Journal of Applied Probability and Statistics
2024, Vol. 19, No. 3, pp. 87-100

A COMPARATIVE ANALYSIS OF PARAMETRIC AND TREE-BASED IMPUTATION TECHNIQUES FOR MISSING DATA IN EPIDEMIOLOGICAL RESEARCH

Sara Javadi[1], Mohammad Mehdi Saber[2], Mehrdad Taghipour[3], Abdussalam Aljadani[4*], Mahmoud M. Mansour[5], Mohamed S. Hamed[6] and Haitham M. Yousof[7]

[1]*Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran*

[2]*Department of Statistics, Higher Education Center of Eghlid, Eghlid, Iran*

[3]*Department of, Faculty of Sciences, University of Qom, Qom, Iran*

[4*]*Department of Management, College of Business Administration in Yanbu, Taibah University, Al-Madinah, Al-Munawarah 41411, Kingdom of Saudi Arabia*
Email: ajadani@taibahu.edu.sa

[5]*Department of Management Information Systems, Yanbu, Taibah University, Yanbu 46421, Saudi Arabia*

[5,6,7]*Department of Statistics, Mathematics and Insurance, Faculty of Commerce, Benha University,Egypt;*

[6]*Department of Business Administration, Gulf Colleges, KSA;*

SUMMARY

Missing data presents a common challenge for researchers and data scientists, prompting the use of multiple imputations by chained equations in epidemiologic research. This method is highly favored for its practicality and reliable aptitude to generate unbiased effect estimates and make valid inferences. When employing multiple imputation by chained equations, researchers can choose from various imputation techniques, both parametric and nonparametric. Recent studies indicate that nonparametric tree-based methods may outperform parametric approaches, especially when dealing with interactions or nonlinear effects among predictor variables. Yet, these comparisons can be misleading if the parametric model does not include all effects present in the final analysis model, including interactions. Based on simulation results, it has been shown that integrating interactions into the parametric imputation model enhances its effectiveness in handling missing binary outcomes. While parametric imputation generally results in lower bias and slightly higher coverage probability for interaction effects, it tends to yield wider confidence intervals compared to tree-based methods, such as classification and regression trees. Furthermore, parametric imputation requires careful specification of the imputation model. Epidemiologists must be diligent in defining their imputation models within multiple imputations by chained equations. This study contributes to the field by offering a balanced comparison between parametric and tree-based imputation methods for data sets featuring binary outcomes.

Keywords and phrases: Imputation; Interaction; Missing Data; Missing at Random; Missing Completely at Random ; Regression Tree; Regression Analysis.

2020 Mathematics Subject Classification: Primary 62H10, secondary 62J12.

1 Introduction

Effectively addressing missing data is a major challenge for researchers in epidemiology, as it can stem from issues like survey nonresponse, data collection errors, or participant loss in longitudinal studies. If not managed properly, missing data can result in biased estimates and weakened statistical power. To maintain the integrity of analyses, researchers must use suitable methods to handle these gaps. Multiple Imputation (MI) has become a favored technique across various fields, including epidemiology and the social sciences. MI generates multiple plausible values for missing data based on observed data, analyzes each completed dataset, and combines the results for final estimates. It is particularly effective for missing data that is missing at random (MAR) or missing completely at random (MCAR), with the distinction between these two types being important for selecting the right handling method. MCAR refers to cases where the likelihood of missing data is unrelated to both observed and unobserved data. In other words, the missingness does not depend on any known or unknown information. For instance, if survey respondents fail to answer a question due to a system malfunction, this would be considered MCAR because the missingness is purely random and not related to any specific variables in the dataset. On the other hand, Missing at Random (MAR) refers to cases where the probability of data being missing depends on the observed data, but not on the unobserved data. For example, in a study investigating health outcomes, if patients with higher levels of income are less likely to report certain lifestyle factors, the missingness is dependent on income (an observed variable) but not on the unreported lifestyle data itself. MI is particularly effective in addressing both MAR and MCAR data because it utilizes observed patterns to generate plausible values for missing data, allowing for more accurate and unbiased results. Importantly, the MI technique enables researchers to account for the uncertainty introduced by missing data by generating multiple imputations. This not only provides more reliable estimates but also allows for valid statistical inference, as it incorporates variability across the multiple imputed datasets. By combining the results from these datasets using Rubin's rules, MI helps produce more precise estimates of the parameters of interest, along with proper measures of uncertainty. While MI is a powerful tool, its effectiveness relies on the assumption that the missing data mechanism falls within the MAR or MCAR categories. If data is missing in a non-random manner (i.e., Missing Not at Random, or MNAR), MI may not fully correct for biases. In such cases, alternative strategies, such as sensitivity analysis or model extensions, may be necessary to assess the potential impact of the missing data (see [1], [2], [3]).

