of the four imputation methods, this leads to a set of point estimates and 95% CIs, one for each coefficient.

2.5 | Calculation of bias, coverage, and CI width

For each coefficient, we computed bias as the estimated coefficient minus the true value. For a single repetition of the simulation, coverage is 0 if the estimated 95% CI does not contain the true value and 1 if the estimated interval does contain the truth. Confidence interval width refers to the width of the estimated 95% CI for each coefficient.

Steps 1 to 5 are repeated 10,000 times. To mirror the comparisons made in Doove et al.[12] and Shah et al.,[13] we report the mean bias, coverage, and mean 95% CI width across the replications for each of the four imputation methods. Additionally, we present empirical mean absolute error (MAE) for each coefficient, which is computed as the average absolute bias over the 10,000 simulation replications. A link to the source code used for this simulation can be found in Supplementary Material B in Appendix S1.

3 | DATA APPLICATION

To demonstrate application of these methods in practice, data were obtained from the ADNI database.[21] The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early Alzheimer's disease. Restrictions apply to the availability of these data, which were used under license for this study. Data are available at adni.loni.usc.edu with the permission of the ADNI.[21]

For this analysis, we sought to examine the effects of HV and age on cognitive function, allowing for possible interaction between HV and age. Our specific outcome of interest is cognitive function assessed with the modified Alzheimer's Disease Assessment Scale - Cognition 13-item (ADAS-Cog 13) scale where higher scores indicate greater cognitive

impairment.[22] When using a regional volume in the brain such as HV, the literature suggests normalizing the raw volume to intracranial volume which eliminates the correlation between HV and intracranial volume while not affecting the association between HV and cognition.[23] As such, instead of using the raw HV in our model, we use the normalized HV calculated from the residuals of a linear model regressing raw HV on intracranial volume.[23] ADAS-Cog 13, normalized HV, and age were all standardized prior to imputation and analysis. Normalized HV and age are approximately normally distributed, and ADAS-Cog 13 is slightly skewed right. Bivariate relationships between each predictor and ADAS-Cog 13 reveal no concerning departures from linearity. Plots showing the distributions of these variables in the ADNI sample can be found in Supplementary Material C in Appendix S1.

The full ADNI dataset at baseline includes 1737 subjects, 1479 (85.1%) of which have complete data on HV, age, and ADAS-Cog 13. The goal of this analysis is to utilize imputation methods to include all subjects in the final analysis while preserving underlying relationships between the variables of interest. A preliminary complete case analysis revealed that there may be some evidence of an HV-age interaction ($P = .126$), so preserving this potential relationship is important in the analysis. Despite the fact that we do not know the true underlying effects in these real data, we implement PMM-Naive, PMM-Int, CART, and RF to compare the results that would be obtained using each of these methods. We hypothesize that PMM-Naive will lead to underestimation of the HV-age interaction while PMM-Int and tree-based methods will be able to preserve the interaction effect in the final analysis.

4 | RESULTS

4.1 | Scenario 1

Scenario 1 represents a study of moderate complexity with a true interaction. For this scenario, Figure 1 displays the distribution of bias across the 10,000 replications, and Tables 1 and 2 display the coverage and average 95% CI width, respectively, for each imputation method. For the interaction effect, PMM-Naive leads to estimates with higher mean bias than the tree-based methods, corroborating the findings in Doove et al.[12] Further, coverage of the interaction effect is smallest (only 68.0%) using PMM-Naive. These findings are not surprising, as the imputation model for PMM-Naive includes only the main effects, ignoring the true underlying interaction. PMM-Int, which correctly includes the interaction term in the imputation model, has smaller mean bias and larger coverage (93.3%) for the interaction effect than the tree-based methods. Thus, for estimating a true interaction effect, including the interaction term in the parametric imputation model avoids large mean bias and low coverage.

For the main effects, PMM-Int also leads to small mean bias and approximately 95% coverage. However, RF imputation, which leads to greater mean bias than PMM-Int, also attains at least 95% coverage with the narrowest 95% CIs. At first, this result may seem counterintuitive; however, the 5th and 95th percentiles of the empirical bias distribution in Figure 1 reveal why it is the case. Across replications of this simulation, the biases of the main effects from the RF method have a narrower distribution than those from the PMM-Naive, PMM-Int, and CART methods. Thus, despite the fact that the RF method leads to point estimates with higher mean bias than PMM-Int, it maintains valid coverage for the main effects with the narrowest CIs. We observed this phenomenon for all main effects in Scenario 1, but as shown in Figure 1, this phenomenon does not occur for the interaction effect.

