Parametric and non-parametric methods for Missing Value Imputation in Observational Studies

Abstract

Observational studies often suffer from missing data, which can lead to biased results and reduced statistical power. This study explores the efficacy of parametric (Multiple Imputation - MI, and Full Information Maximum Likelihood - FIML) and non-parametric (K-Nearest Neighbors - KNN, and Random Forest - RF) imputation methods in handling missing data within observational studies. Monte Carlo simulations were conducted across different structures of Directed Acyclic Graphs (DAGs) with binary treatments and continuous outcomes, investigating the effects of sample size, missingness rate, and missing data mechanism (MCAR and MNAR) on model estimation. The findings reveal that no single imputation technique consistently excels across all scenarios. Parametric methods perform well in high sample settings even in MNAR conditions, while non-parametric methods can be considered as an option for better adaptability in low sample settings, particularly under MNAR conditions. This study also finds that in smaller sample sizes, it is crucial to conduct sensitivity analyses and report the estimates rather than relying heavily on imputation methods. This study highlights the importance of choosing the appropriate imputation method based on specific study conditions and assumptions, pointing out the strengths and limitations of both parametric and non-parametric approaches. Future research aims to extend these analyses to scenarios with non-linear relationships and non-parametric outcome regression models.

Introduction

In observational studies, the issue of missing data is a common challenge that can significantly impact the validity of the research findings. Traditionally, researchers have relied on case-wise deletion to manage missing entries, a method where any case with a missing value is excluded from analysis. This approach can lead to biased results, especially if the missing data mechanism is not Missing Completely at Random (MCAR), where data is missing irrespective of both observed and unobserved data.

More commonly in biomedical research, data is Missing Not at Random (MNAR), where the missingness is related to the unobserved data itself, making case-wise deletion inappropriate and potentially misleading. The prevalence of MNAR in this domain necessitates more sophisticated imputation techniques to properly handle missing data without introducing bias.

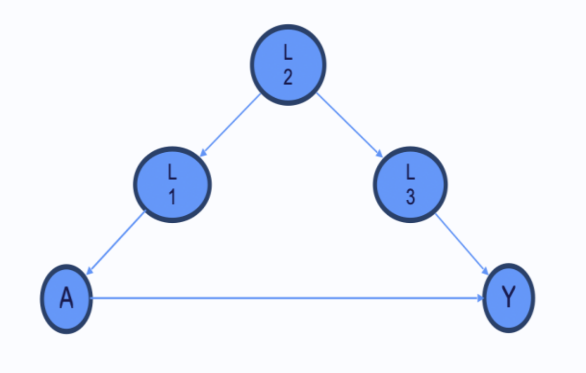
Several advanced imputation methods are used to address these issues. Full Information Maximum Likelihood (FIML) and Multiple Imputation (MI) are parametric methods; FIML makes use of all available data to estimate model parameters under the assumption that the data are MCAR or Missing at Random (MAR), while MI creates multiple complete datasets under a specified model, analyzing each to account for the uncertainty introduced by the missing data. On the non-parametric front, K-Nearest Neighbors (KNN) and Random Forest (RF) imputation offer flexibility by imputing missing values based on the resemblance of observed data points, capturing non-linear relationships and interactions without assuming a particular data distribution.

This study aims to explore the performance of these imputation methods across different Directed Acyclic Graph (DAG) scenarios, which represent structured models of the relationships between treatment, confounders, and outcomes in observational studies. The DAGs examined vary in complexity, from simple structures with direct paths to more intricate configurations involving mediator variables. By analyzing these diverse scenarios, we seek to determine how each imputation method handles the subtleties of structured data in the presence of missing values, providing insights that could guide researchers in selecting the most appropriate imputation strategy for their specific study conditions.

Scenarios of the DAGs investigated

Scenario 1

In the provided DAG, the variable **A** represents a treatment intervention, and **L1, L2,** and **L3** are confounders that influence both the treatment and the outcome **Y**.



Examples of Missing Data Mechanisms in this DAG

MCAR (Missing Completely at Random):

A study investigates the effectiveness of a new weight loss medication on patients with obesity.