The Multiple Imputation by Chained Equations (MICE) method, also known as fully conditional specification, is a commonly utilized technique for handling missing data [4], [5]. This algorithm, which is a Gibbs Sampler and Bayesian simulation approach, generates random draws from a predictive model while considering all other variables, and performs univariate imputations sequentially until convergence is reached. One of the main advantages of MICE is that it does not require the specification of a joint distribution for all variables. However, it is crucial to ensure

that the imputation model is compatible with the final analysis model to avoid biased parameter estimates and invalid inferences [6]. Default software for MICE typically includes each variable as a linear predictor in the imputation model without considering interactions or nonlinearities. This may lead to biased parameter estimations, especially in cases where there are interactions between variables [7], [8], [9]. In such scenarios where including interactions in a parametric imputation model is not feasible, recursive partitioning methods like Classification and Regression Trees (CART) and Random Forests (RF) can be incorporated into the MICE algorithm [10]. By utilizing tree-based methods, MICE can effectively handle situations with numerous predictors, small sample sizes, or highly correlated predictors.

The advantages of CART and RF lie in their nonparametric nature, eliminating the need for users to specify an imputation model. These models are also flexible in capturing interaction effects and non-linear relationships. In the MICE algorithm, missing values are imputed by constructing a tree with all other variables as predictors for incomplete variables. While tree-based methods may pose challenges in result interpretation, this is insignificant in the imputation process, where the main goal is to maintain the data structure for unbiased parameter estimates and valid inferences [10], [11], [12]. This study seeks to evaluate the effectiveness of partially parametric imputation methods in comparison to recursive partitioning methods in two simulation studies with binary outcomes as a measure of performance.

The main goal is to determine which imputation method within the MICE algorithm is more effective in preserving interaction effects. The paper is structured as follows: it begins by detailing the simulation designs, followed by the presentation of four MICE imputation methods used in the investigation of preserving interaction effects. Finally, the findings from the simulation studies are deliberated in the concluding section.

2 Materials and Methods

2.1 Scenarios

We carried out a simulation study to evaluate the efficacy of parametric and tree-based imputation methods, utilizing data sourced from two distinct models for comparison. For each pairing of the data generation model and imputation method, we executed a systematic process consisting of the following steps: data generation, deletion of observations according to missing at random (MAR) and missing completely at random (MCAR) criteria, imputation, logistic regression analysis, and assessment of Bias, Coverage Probability (CP), and Confidence Interval (CI) width for each coefficient. A high-quality imputation method should exhibit minimal Bias, a Cross-Validation Prediction of at least 95%, and precise Confidence Intervals. In addition, we computed both model-based and empirical Standard Error (SE) values – the former representing the average SE estimated across simulations, and the latter indicating the Standard Deviation of the estimates. We also determined the proportion of variance attributable to missing data ($\hat{\lambda}$), which is given in the following sections.

2.2 Data generation

A rigorous examination was carried out by creating 1000 simulated datasets, each consisting of 1000 observations Two logistic regression models were employed in this study, with one model incorporating an interaction term between two continuous variables, and the other model incorporating an interaction term between two binary variables. These specific models are given in equations 1 and 2, respectively.

$$\log it[P(Y_1 = 1)] = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + a_1 X_1 + a_2 X_2 + a_3 X_3 + a_4 X_4 + a_5 X_5 + a_6 X_1 X_2, \quad (2.1)$$

$$\log it[P(Y_2 = 1)] = \gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + b_5 X_5 + \gamma_3 Z_1 Z_2. \quad (2.2)$$

The dataset was generated from a multivariate normal distribution with a mean of 0 and a standard deviation of 1 for each of the five continuous variables X_1 to X_5. The correlation structure included correlations of 0.5 between X_1 and X_2, X_1 and X_3, and X_2 and X_3, and a correlation of 0.3 between X_4 and X_5, with a correlation of 0 between X_3 and X_4. Additionally, two binary variables, Z_1 and Z_2, were randomly drawn from a binomial distribution. Model parameters were set such that intercepts (β_0) were zero, β_1, β_2, γ_1, and γ_2 were 0.25, and a_1 to a_5, b_1 to b_5 were 0.5, with a_6 and γ_3 equal to 1.

2.3 Removal of observations

The removal of observations refers to the process of deleting or excluding certain data points from a dataset to improve the accuracy and reliability of the analysis. This may be necessary if the observations are deemed to be outliers, errors, or otherwise skewing the results. By removing these problematic observations, researchers can ensure that their conclusions are based on more representative and trustworthy data. However, it is important to be transparent about the criteria used for removal and to consider the potential impact on the overall validity of the findings. In this study, for each data set, we introduced missing values through missing at random (MAR) and missing at random (MCAR) mechanisms, which resulted in different proportions of missing data for the outcome variable: 10, 20, 30, 40, and 50 percent.