Despite the fact that others have previously used mean bias to compare MICE imputation methods,[12,13] variability in the spread of the bias from each replication makes the mean bias a poor summary measure for considering precision as well as accuracy of the coefficient estimates. A distribution of bias that is centered around 0 but very wide will have lower mean bias than a narrow but slightly off-center distribution. However, the latter arguably provides better estimation of the coefficient than the former. Due to this phenomenon, we find it important to report a measure such as the MAE or mean squared error rather than relying solely on the mean bias for this comparison.

For the main effects, RF leads to smaller MAE than PMM-Naive, PMM-Int, and CART. Assessing the accuracy of coefficient estimation with MAE allows one to see that, for main effects, RF produces estimates that are more accurate, have valid coverage, and have narrower 95% CIs than the other three methods.

For the interaction term, PMM-Naive leads to the least accurate coefficient estimation regardless of whether accuracy is assessed as mean bias or MAE. However, using MAE, PMM-Int no longer appears to have more accurate coefficient estimates than tree-based methods.

For results in this scenario, we also investigated the selection of tree-based tuning parameters. For CART, we varied the complexity parameter (*cp*) from 10^{-6} to 0.1 and the minimum number of observations in each leaf (*minbucket*) from

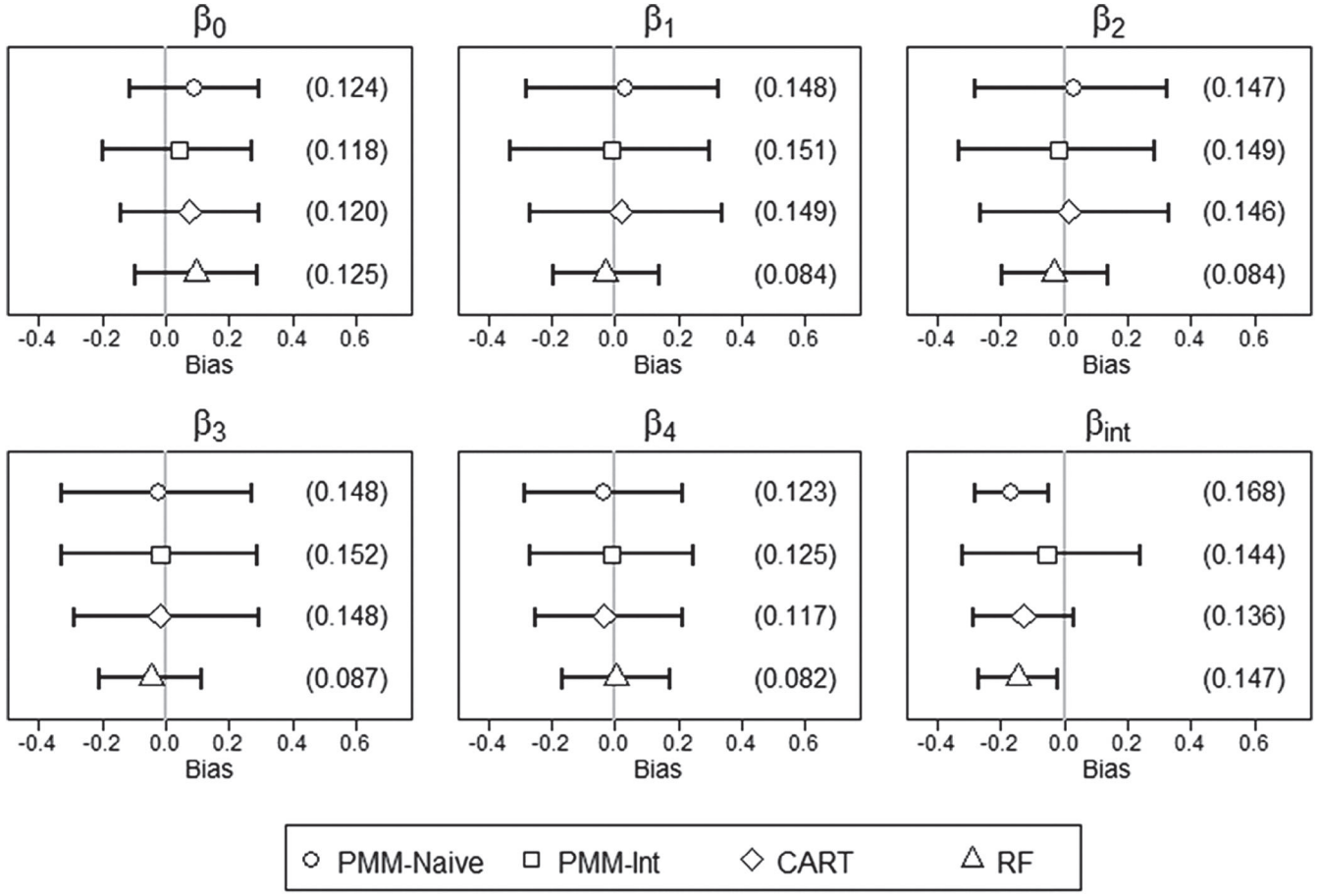

FIGURE 1 Scenario 1: Distribution of bias of the estimated coefficients from each imputation model (10,000 replications). Tick marks indicate the 5th and 95th percentiles. Mean absolute error is shown in parentheses.

TABLE 1 Scenario 1: Coverage (%) for each coefficient

Imputation model	β_0	β_1	β_2	β_3	β_4	β_{int}
PMM-Naive	90.2	94.2	94.4	95.1[a]	95.0[a]	68.0
PMM-Int	94.2	95.3[a]	95.6[a]	95.7[a]	95.6[a]	93.3
CART	89.4	92.3	92.8	92.8	94.0	85.7