Variables:

- A - New weight loss medication
- L1 - Physical activity level
- L2 - Socioeconomic status
- L3 - Diet quality
- Y - Change in body weight

Missingness in **L1** (Physical activity level) occurs independently of any other factors, such as a sudden loss of data due to errors in data handling or system failures during the collection process. This is expressed mathematically as $P(RL1=1/A,L1,L2,L3,Y)=P(RL1=1) =P(RL1 =1)$, *where* $RL1=1$ indicates missing data.

MNAR (Missing Not at Random):

A study evaluates the effectiveness of a new antidepressant on patients with major depressive disorder.

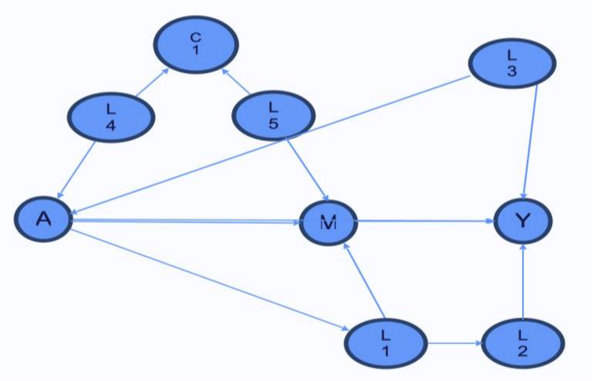
Variables:

- A - New antidepressant medication
- L1 - Severity of depression symptoms at baseline
- L2 - Duration of major depressive disorder
- L3 - Presence of co-occurring anxiety disorders
- Y - Reduction in depression symptoms

Missingness in **L1** (Severity of depression symptoms at baseline) is influenced by the symptoms' severity. For instance, patients with more severe symptoms might be less likely to follow up or provide complete data due to their condition. This dependency is modeled as $P(L1missing/L1observed) = pnorm(L1observed)$, indicating that the higher the observed severity, the more likely the data is to be missing.

Scenario 2

This DAG presents a more intricate model involving multiple confounders, a treatment variable, mediator, and an outcome. The DAG illustrates relationships that could represent complex scenarios often seen in health-related studies. **A** represents a medical treatment or an intervention. **M** represents a mediator variable, such as an immediate effect or measurement influenced by the treatment. **L4** and **L5** are treatment mediator confounders that are blocked by collider **C1**. **L3** is the direct confounder between treatment and the outcome. **L1** and **L2** are mediator outcome confounders. **L1** being directly affected by treatment **A**.



Examples of Missing Data Mechanisms in this DAG

MCAR (Missing Completely at Random):

A study evaluating the effectiveness of cardiac medications in improving long-term cardiac health.

Variables:

- **A** - Prescription of Cardiac Medication
- **L1** - Blood Pressure Level
- **L2** - Cholesterol Level
- **L3** - Genetic Predisposition to Cardiovascular Disease
- **L4** - BMI
- **L5** - Age
- **M** - Reduction in Systolic Blood Pressure
- **C1** - Cardiovascular Risk Score
- **Y** - Cardiac Health Score

Missingness in **L1** (Blood Pressure Level) occurs randomly due to data entry errors or data loss during collection, unrelated to any other variable or its own value. This missingness pattern represents a scenario where the missing data does not depend on the observed or unobserved data, characterizing a true MCAR situation.

MNAR (Missing Not at Random):

A study assessing the efficacy of mental health interventions, specifically cognitive behavioral therapy, in improving mood and mental health outcomes.

Variables:

- **A** - Enrollment in a Mental Health Program
- **L1** - Severity of Depressive Symptoms
- **L2** - Anxiety Level
- **L3** - Genetic Predisposition to Mental Health Disorders
- **L4** - Physical Activity Level
- **L5** - Social Support Level
- **M** - Improvement in Mood Scores
- **C1** - Mental Well-being Index
- **Y** - Overall Mental Health Outcome

Missingness in **L1** (Severity of Depressive Symptoms) is influenced by the severity of the symptoms themselves, where patients with more severe symptoms may avoid follow-up due to stigma or the debilitating nature of their depression. This results in a missing data pattern where the likelihood of missing data increases with the severity of the observed symptoms, typical of an MNAR scenario.