2.4 Imputation of missing data

The imputation of missing data is a common technique used in statistics to deal with missing or incomplete data in a dataset. This process involves estimating or predicting the missing values based on the information available in the dataset. Imputation can help to reduce bias, improve the accuracy of statistical analyses, and maintain the sample size of the dataset. There are various methods for imputing missing data, such as mean imputation, regression imputation, and multiple imputation. It is important to consider the assumptions and limitations of the imputation method chosen, as well as the potential impact on the results of the analysis. Overall, imputation is a valuable tool for handling missing data and ensuring the robustness of statistical analyses.

During the analysis, we conducted Multiple Imputations by Chained Equations (MICE) using the `mice` package in R version 3.1.0 [5], [13]. Missing values were imputed in each simulated dataset using various methods within the MICE algorithm, including Predictive Mean Matching (PMM)

with an interaction term (MICE-Interaction), Classification and Regression Trees (CART), and Random Forest (RF). In the following parts, we briefly describe each of the methods.

2.4.1 Predictive Mean Matching (PMM) method

To fill in missing values, the PMM method begins by analyzing a parametric model to locate cases with comparable predictive means, and then randomly selects an observed value from this group of similar subjects [14]. PMM is generally favored over traditional regression because it generates imputations from the available data, preserving data structure like skewness and preventing issues like unrealistic imputations [9], [15]. The standard implementation of PMM in the `mice` package includes only the main effects in the imputation model. More information on the conventional PMM method can be found in reference [12].

2.4.2 MICE-Interaction

The key distinction between MICE-Interaction and PMM (MICE-PMM) lies in the incorporation of an interaction term within the imputation model for MICE-Interaction. Specifically, the interaction term was included as an additional variable in the MICE-Interaction approach, with the default predictor matrix being utilized within the `mice` function [16], [17].

2.4.3 CART and MICE-CART

CART is a sophisticated tree-based imputation method that removes the necessity of defining an imputation model. Essentially, this technique constructs a decision tree by utilizing binary decision rules and a single predictor variable to split the data into two nodes, consequently reducing the variance of the outcome within each node. This process involves using predictors to classify subjects based on the outcome. The subgroups are determined by the optimal split, typically measured using the Gini index [18]. To prevent overfitting, the partitioning occurs until a specific criterion is met, such as a set number of observations in the final subsets [18], [19].

The dataset Y is divided into Y^{obs} and Y^{miss}, where Y^{obs} contains fully observed columns and Y^{miss} includes partially observed columns. The MICE-CART procedure follows these steps when imputing incomplete variables (with k representing the number of incomplete variables, and $\dot{Y}$ representing the current imputed data matrix Y) [11]:

1. The initial values of Y_j for $j = 1, \ldots, k$, are randomly sampled from Y_j^{obs}, creating a data matrix Z.
2. The CART model is applied to each missing response variable Y_j^{miss} by using the remaining variables in Z as predictors. Only subjects with observed values for Y_j^{miss} are included in this modeling process.
3. For individuals in the group Y_j^{miss}, the terminal node is identified based on the decision tree fitted in step 2. An observed value for Y_j^{miss} is randomly chosen from the subset within this node and utilized for imputation.
4. Steps 2 and 3 are repeated several times. This process is done for each variable that has missing values to obtain a complete data set.

5. Repeat steps 1-4 [12].

We utilized the `rpart` package to implement the Classification and Regression Trees (CART) algorithm in Multiple Imputation by Chained Equations (MICE), setting a complexity parameter of 10^{-4} and requiring a minimum of five observations at each terminal node [5], [20].

2.4.4 Random Forest (RF) and MICE-RF

The RF is a robust supervised machine learning algorithm known for its ability to construct recursively partitioned trees without the need for pruning [21]. Considered an enhancement of the Classification and Regression Trees (CART) algorithm, RF generates multiple trees by randomly selecting samples with replacements from the initial dataset. To address potential overfitting, RF incorporates a random selection of variables at each node to identify the optimal split. In the RF imputation method, multiple regression trees are built by randomly selecting B bootstrap samples from the original data. A subset of independent variables is randomly chosen for each split, and a tree is built using the CART algorithm.

When applying this method in the Multiple Imputation by Chained Equations (MICE) framework, B bootstrap samples are first generated, and a tree is fitted on each sample. Within each tree, leaves contain donors for the missing values of variable j.

For a missing value Y_j^{miss}, a random value is selected from the donors in the leaf associated with that value. This process is repeated for each variable with missing data to create a complete observation set. The remainder of the imputation process follows the same principles as the CART method in MICE [12].

In our study, we utilized the `mice` package to incorporate RF into the MICE framework [5]. We used 10 bootstrap samples and one-third of the predictors for each split candidate [5], [12]. While we did not vary the number of trees (n_{tree}), previous research indicates that the imputation quality is consistent between $n_{\text{tree}} = 10$ and $n_{\text{tree}} = 100$.

3 Regression Analysis

For each of the 1000 imputed datasets, a final analysis model was accurately fitted, and the combined results were synthesized using Rubin's rules [**?**].