RF	87.1	99.4[a]	99.4[a]	99.2[a]	98.9[a]	88.0

Abbreviations: CART, classification and regression trees; PMM-Int, predictive mean matching with an interaction term in the imputation model; PMM-Naive, predictive mean matching using default *mice* settings; RF, random forests.
[a]Coverage ≥ 95.0%.

TABLE 2 Scenario 1: 95% confidence interval width for each coefficient

Imputation model	β_0	β_1	β_2	β_3	β_4	β_{int}
PMM-Naive	0.527	0.790	0.795	0.804	0.640	0.421
PMM-Int	0.602	0.827	0.827	0.837	0.676	0.757
CART	0.503	0.740	0.743	0.736	0.587	0.496
RF	0.477	0.635	0.635	0.626	0.548	0.501

Abbreviations: CART, classification and regression trees; PMM-Int, predictive mean matching with an interaction term in the imputation model; PMM-Naive, predictive mean matching using default *mice* settings; RF, random forests.

1 to 15. We found that $cp > 0.01$ leads to increased mean bias and decreased coverage for all effects, but the default value of 10^{-4} is appropriate. Smaller values of *minbucket* lead to slightly decreased mean bias, but the default value of 5 seemed reasonable. For RF, we varied the number of predictors considered for splitting at each node (*mtry*) from 1 to 4, but it had very little effect on results in Scenario 1. We did not consider varying the number of trees (*ntree*) since simulations in Shah et al.[13] suggested that imputation quality is equivalent for *ntree* = 10 and *ntree* = 100. Since the performance of CART and RF could not be notably improved by altering the tuning parameters, we present results using the default tuning parameters ($cp = 10^{-4}$, *minbucket* = 5, *mtry* = 1). Simulation results for different values of these tuning parameters can be found in Supplementary Material D in Appendix S1.

The relative performance of these imputation methods does not change based on varying the sample size ($n \in [100, 1000]$), interaction effect size ($\beta_{\text{int}} \in [0.1, 0.5]$), and number of imputations ($m \in [10, 50]$).