Imputation Methods:

Multiple Imputation Method

Multiple Imputation involves creating several independent datasets by replacing missing values with estimates and analyzing each data set separately. This technique considers the uncertainty inherent in the imputation process by using random variations in the predictions, thus generating multiple complete datasets. These datasets are then analyzed using standard procedures, and the results are combined to produce overall estimates.

This method is particularly effective assuming data are Missing at Random (MAR) or Missing completely at Random (MCAR), allowing for statistically valid inferences that acknowledge missing data.

Full Information Maximum Likelihood (FIML) Method

The Full Information Maximum Likelihood method capitalizes on all available data, whether fully or partially observed, to derive parameter estimates that optimize the model's likelihood function. This function measures how well the statistical model aligns with the observed data. Research has demonstrated that FIML provides unbiased estimates when dealing with normally distributed data under the MAR or MCAR mechanism.

Random Forest (RF) Imputation Method

Random Forest imputation, a robust, nonparametric machine learning approach, predicts missing values using a random forest trained on complete cases. It iteratively updates and uses the newly estimated data to refine further predictions until convergence or a set number of iterations is reached. This method is adaptable to various data types and complexities but is computationally demanding.

K-Nearest Neighbors (KNN) Imputation Method

K-Nearest Neighbors imputation estimates missing values by identifying and averaging the nearest observed data points using distance metrics like Euclidean distance. This method does not rely on distribution assumptions and is sensitive to the number of neighbors 'k' used; inappropriate 'k' values can either dilute the accuracy or overfit the data.

Simulations

Based on the structures of the Directed Acyclic Graphs (DAGs) detailed earlier, we have conducted simulations to generate data for each scenario. In DAG 1, the data generation process incorporates a population parameter for the Average Treatment Effect (ATE) set at 1.5. For DAG 2, the simulations involve population parameters with the Natural Direct Effect (NDE) set at 0.9 and the Natural Indirect Effect (NIE) set at 0.6. These parameters reflect the predefined theoretical relationships within each respective DAG configuration.

In our simulations, we manipulated several parameters across two DAG scenarios including sample size, rate of missingness, and the mechanism of missing data. We varied the sample sizes to include small (N = 50) and large (N = 1000) groups to assess the impact of sample size on imputation effectiveness. The rate of missingness ranged from 5 % to 30 % to observe the performance of imputation methods under various levels of missingness. Two missing data mechanisms were employed: Missing Completely at Random (MCAR), where missingness in the confounder L1 was introduced at random, and Missing Not at Random (MNAR), where missingness was dependent on the cumulative distribution function of the observed values of L1.

For each testing scenario, 100 datasets were generated. Each of these datasets was then analyzed using four distinct methods for managing missing data. The `lavaan` R package facilitated the use of Full Information Maximum Likelihood (FIML) for our analyses. Additionally, Random Forests and K-Nearest Neighbors (KNN) imputation techniques were employed using the `missForest` and `VIM` packages respectively. For multiple imputation, we utilized the `mice` package, thereby ensuring a thorough and varied approach to handling missing data throughout our simulations. The detailed code is provided on my GitHub site https://github.com/Blueeyes27/Missing-Value-Imputation-in-Causal-Inference/tree/main

Results

To evaluate the effectiveness of the four methods, we calculated the relative bias, which indicates how much the estimated values deviate from the true population parameters on an average across 100 simulations. Additionally, we assessed the coverage rate of the confidence intervals, determining the likelihood that the true values fall within the 95% confidence intervals across 100 simulations. The relative bias for the estimated average treatment effect in the first scenario is shown in Figure 1, while Figure 2 displays the coverage probabilities for the same parameter. For the second scenario, Figure 3 illustrates the relative bias for the estimated Natural Indirect Effect, and Figure 4 shows the coverage probabilities for this effect. Detailed results for the Natural Direct Effects in the second scenario, including relative bias and coverage rate findings, are available on my GitHub page.