3.1 Calculation of Bias, CP, and CI width

We calculated Bias as the difference between the estimated coefficient and the true value. The Confidence Probability (CP) is 0 if the 95% CI does not contain the true value, and 1 if it does. Average Confidence Interval Length (AL) measures the width of the 95% CI for each coefficient. We repeated Steps 1 to 5 10,000 times and reported the mean Bias, CP, and CI width across the replications for each imputation method. Additionally, we include the empirical standard error for each coefficient, which is the SD of the estimates across 1,000 simulation replications. The source code for this simulation can be found in the supplementary material [12].

4 Results

Scenario 1 (Interaction between two continuous variables)

Scenario 1 entails a study of moderate complexity involving a genuine interaction. Table 1 displays the Biases observed across 1000 replications, while Tables 2 and 3 present the Coverage Probability (CP) and average 95% confidence interval (CI) width for each imputation method. It is noted that MICE-Stratified produces estimates with a higher mean Bias for the interaction effect when compared to tree-based methods. Interestingly, MICE-Stratified shows the smallest CP values for interaction effects at each missing percentage, indicating its limitations under the MAR mechanism.

On the other hand, MICE-Interaction, which incorporates the interaction term in the imputation model, exhibits lower mean Bias and higher CP (>99.0%) for the interaction effect across all missing percentages compared to tree-based methods. Including the interaction term in the parametric imputation model is essential for accurate estimation of a true interaction effect, preventing significant Bias and ensuring high CP. For main effects, MICE-Interaction performs well with small Bias and CP >0.95 at all missing percentages. However, RF and CART at 10% and 20% missing percentages show higher mean Bias than MICE-Interaction but achieve at least 95% CP with narrower 95% CIs. The widths of the coefficient distributions in Table 3 help to explain this phenomenon, as they show that RF and CART methods have narrower Bias widths compared to MICE-Interaction when missing percentages are below 30%.

While the MICE-Interaction method maintains CP values >0.95, recursive partitioning methods fall below this threshold for 40% and 50% missing percentages. MICE-Stratified also shows CP values below 0.95 for most effects. Instead of relying solely on mean Bias, it is crucial to consider measures such as standard error of the coefficient and empirical standard error to evaluate precision accurately. Additionally, for interaction between continuous variables, CART exhibits smaller standard errors than MICE-Interaction and MICE-RF.

Following Emily Slade's research [9], we explored tree-based tuning parameters and found that default values are appropriate for CART. Reducing the "minbucket" values resulted in a slight decrease in mean bias, however, we found the default value of 5 to be a suitable choice. We chose not to adjust the number of trees (`ntree`) in our analysis, as previous simulations indicated that imputation quality remains consistent between `ntree` values of 10 and 100. Additionally, the performance of decision trees (CART) and random forests (RF) did not show significant improvement when modifying tuning parameters. Therefore, according to [9], we present our results using the default tuning parameters, where

$$\texttt{cp} = 10^{-4}, \texttt{minbucket} = 5, \texttt{mtry} = 1$$

Scenario 2 (Interaction between two binary variables)

We then proceeded to examine the interaction term between two binary variables in the imputation model, specifically in Scenario 2. In Scenario 2, Table 3 presents the CP width corresponding to each assignment method. Overall, the Bias values for the MICE-Interaction method were observed to be lower compared to other methods. Recursive partitioning methods tended to underestimate

all coefficients except for one of the main effects. As expected, the mean Bias values generally increased with the proportion of missing values, although there were some exceptions. Particularly, for interactions, recursive partitioning methods generated Bias values exceeding 0.05. Significant bias values were observed for recursive partitioning methods, exceeding 0.40 in cases of missing data at percentages of 40% and 50%. In terms of estimating the main effects, the MICE-Interaction method exhibited the least Bias among the four methods. Among the two recursive partitioning methods, CART showed lower Bias values for the main coefficients. The CP values for the MICE-Stratified method were consistently low across most scenarios, especially under the MCAR mechanism. On the other hand, the CP values for the MICE-Interaction method were at least 0.95. For missing percentages above 40%, the CP values for recursive partitioning methods fell below 0.95. Despite the MICE-Stratified method having the shortest CI length, lowest standard error, and ratio of variations, it was not deemed an acceptable model due to the values of Bias, relative Bias, and CP.

5 Numerical Illustration

Two recursive partitioning techniques, namely CART and RF, were compared with the compatible parametric model, MICE-Interaction, to assess their ability to preserve interactions and main effects. Simulated datasets were used, featuring a binary outcome and a mix of continuous and binary predictors. The results showed that, across all missing percentages, MICE-Interaction outperformed the tree-based methods in estimating the true interaction effect, exhibiting lower mean Bias and higher CP. When the true interaction term was omitted from the parametric imputation model, it led to the largest mean Bias and the smallest CP [9].