4.2 | Scenario 2

Next, we sought to determine if inclusion of the interaction term in the imputation model is detrimental when no true interaction is present, as in Scenario 2. Under Scenario 2, Figure 2 displays the distribution of bias across the replications, and Tables 3 and 4 display the coverage and average 95% CI width, respectively, for each imputation method. When there are no true interactions between the variables in their effect on the outcome, parametric imputation models with and without an interaction term (PMM-Naive and PMM-Int, respectively) perform very similarly in terms of accuracy of coefficient estimates (as assessed by mean bias and MAE), coverage, and average 95% CI width. In this scenario, CART has MAE approximately equal to that of the parametric methods but coverage below 95%. RF has the highest mean bias but lowest MAE. Figure 2 reveals that the distribution of bias across simulation replicates is much narrower for RF than

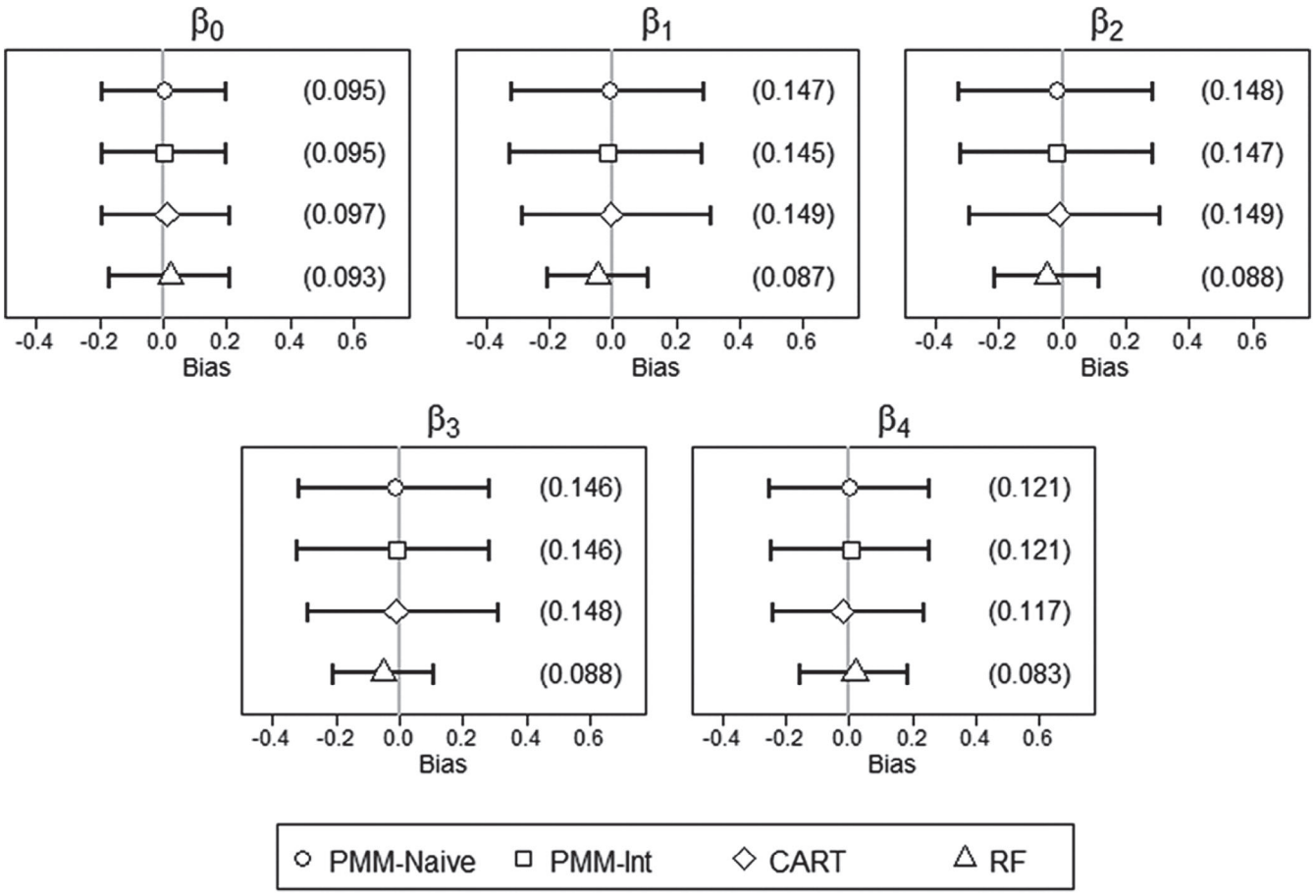

FIGURE 2 Scenario 2: Distribution of bias of the estimated coefficients from each imputation model (10,000 replications). Tick marks indicate the 5th and 95th percentiles. Mean absolute error is shown in parentheses.