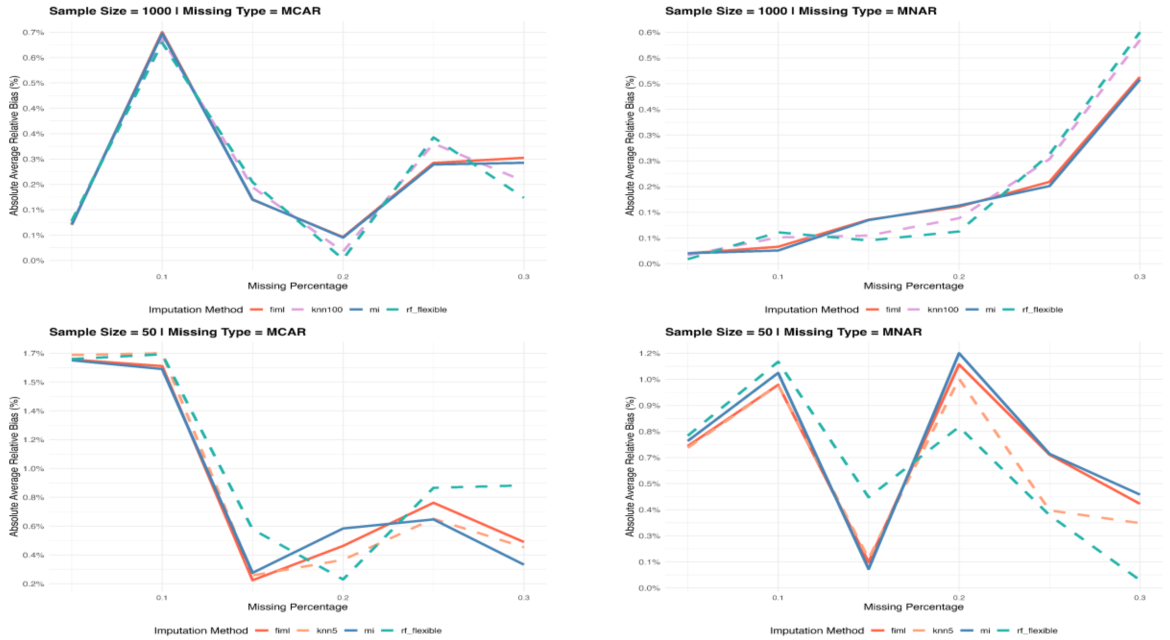


Fig 1: Average Relative bias plots for ATE under Scenario 1 DAG

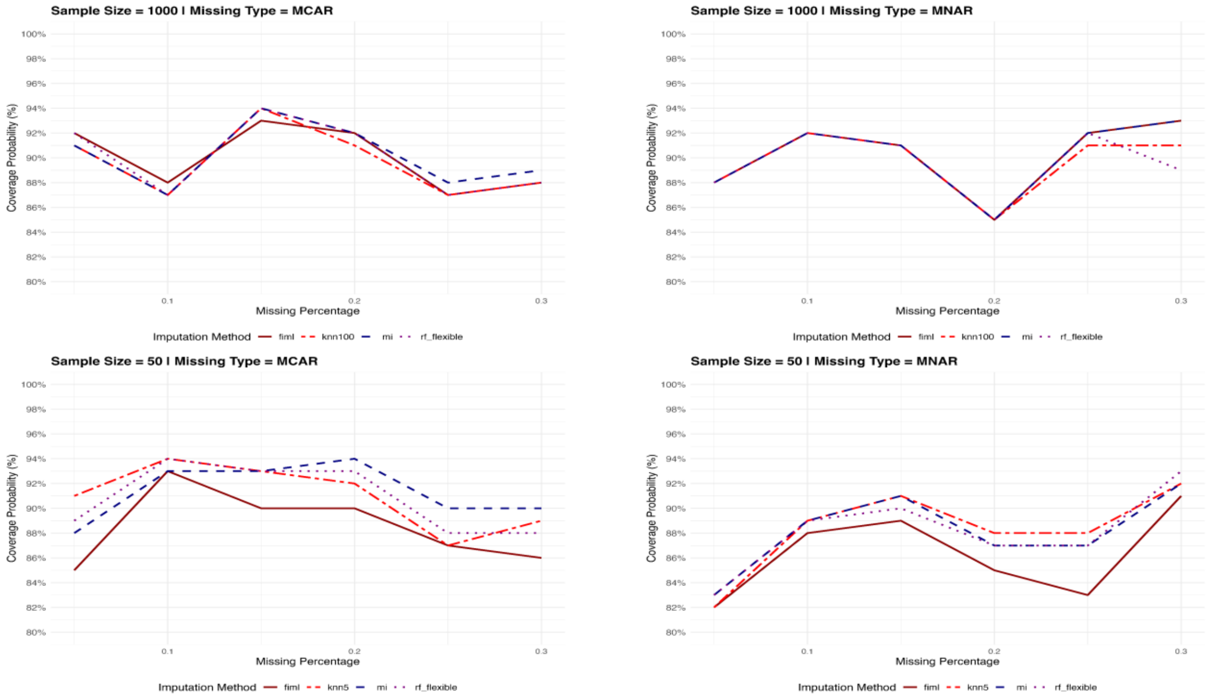


Fig 2: Coverage Probability plots for ATE under Scenario 1 DAG

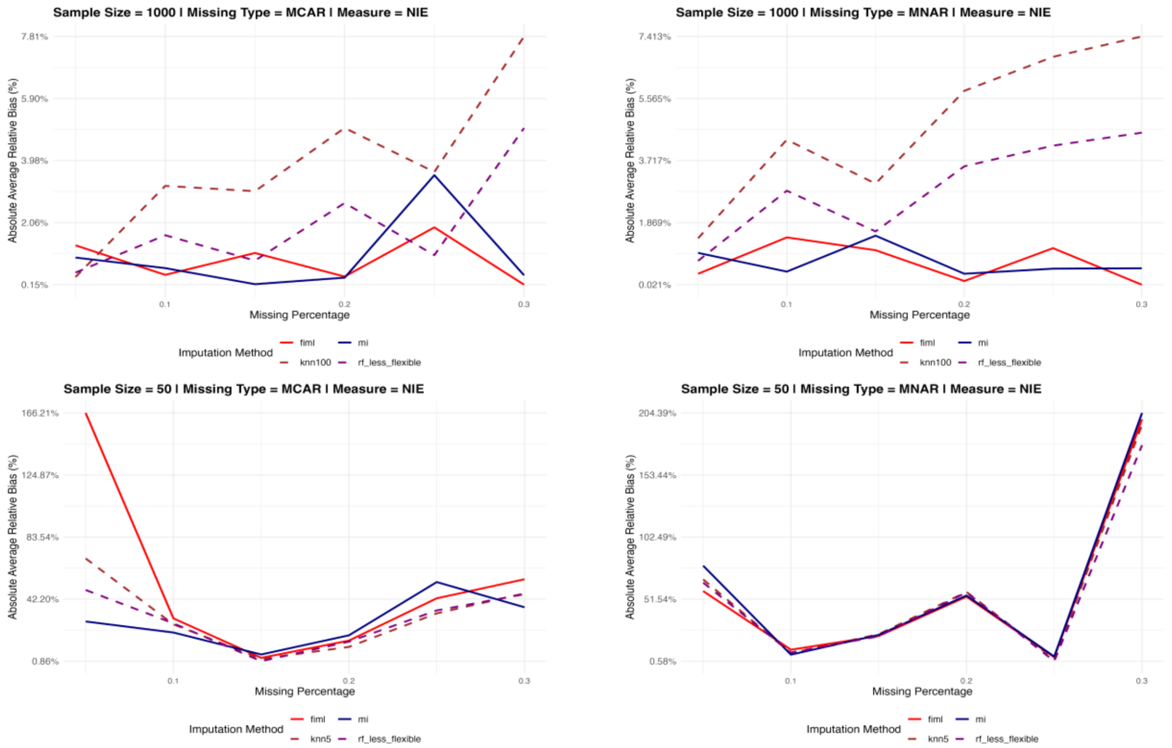


Fig 3: Average Relative bias plots for NIE under Scenario 2 DAG

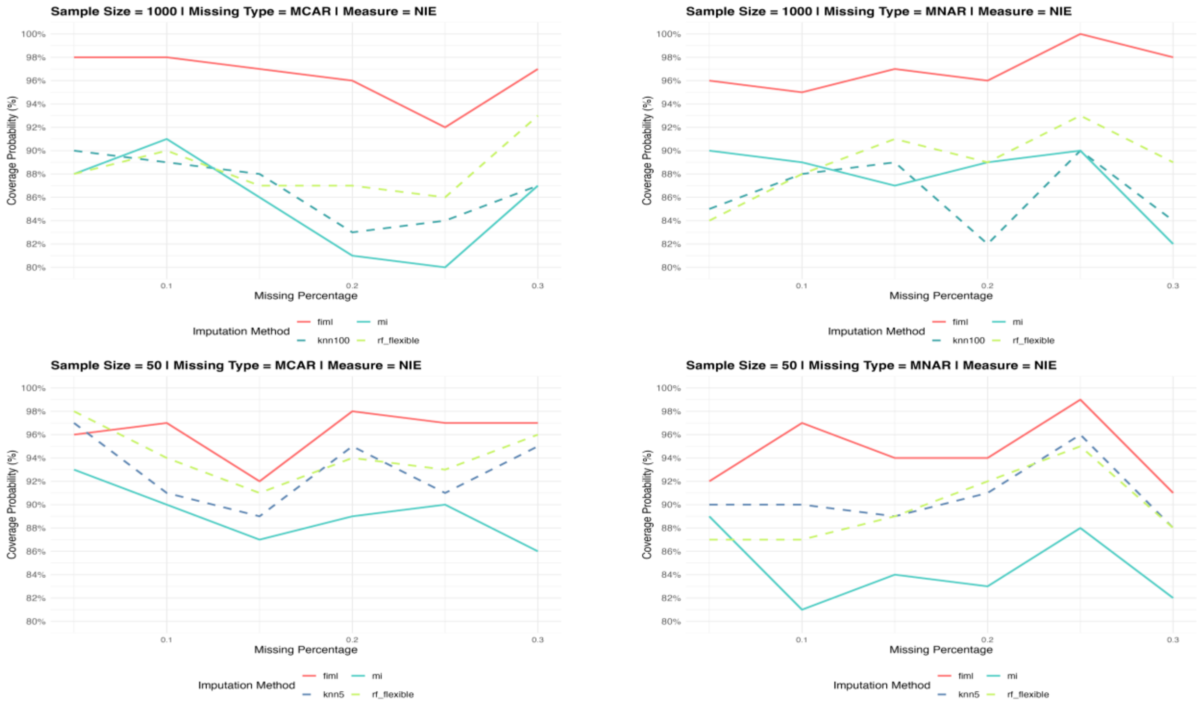


Fig 4: Coverage Probability plots for NIE under Scenario 2 DAG

Discussion

In our analysis of two distinct DAG scenarios, we observed varying performances of missing data imputation methods based on sample size, missing data mechanism (MCAR vs. MNAR), and the complexity of the DAG structure.

Scenario 1 Analysis:

- **Large Sample Size (N=1000) with MCAR:** All methods maintain low bias across varying missing percentages, with a notable peak around 