Recent studies have indicated that the parametric imputation model performs better than nonparametric methods like CART and RF [9] when the imputation model aligns with the analysis model. However, there is a lack of research comparing these methods in binary responses with both binary and continuous predictors. Our findings suggest that the nature of the interaction term significantly impacts the performance of imputation methods. Specifically, MICE-Interaction was shown to preserve the interaction effect better when two continuous variables influenced the interaction term, displaying higher CP and lower mean Bias compared to CART and RF.

In situations where the interaction involves two continuous variables, when the primary focus is on main effects, or when no true interaction effects exist between predictors and outcomes, RF and CART imputations may be preferred over parametric imputation. While CART imputation had a larger mean Bias for main effects compared to MICE-Interaction, it also yielded the narrowest 95% CIs. This indicates that CART imputation, despite its bias, produces estimates closer to the truth compared to other methods, making it more accurate for estimating main effects. Additionally, RF imputation was found to generate more efficient parameter estimates than parametric imputation in MICE.

However, MICE-Interaction imputation at times resulted in CIs with higher than nominal CP, though still narrower than those from parametric or RF imputation. Future research could aim to refine MICE imputations further to produce narrower, yet still reliable, CIs. This study demonstrates several key strengths, such as the comprehensive comparison of tree-based and parametric

imputation methods within the MICE framework. Furthermore, the use of advanced simulation techniques to generate missing data allows for a thorough evaluation of the effectiveness of these imputation methods under various missing data mechanisms. Some limitations include the focus on one MAR condition, the absence of true data for comparison, and the use of a fixed sample size. Future research should explore the misspecification of final models, consider more complex and larger datasets, and address MNAR data with expert knowledge. Sensitivity analysis of underlying missingness mechanisms and parameter tuning in imputation models may also be beneficial for enhancing imputation performance.

In this study, we conducted a comprehensive comparison of parametric and tree-based imputation methods in the MICE algorithm, using a well-defined parametric model for a thorough evaluation. To our knowledge, this is the first study to examine tree-based imputation in MICE against a parametric model with a true interaction effect and a binary outcome. Our findings show that MICE-Interaction is the preferred method for estimating interactions between two continuous variables, as it has the lowest Bias, highest CP, and most precise 95% CIs for interaction effects across varying levels of missing data. We observed a trade-off between tree-based and parametric imputations for estimating main effects, particularly at missing percentages below 30%. Despite wider 95% CIs, MICE-Interaction outperformed other methods in terms of CP for interaction effects. In cases involving interaction between two binary variables, MICE-Interaction is advised for its low Bias and acceptable CP for interaction effects across all levels of missing data. It is important to note that parametric imputation should only be used when there is sufficient information to ensure all necessary interaction terms are incorporated into the imputation model.

Based on Table 1, Table 2, Table 3, it is noted that the MI performs the best overall, with low bias, high coverage probability, and acceptable confidence interval widths across different levels of missingness. It is the most robust method among those tested. The MCR offers a reasonable trade-off but suffers from increasing bias and reduced coverage as missingness increases. The MRF shows high bias and poor coverage at higher levels of missingness, indicating that it may not be suitable for datasets with substantial missing data, especially when dealing with interactions. The MS performs the worst, with significant bias, poor coverage, and potentially misleading narrow confidence intervals. This method should generally be avoided, especially in cases of high missingness. These observations provide clear guidance on the strengths and limitations of each method in handling missing data. The results suggest that while MI is the most reliable approach, careful consideration is needed when choosing a method, particularly in the presence of high levels of missing data and complex variable interactions.

6 Conclusions

Missing data presents a significant challenge in research, prompting the adoption of multiple imputations by chained equations (MI) in epidemiologic studies due to its practicality and ability to yield unbiased effect estimates. Recent investigations suggest that nonparametric tree-based methods may outperform parametric approaches, especially with interactions or nonlinear effects among predictors; however, this can be misleading if the parametric model lacks necessary interac-

tions. Simulation results indicate that incorporating these interactions enhances the effectiveness of parametric imputation, which generally shows lower bias and slightly better coverage probability for interaction effects but produces wider confidence intervals compared to tree-based methods. The study emphasizes that MI consistently offers the best performance across various missingness levels, maintaining low bias and high coverage probability, while the MCR method struggles as missingness increases, and the MRF method shows high bias and poor performance in complex scenarios. The MS method is the least effective, with significant bias and unreliable inferences, suggesting it should be avoided in cases with substantial missing data. Overall, the findings underscore MI's robustness and the importance of careful method selection in data with high missingness and complex interactions.

References

[1] Sterne, J.A., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, **338**.

[2] Rubin, D. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons Inc.

[3] Lee, K.J., Carlin, J.B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*, **171**(5), 624-632.

[4] Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., Solenberger, P. (2001). A multivariate technique for multiplying imputing missing values using a sequence of regression models. *Survey Methodology*, **27**(1), 85-96.