TABLE 3 Scenario 2: Coverage (%) for each coefficient

Imputation model	β_0	β_1	β_2	β_3	β_4
PMM-Naive	94.7	94.1	94.5	95.1[a]	95.3[a]
PMM-Int	94.8	94.7	95.3[a]	95.2[a]	95.3[a]
CART	91.7	92.0	92.1	92.2	93.8
RF	92.6	99.2[a]	99.0[a]	99.1[a]	98.7[a]

Abbreviations: CART, classification and regression trees; PMM-Int, predictive mean matching with an interaction term in the imputation model; PMM-Naive, predictive mean matching using default *mice* settings; RF, random forests.
[a]Coverage ≥ 95%.

TABLE 4 Scenario 2: 95% confidence interval width for each coefficient

Imputation model	β_0	β_1	β_2	β_3	β_4
PMM-Naive	0.494	0.792	0.796	0.798	0.638
PMM-Int	0.496	0.801	0.802	0.805	0.643
CART	0.444	0.732	0.731	0.732	0.588
RF	0.431	0.621	0.621	0.622	0.545

Abbreviations: CART, classification and regression trees; PMM-Int, predictive mean matching with an interaction term in the imputation model; PMM-Naive, predictive mean matching using default *mice* settings; RF, random forests.

any other method, which is not captured in the mean bias. Thus, like Scenario 1, MAE appears to be a better measure of coefficient estimation accuracy than mean bias. When there is no true interaction in the model, RF has the lowest MAE, highest coverage, and narrowest 95% CIs.

5 | DATA APPLICATION

Of the 1737 subjects in the ADNI cohort at baseline, 14.3% were missing information on the normalized HV, 0.8% were missing the ADAS-Cog 13 score, and age was fully observed. Table 5 shows the results of this analysis after imputation by each method as well as using complete case analysis. As expected, the effect of HV-age interaction on cognition was estimated to be much smaller after imputation by PMM-Naive than after imputation by PMM-Int, CART, or RF. Although this interaction is not significant at the $\alpha = 0.05$ level using any of these imputation models, the notable difference in effect size between the imputation methods demonstrates the need to appropriately account for interactions in the imputation model for this analysis.

6 | DISCUSSION

In this simulation study comparing parametric and tree-based imputation models within the MICE algorithm, the compatible parametric model (PMM-Int) led to estimating the true interaction effect with lower mean bias and higher coverage than the tree-based methods. However, of the methods considered, RF had the highest coverage, lowest MAE, and narrowest 95% CIs for all main effects. CART imputation led to coverage below 95% for all effects and thus is not recommended for use. Application of these imputation methods to data from the ADNI cohort demonstrates that estimation of the interaction effect in real data can vary greatly based on the choice of imputation method, warranting the need to understand their relative performance.

Incorrectly specifying the parametric imputation model by omitting the true interaction term (PMM-Naive) resulted in estimating the interaction effect with the largest mean bias and smallest coverage. This result is also reported by Doove et al. who claim that tree-based methods preserve the interaction effect better than the standard parametric implementation of MICE.[12] However, the "standard" parametric method in Doove et al. omits the interaction term from the imputation model then goes on to estimate the interaction in the analysis model.[12] While including only main effects in

TABLE 5 Effects on ADAS-Cog 13 from linear regression analysis of the ADNI cohort at baseline