10% missing data (although the value is minimal) before stabilizing. This suggests that despite the spike, all methods handle MCAR effectively in large samples. The coverage probabilities for all methods are mostly above 90%, dipping slightly at certain points but providing reliable interval estimates.
- **Large Sample Size (N=1000) with MNAR:** Although it looks like the bias increases steadily with the missing percentage for all methods, the values of bias are very minimal indicating that in high sample settings all the methods perform equally good with respect to bias. Coverage probabilities are lower than in the MCAR case, particularly for higher missing percentages. This indicates a decrease in the reliability of confidence intervals under MNAR conditions.
- **Small Sample Size (N=50) with MCAR:** There is a sharp increase in bias at lower missing percentages, which then generally decreases or stabilizes as the missing percentage increases. This pattern suggests initial sensitivity to lesser amounts of missing data, which is mitigated as missing data increases. The coverage probabilities vary more widely than in larger samples, particularly for FIML, which dips below 90% at several points and has significantly lower coverage probabilities due to the lower sample size.
- **Small Sample Size (N=50) with MNAR:** KNN and Random Forest imputation methods show a slightly better performance compared to FIML and MI with respect to bias (although the difference is very minimal). FIML again shows significantly lesser coverage rates due to the violation in MAR/MCAR assumption in addition to lower sample size.