[5] Buuren, S.V., Groothuis-Oudshoorn, K. (2010). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 1-68.

[6] Gelfand, A.E., Smith, A.F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**(410), 398-409.

[7] Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, **16**(3), 219-242.

[8] Seaman, S.R., Bartlett, J.W., White, I.R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Medical Research Methodology*, **12**(1), 46.

[9] Slade, E., Naylor, M.G. (2020). A fair comparison of tree-based and parametric methods in multiple imputation by chained equations. *Statistics in Medicine*, **39**(8), 1156-1166.

[10] Strobl, C., Malley, J., Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, **14**(4), 323.

[11] Burgette, L.F., Reiter, J.P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, **172**(9), 1070-1076.

[12] Doove, L.L., Van Buuren, S., Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, **72**, 92-104.

[13] Team, R.C. (2013). R: A language and environment for statistical computing.

[14] Little, R.J. (1988). Missing data adjustments in large surveys. *Journal of Business & Economic Statistics*, **6**(3), 287-296.

[15] Morris, T.P., White, I.R., Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, **14**(1), 1-13.

[16] Von Hippel, P.T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, **39**(1), 265-291.

[17] Seaman, S.R., Bartlett, J.W., White, I.R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Medical Research Methodology*, **12**(1), 1-13.

[18] Breiman, L., Friedman, J.H., Olshen, R.A., et al. (1984). Classification and Regression Trees. Boca Raton, FL: Chapman and Hall/CRC.

[19] Friedman, J., Hastie, T., Tibshirani, R. (2001). The Elements of Statistical Learning: Springer Series in Statistics. New York, NY, USA.

[20] Therneau, T., Atkinson, B., Ripley, B. (2018). Part: Recursive Partitioning and Regression Trees. R package.

[21] Hayes, T., Usami, S., Jacobucci, R., McArdle, J.J. (2015). Using Classification and Regression Trees (CART) and random forests to analyze attrition: Results from two simulations. *Psychology and Aging*, **30**(4), 911.

Table 1: Scenario 1: Bias of each coefficient

Variable	Model	MAR					MCAR				
		10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
$\boldsymbol{X_1}$	MCR	-0.022	-0.031	-0.056	-0.067	-0.093	-0.015	-0.031	-0.045	-0.066	-0.079
	MI	-0.001	0.005	0.001	0.005	0.002	0.001	0.000	0.001	0.002	0.003
	MRF	-0.054	-0.096	-0.142	-0.175	-0.211	-0.052	-0.099	-0.138	-0.178	-0.209
	MS	-0.116	-0.192	-0.255	-0.304	-0.346	-0.118	-0.204	-0.267	-0.317	-0.357
$\boldsymbol{X_1 * X_2}$	MCR	-0.057	-0.112	-0.168	-0.228	-0.281	-0.053	-0.110	-0.169	-0.230	-0.289
	MI	0.000	0.001	0.006	0.006	0.006	0.003	0.006	0.002	0.009	0.008
	MRF	-0.119	-0.229	-0.323	-0.413	-0.494	-0.121	-0.226	-0.324	-0.413	-0.496
	MS	-0.231	-0.399	-0.522	-0.620	-0.703	-0.235	-0.405	-0.536	-0.638	-0.725
$\boldsymbol{X_2}$	MCR	-0.018	-0.041	-0.057	-0.072	-0.093	-0.015	-0.032	-0.051	-0.065	-0.084
	MI	0.001	-0.003	0.002	0.002	0.000	0.000	0.004	-0.002	0.000	0.002
	MRF	-0.051	-0.103	-0.140	-0.177	-0.213	-0.054	-0.099	-0.143	-0.179	-0.213
	MS	-0.115	-0.200	-0.258	-0.306	-0.348	-0.120	-0.206	-0.269	-0.322	-0.361
$\boldsymbol{X_3}$	MCR	-0.014	-0.027	-0.036	-0.061	-0.068	-0.014	-0.027	-0.042	-0.054	-0.072
	MI	0.002	0.000	0.007	0.004	0.010	-0.001	0.002	0.002	0.001	0.004
	MRF	-0.040	-0.077	-0.105	-0.139	-0.168	-0.041	-0.074	-0.107	-0.136	-0.164
	MS	-0.087	-0.156	-0.208	-0.259	-0.305	-0.086	-0.151	-0.209	-0.262	-0.307
$\boldsymbol{X_4}$	MCR	-0.021	-0.036	-0.052	-0.068	-0.090	-0.011	-0.024	-0.036	-0.050	-0.061
	MI	-0.002	-0.001	0.001	0.002	-0.001	0.001	0.002	0.002	0.001	0.006
	MRF	-0.047	-0.086	-0.119	-0.154	-0.190	-0.038	-0.074	-0.107	-0.138	-0.166
	MS	-0.111	-0.182	-0.229	-0.272	-0.307	-0.082	-0.151	-0.208	-0.260	-0.304
$\boldsymbol{X_5}$	MCR	-0.020	-0.038	-0.058	-0.075	-0.091	-0.011	-0.022	-0.036	-0.052	-0.060
	MI	-0.002	-0.003	-0.003	-0.003	-0.006	0.002	0.003	0.002	0.002	0.005
	MRF	-0.046	-0.086	-0.125	-0.159	-0.192	-0.038	-0.073	-0.106	-0.138	-0.163
	MS	-0.113	-0.180	-0.234	-0.274	-0.310	-0.082	-0.147	-0.208	-0.261	-0.305
$\boldsymbol{Z_1}$	MCR	-0.022	-0.040	-0.065	-0.079	-0.103	-0.023	-0.046	-0.065	-0.090	-0.109
	MI	-0.003	-0.003	-0.004	-0.005	0.002	0.001	0.000	-0.001	-0.002	-0.007
	MRF	-0.026	-0.048	-0.071	-0.088	-0.108	-0.024	-0.046	-0.065	-0.090	-0.111
	MS	-0.043	-0.073	-0.105	-0.130	-0.151	-0.042	-0.077	-0.102	-0.129	-0.153
$\boldsymbol{Z_2}$	MCR	-0.021	-0.040	-0.064	-0.073	-0.108	-0.020	-0.044	-0.058	-0.087	-0.098
	MI	-0.002	0.002	-0.006	0.005	-0.003	-0.001	0.000	0.006	-0.002	0.004
	MRF	-0.025	-0.052	-0.071	-0.097	-0.123	-0.022	-0.050	-0.071	-0.095	-0.118
	MS	-0.048	-0.077	-0.111	-0.136	-0.157	-0.044	-0.085	-0.107	-0.137	-0.158