Exposure	Imputation model	Estimate (95% CI)	*P*-value
Normalized HV	Complete case	−0.619 (−0.666, −0.572)	<.001
	PMM-Naive	−0.608 (−0.653, −0.564)	<.001
	PMM-Int	−0.612 (−0.658, −0.567)	<.001
	CART	−0.610 (−0.654, −0.566)	<.001
	RF	−0.610 (−0.658, −0.563)	<.001
Age	Complete case	−0.099 (−0.145, −0.053)	<.001
	PMM-Naive	−0.087 (−0.130, −0.044)	<.001
	PMM-Int	−0.089 (−0.133, −0.044)	<.001
	CART	−0.088 (−0.131, −0.045)	<.001
	RF	−0.085 (−0.129, −0.041)	<.001
Normalized HV-age interaction	Complete case	−0.034 (−0.079, 0.010)	.126
	PMM-Naive	−0.016 (−0.061, 0.029)	.484
	PMM-Int	−0.039 (−0.090, 0.012)	.126
	CART	−0.032 (−0.075, 0.011)	.141
	RF	−0.028 (−0.071, 0.014)	.186

Abbreviations: ADNI, Alzheimer's Disease Neuroimaging Initiative; CART, classification and regression trees; CI, confidence interval; HV, hippocampal volume; PMM-Int, predictive mean matching with an interaction term in the imputation model; PMM-Naive, predictive mean matching using default *mice* settings; RF, random forests.

the imputation model may be the default implementation of MICE in R, it has been well-established that the imputation model must be at least as complex as the analysis model.[3,6-9] As such, a well-informed user of MICE would not implement the default approach used by Doove et al. when the final analysis model includes an interaction term. In this case, we argue that the "standard" parametric imputation model would include each effect in the final analysis model, including the interaction term. This approach, which we call PMM-Int, actually preserves the interaction effect best by providing higher coverage and lower mean bias of the interaction effect than CART and RF.

Selection of the predictors and the form of their effects to include in the final analysis model is an issue beyond the scope of this paper, and as such, our recommendations for the imputation model are based on assuming that the final analysis model is correctly specified. Future research could consider the interplay between misspecification of the final analysis model and selection of a MICE imputation model. Further, the focus of this paper is on scenarios of moderate complexity, that is, several correlated variables with a multiplicative interaction. Of course, data could be much larger and more complex, but this is also beyond the scope of this paper. There are even situations in which it could be difficult or impossible to include all necessary interactions in a parametric imputation model. For instance, Shah et al. use data from electronic health records with a large number of predictors and unknown correlation structure.[13] In these cases, RF imputation would be preferred since it has the ability to capture complex relationships such as interactions and nonlinearities without the need to specify an imputation model.

Based on our results, if interest lies primarily in the main effects, or if there are no true interaction effects between the predictors on the outcome, RF imputation would also be recommended over a parametric imputation model since it led to estimation of the main effects in our simulation with the lowest MAE, highest coverage, and narrowest 95% CIs. Even though RF imputation led to the largest mean bias for the main effects, it also had the smallest MAE. This means that even though RF imputation leads to biased estimates on average, the estimates after RF imputation still tend to be closer to the truth than other imputation methods. We argue that this makes RF a more accurate imputation method for estimating the main effects. While others have considered only the mean bias and empirical SD of coefficient estimates, our study adds the important and novel comparison of MAE of the coefficient estimates for each imputation method. Shah et al. demonstrated that RF imputation, as opposed to parametric imputation, in MICE produced more efficient parameter estimates, with efficiency defined by the empirical SD.[13] Our findings of low MAE of the coefficients when using RF imputation are highly related to this finding. However, the use of MAE demonstrates that the coefficient estimates not

only have low variability, but they are also close to the truth. We recommend the use of MAE or mean squared error to capture this phenomenon in future studies.

In some scenarios, RF imputation resulted in CIs with higher than nominal coverage. Despite the fact that these conservative CIs are wider than they need to be, they were still narrower than the CIs that resulted after performing parametric or CART imputation. Others have also noted the conservative nature of CIs after performing RF imputation in MICE, and future work could focus on refining RF imputation in MICE to produce even narrower CIs that still attain nominal coverage.[13]

Recommendations based on the results of this study are limited to similar types of data. Performance of the imputation methods may differ when the final analysis model takes on different forms such as logistic regression or survival analysis. Furthermore, tree-based imputation methods may perform poorly when the signal-to-noise ratio is low due to the increased difficulty in selecting the appropriate variable on which to make splits. Future work should extend to a more diverse range of data including noncontinuous variables with varying complexity of correlation structure.