Scenario 2 Analysis:

- **Large Sample Size (N=1000) with MCAR:** The values of biases are significantly higher when compared to Scenario 1 with a simple DAG structure. FIML and MI methods perform better when compared to non-parametric imputation methods which display larger biases at higher missing percentages. Interestingly, FIML yields conservative (More than 95%) coverage probabilities with wider confidence intervals.

- **Large Sample Size (N=1000) with MNAR:** Like the MCAR case, KNN and RF imputation methods display high biases. Even though there is a violation in the assumption, FIML and MI display exceptionally low biases due to the high sample size. FIML continues to provide higher coverage rates than the nominal level (95%).
- **Small Sample Size (N=50) with MCAR:** Like the previous Scenario, the bias for FIML at lower missing percentage are high. This is because at lower missing percentages, FIML tries to maximize the likelihood based on a limited subset of missing data. Other than this, for the rest of the missing percentages, all the methods tend to perform the same with high biases overall. It is important to conduct sensitivity analyses in such situations. The FIML method continues to provide high coverage rates here as well.
- **Small Sample Size (N=50) with MNAR:** All the methods show the same pattern of significant increase in bias with increase in missing percentage, it is better to increase the sample size and get better estimates or do sensitivity analyses in such situations rather than relying on any imputation methods.

Conclusion:

This study highlights that no single imputation method universally outperforms others; instead, the effectiveness of each method is contingent upon the assumptions they are based on. Traditional parametric methods, such as Multiple Imputation (MI) and Full Information Maximum Likelihood (FIML), perform better in scenarios with large sample sizes and under various missing data mechanisms (MCAR, MNAR, MAR). These methods are robust in handling missing data when the assumptions about data distribution and relationships are met.

Conversely, non-parametric imputation methods can be explored in small sample settings or when data missingness is not at random (MNAR). It is especially important to conduct sensitivity analyses or increase the sample size in small sample scenarios if it is possible. The advantage of non-parametric imputation methods is particularly notable in situations where the data exhibits non-linear relationships and includes categorical predictors, as non-parametric methods do not rely on the assumption of normality. These methods, such as Random Forests and K-Nearest Neighbors, can adeptly capture the complexities and interactions inherent in non-linear and non-normal data, making them preferable for more complex or less structured data environments.

Given that many real-world applications involve complex interactions and non-linear relationships among variables, an immediate extension of this research could involve simulating data that mimic these conditions. This could include using Generalized Additive Models (GAMs) or Neural Networks in the outcome regression models to better capture the underlying patterns and interactions. Such an approach is anticipated to reveal that non-parametric methods may outshine traditional imputation methods under these conditions, thereby providing a more effective toolset for dealing with the intricacies of real-world data in missing data imputation.

References

1) Tang, D., & Tong, X. (2023). A Comparison of Full Information Maximum Likelihood and Machine Learning Missing Data Analytical Methods in Growth Curve Modeling. arXiv preprint arXiv:2312.17363. https://doi.org/10.48550/arXiv.2312.17363
2) Maarten van Smeden, Bas B.L. Penning de Vries, Linda Nab, Rolf H.H. Groenwold, Approaches to addressing missing values, measurement error, and confounding in epidemiologic studies, Journal of Clinical Epidemiology, Volume 131, 2021, Pages 89-100, ISSN 0895-4356, https://doi.org/10.1016/j.jclinepi.2020.11.006.(https://www.sciencedirect.com/science/article/pii/S0895435620311756)
3) Batista, G. E., & Monard, M. C. (2002). A study of K-nearest neighbor as an imputation method. His, 87(251-260), 48.
4) Rosseel, Y. (2012). lavaan: a brief user’s guide.
5) Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118.
6) Templ, M., Alfons, A., Kowarik, A., Prantner, B., & Templ, M. M. (2022). Package ‘VIM.’
7) Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. Structural equation modeling, 8(3), 430-457.
8) van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1-67. https://www.jstatsoft.org/article/view/v045i03

9) Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R package for causal mediation analysis (Version 4.5.0) [Computer software]. https://cran.r-project.org/web/packages/mediation/vignettes/mediation.pdf

10) https://ggplot2.tidyverse.org/