Table 2: Scenario 1: CP of each coefficient.

Variable	Model	MAR					MCAR				
		10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
$\boldsymbol{X_1}$	MCR	1.000	0.999	0.972	0.928	0.844	1.000	0.994	0.972	0.922	0.843
	MI	1.000	1.000	1.000	0.996	0.980	1.000	1.000	0.998	0.993	0.983
	MRF	1.000	0.988	0.920	0.787	0.649	1.000	0.980	0.893	0.779	0.666
	MS	0.908	0.434	0.141	0.052	0.019	0.888	0.393	0.121	0.037	0.016
$\boldsymbol{X_1 * X_2}$	MCR	0.997	0.949	0.766	0.560	0.391	0.998	0.951	0.780	0.550	0.409
	MI	1.000	1.000	1.000	0.997	0.991	1.000	1.000	1.000	0.997	0.989
	MRF	0.987	0.619	0.214	0.063	0.014	0.997	0.668	0.236	0.067	0.018
	MS	0.338	0.006	0.000	0.000	0.000	0.311	0.004	0.000	0.000	0.000
$\boldsymbol{X_2}$	MCR	1.000	0.996	0.973	0.908	0.873	1.000	0.994	0.970	0.928	0.858
	MI	1.000	1.000	0.997	0.996	0.992	1.000	1.000	0.999	0.996	0.988
	MRF	1.000	0.982	0.914	0.778	0.659	1.000	0.988	0.901	0.814	0.644
	MS	0.909	0.400	0.153	0.043	0.016	0.891	0.377	0.106	0.038	0.011
$\boldsymbol{X_3}$	MCR	1.000	0.997	0.972	0.936	0.886	1.000	0.999	0.983	0.927	0.858
	MI	1.000	1.000	0.998	0.994	0.987	1.000	1.000	0.997	0.996	0.996
	MRF	1.000	0.996	0.966	0.882	0.787	1.000	0.996	0.961	0.907	0.789
	MS	0.979	0.640	0.297	0.118	0.044	0.979	0.671	0.294	0.111	0.033
$\boldsymbol{X_4}$	MCR	1.000	0.987	0.956	0.868	0.770	1.000	0.990	0.979	0.927	0.894
	MI	1.000	0.999	0.999	0.987	0.988	1.000	1.000	0.999	0.997	0.983
	MRF	1.000	0.971	0.911	0.723	0.558	1.000	0.987	0.920	0.821	0.707
	MS	0.839	0.273	0.092	0.032	0.012	0.967	0.523	0.178	0.053	0.012
$\boldsymbol{X_5}$	MCR	1.000	0.988	0.931	0.854	0.781	1.000	0.994	0.977	0.940	0.879
	MI	1.000	1.000	0.997	0.990	0.989	1.000	1.000	0.998	0.996	0.982
	MRF	1.000	0.977	0.846	0.701	0.558	1.000	0.992	0.939	0.802	0.703
	MS	0.826	0.292	0.071	0.024	0.008	0.962	0.533	0.151	0.047	0.006
$\boldsymbol{Z_1}$	MCR	1.000	1.000	0.993	0.984	0.961	1.000	0.999	0.991	0.980	0.962
	MI	1.000	1.000	0.999	0.996	0.986	1.000	1.000	1.000	0.988	0.984
	MRF	1.000	0.999	0.999	0.993	0.990	1.000	1.000	0.998	0.991	0.986
	MS	1.000	0.992	0.954	0.930	0.909	0.999	0.989	0.956	0.926	0.899
$\boldsymbol{Z_2}$	MCR	1.000	1.000	0.997	0.981	0.968	1.000	0.999	0.997	0.986	0.969
	MI	1.000	1.000	0.998	0.993	0.990	1.000	1.000	1.000	0.994	0.987
	MRF	1.000	1.000	0.999	0.994	0.989	1.000	1.000	0.998	0.996	0.990
	MS	1.000	0.989	0.954	0.914	0.890	1.000	0.976	0.952	0.931	0.891