We examined the performance of methods when the data are MAR. If data are missing not at random, expert knowledge must be used to inform the imputation procedure, which is beyond the scope of this paper. It is important to perform analyses examining the sensitivity of results to the assumptions made regarding the underlying missingness mechanism.[24]

In summary, this paper offers a fair comparison of parametric and tree-based imputation methods within the MICE algorithm by including a correctly specified parametric model for comparison. To our knowledge, this is the first paper to compare tree-based imputation in MICE to a parametric model that includes a true interaction effect. If interest lies primarily in estimation of main effects, we recommend utilizing RF as it leads to the lowest MAE, highest coverage, and narrowest 95% CIs for the main effects. If interest lies primarily in estimation of an interaction effect, there is a trade-off between RF and parametric imputation. Both methods have approximately equal MAE for estimating the interaction effect. PMM-Int has the highest coverage of the interaction effect, but it is at the expense of wide 95% CIs. Importantly, parametric imputation should only be utilized if there is enough information to ensure that all necessary interaction terms are included in the imputation model. If one can accept the reduction of coverage for the interaction effect, RF imputation is recommended as it does not require specification of the imputation model.

ACKNOWLEDGEMENTS

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; Euroimmun US, Inc; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio US; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

DATA AVAILABILITY STATEMENT

Restrictions apply to the availability of these data, which were used under license for this study. Data are available at adni.loni.usc.edu with the permission of the ADNI.

ORCID

Emily Slade https://orcid.org/0000-0002-1654-3822

REFERENCES

1. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: Wiley-IEEE; 1987.

2. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393.
3. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30(4):377-399.
4. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45(3):1-67.
5. Royston P, White IR. Multiple imputation by chained equations (MICE): implementation in Stata. *J Stat Softw*. 2011;45(4):1-20.
6. Tilling K, Williamson EJ, Spratt M, Sterne JAC, Carpenter JR. Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. *J Clin Epidemiol*. 2016;80:107-115.
7. Seaman SR, Bartlett JW, White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Med Res Methodol*. 2012;12:46.
8. Kenward MG, Carpenter J. Multiple imputation: current perspectives. *Stat Methods Med Res*. 2007;16:199-218.
9. Moons KG, Donders RA, Stijnen T, et al. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol*. 2006;59:1092-1101.
10. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods*. 2009;14(4):323-348.
11. Burgette LF, Reiter JP. Multiple imputation for missing data via sequential regression trees. *Am J Epidemiol*. 2010;172(9):1070-1076.
12. Doove LL, van Buuren S, Dusseldorp E. Recursive partitioning for missing data imputation in the presence of interaction effects. *Comput Stat Data Anal*. 2014;72:92-104.
13. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *Am J Epidemiol*. 2014;179(6):764-774.
14. Team RC. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2014 http://www.R-project.org/.
15. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med Res Methodol*. 2014;14:75.
16. von Hippel PT. How to impute interactions, squares, and other transformed variables. *Sociol Methodol*. 2009;39:265-291.
17. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software; 1984.
18. Therneau T, Atkinson B. rpart: Recursive partitioning and regression trees. R Package Version 4.1-13. 2018. https://CRAN.R-project.org/package=rpart.
19. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
20. Liaw A, Wiener M. Classification and regression by random forest. *R News*. 2002;2(3):18-22.
21. Alzheimer's Disease Neuroimaging Initiative. 2018. http://adni.loni.usc.edu/. Accessed August 1, 2016.
22. Mohs RC, Knopman D, Petersen RC, et al. Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's disease assessment scale that broaden its scope. The Alzheimer's disease cooperative study. *Alzheimer Dis Assoc Disord*. 1997;11(Suppl 2):S13-S21.
23. Voevodskaya O, Simmons A, Nordenskjold R, et al. The effects of intracranial volume adjustment approaches on multiple regional MRI volumes in healthy aging and Alzheimer's disease. *Front Aging Neurosci*. 2014;6:264.
24. White IR, Carpenter J, Evans S, Schroter S. Eliciting and using expert opinions about dropout bias in randomized controlled trials. *Clin Trials*. 2007;4:125-139.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Slade E, Naylor MG. A fair comparison of tree-based and parametric methods in multiple imputation by chained equations. *Statistics in Medicine*. 2020;39:1156–1166. https://doi.org/10.1002/sim.8468