Table 3: Scenario 1: 95% CI width of each coefficient.

Variable	Model	MAR					MCAR				
		10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
$\boldsymbol{X_1}$	MCR	0.421	0.437	0.447	0.461	0.473	0.424	0.440	0.455	0.467	0.478
	MI	0.435	0.467	0.515	0.566	0.644	0.438	0.476	0.520	0.587	0.670
	MRF	0.433	0.452	0.472	0.481	0.496	0.435	0.459	0.476	0.487	0.502
	MS	0.381	0.367	0.362	0.363	0.370	0.383	0.371	0.366	0.367	0.371
$\boldsymbol{X_1 * X_2}$	MCR	0.460	0.481	0.490	0.500	0.499	0.465	0.482	0.498	0.504	0.514
	MI	0.467	0.501	0.547	0.610	0.674	0.472	0.513	0.566	0.627	0.738
	MRF	0.484	0.514	0.517	0.516	0.510	0.489	0.516	0.526	0.525	0.525
	MS	0.387	0.358	0.341	0.329	0.320	0.391	0.364	0.345	0.333	0.326
$\boldsymbol{X_2}$	MCR	0.422	0.436	0.450	0.464	0.476	0.425	0.438	0.456	0.464	0.479
	MI	0.434	0.467	0.512	0.578	0.647	0.438	0.476	0.528	0.584	0.673
	MRF	0.432	0.454	0.471	0.488	0.497	0.436	0.459	0.472	0.490	0.502
	MS	0.381	0.367	0.362	0.362	0.369	0.383	0.370	0.365	0.367	0.372
$\boldsymbol{X_3}$	MCR	0.392	0.405	0.418	0.428	0.444	0.393	0.408	0.417	0.433	0.442
	MI	0.403	0.435	0.476	0.532	0.619	0.405	0.440	0.487	0.546	0.613
	MRF	0.399	0.417	0.437	0.456	0.478	0.401	0.420	0.441	0.457	0.471
	MS	0.361	0.352	0.351	0.354	0.363	0.363	0.357	0.356	0.358	0.365
$\boldsymbol{X_4}$	MCR	0.343	0.354	0.365	0.376	0.385	0.341	0.353	0.365	0.374	0.385
	MI	0.355	0.386	0.431	0.478	0.549	0.351	0.382	0.419	0.468	0.532
	MRF	0.352	0.371	0.387	0.395	0.412	0.347	0.364	0.379	0.397	0.416
	MS	0.311	0.304	0.304	0.306	0.312	0.314	0.308	0.305	0.310	0.315
$\boldsymbol{X_5}$	MCR	0.342	0.355	0.364	0.375	0.384	0.341	0.354	0.365	0.374	0.386
	MI	0.354	0.387	0.432	0.484	0.553	0.352	0.381	0.422	0.469	0.533
	MRF	0.353	0.370	0.383	0.397	0.409	0.347	0.366	0.381	0.395	0.413
	MS	0.311	0.304	0.303	0.306	0.314	0.314	0.307	0.305	0.309	0.314
$\boldsymbol{Z_1}$	MCR	0.618	0.639	0.658	0.675	0.691	0.621	0.642	0.661	0.679	0.706
	MI	0.636	0.689	0.760	0.850	0.967	0.642	0.698	0.768	0.860	0.992
	MRF	0.629	0.659	0.687	0.716	0.750	0.632	0.664	0.702	0.725	0.764
	MS	0.576	0.566	0.565	0.573	0.587	0.580	0.571	0.572	0.581	0.595
$\boldsymbol{Z_2}$	MCR	0.620	0.637	0.657	0.675	0.694	0.621	0.642	0.661	0.681	0.706
	MI	0.640	0.689	0.756	0.848	0.948	0.643	0.702	0.774	0.851	0.977
	MRF	0.630	0.657	0.687	0.715	0.752	0.635	0.664	0.700	0.730	0.758
	MS	0.576	0.566	0.566	0.574	0.588	0.580	0.572	0.572	0.580	0.596