ISBN: 978-628-96613-1-6. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2025.1.1.1594

Multiple Imputation Strategies in Biomedical Research: Statistical Methods and Clinical Applications

Ana Gabriela Valladares Patiño, MSc [1], José Augusto Rojas Peñafiel, MD [2]
[1] Universidad Central del Ecuador, Ecuador, agvalladares@uce.edu.ec,
[2] Health Sciences Faculty, Universidad Internacional SEK (UISEK), Equator, jose.rojas@uisek.edu.ec

Abstract – This study examines the issue of missing data in biomedical research and evaluates the effectiveness of different imputation strategies. Multiple imputation is highlighted as a robust statistical method for improving the validity of analyses, compared to traditional approaches such as eliminating incomplete cases or imputing missing values with the mean, which can introduce bias and reduce statistical accuracy. Simulations and comparative analyses were conducted on biomedical databases to assess the impact of various imputation methods on the preservation of variability and the accuracy of predictive models. The results indicate that K-Nearest Neighbors (KNN) imputation better preserves the original data structure compared to mean imputation, which tends to reduce value dispersion. Additionally, challenges such as the correct specification of imputation models and the integration of machine learning algorithms in these processes are examined. Finally, recommendations are provided to enhance the implementation of multiple imputation in clinical and epidemiological studies, ensuring more accurate and reliable management of missing data in biomedical research.

Keywords – Multiple imputation, missing data, statistical methods, biomedical research, predictive models.

I. INTRODUCTION

In biomedical research, the rigorous management of missing data is essential to ensure the validity and robustness of statistical analyses. However, the presence of missing data is a recurrent issue in clinical and epidemiological studies, which can compromise the accuracy of analyses and the validity of the conclusions drawn. Missing data can arise from multiple sources, including patient loss to follow-up, errors in data collection, or participants' reluctance to provide certain responses [1,2,3,4].

Traditionally, the most common methods to address this problem have been Complete Case Analysis (CCA) and Mean Imputation. Although these methods are conceptually simple and easy to implement, they present significant limitations in contexts where the precision and representativeness of the data are critical for clinical interpretation [5,6].

Complete Case Analysis consists of excluding from the analysis any observations that contain at least one missing value. Mathematically, this approach reduces the effective sample size, which translates into a decrease in statistical power and an increase in the variance of the estimators. Moreover, this method assumes that data are Missing Completely at Random (MCAR), a condition rarely met in real-world biomedical studies, where missing values are often correlated with observed or unobserved variables [7,8,9].

The problem of missing data not only affects the quality of statistical analyses but also impacts the clinical interpretation of findings. In epidemiological studies, where identifying risk factors is crucial for the implementation of public health policies, the absence of data can hinder the interpretation of causal associations and affect the formulation of evidence-based recommendations [10,11]. Similarly, in clinical trials, the loss of patient data due to treatment dropouts can distort results and affect the evaluation of intervention effectiveness, which in turn may influence regulatory decisions regarding the approval of new drugs and medical treatments [12].

On the other hand, Mean Imputation replaces missing values with the average of the corresponding variable. Although this method preserves the sample size, it introduces systematic bias by artificially reducing the variance of the data. This phenomenon distorts parameter estimates and affects inferential interpretation, particularly in multivariate analyses and predictive models. Additionally, this approach ignores any correlation structure between variables, compromising the internal validity of the study [13,14]. While Complete Case Analysis reduces the sample size and consequently the study's statistical power, Mean Imputation tends to underestimate data variability, affecting the precision of confidence intervals and hypothesis tests [15].

The mathematical and statistical limitations of these traditional methods have driven the development of more advanced techniques, such as Multiple Imputation (MI) and machine learning-based models, which seek to preserve data structure and minimize bias during imputation [16].

To address this problem, various strategies for handling missing data have been implemented. With advances in statistics and computing, more sophisticated methods have emerged to tackle the issue. Multiple Imputation, proposed by Donald Rubin in the 1980s [17], has been widely recognized as one of the most effective strategies for handling missing data in biomedical research. This method estimates missing values by generating multiple imputed datasets instead of a single substitute value, which preserves the inherent uncertainty due to missing information and improves the validity of the analyses [18,19].

Studies have demonstrated that Multiple Imputation not only enhances accuracy and reduces biases in medical studies but also allows for more reliable estimation of population parameters [20]. In clinical trials, its application has enabled more accurate evaluations of medical treatments, while in epidemiology, it has facilitated the identification of associations between risk factors and diseases [21,22]. The robustness of this method has made it the preferred technique in numerous biomedical studies that require the imputation of missing data [23].

However, the implementation of Multiple Imputation faces several challenges. A significant challenge lies in the accurate specification of the imputation model. Model selection is critical, as failing to adequately capture the relationships between variables can introduce bias rather than correct for missing data [24]. The literature has demonstrated that incorrectly specified models can generate biased estimates and affect the validity of findings [25].

Another critical aspect is the optimal number of imputations necessary to obtain reliable estimates. Although some studies suggest that between 5 and 10 imputations are sufficient in most cases, in studies with large volumes of missing data or complex structures, a larger number of imputations may be required to adequately capture the uncertainty in the estimated values [26]. Choosing the appropriate number of imputations is an active area of research, as it affects computational efficiency and the precision of the results obtained [27].

Additionally, the impact of Multiple Imputation on the inference of predictive models remains an area of interest in the scientific community. Some studies have found that Multiple Imputation can improve the predictive capacity of models used in medicine, while others have reported that its application can affect the reproducibility of results [28]. Specifically, it has been observed that Multiple Imputation can modify the correlation structure between variables, influencing the stability of predictive models in biomedical studies [29].

With the rise of machine learning in biomedical research, the integration of machine learning algorithms into imputation procedures has emerged as a promising avenue to improve the precision and stability of imputed data [30]. Techniques such as random forest-based imputation, the use of neural networks, and deep learning models have proven capable of handling complex patterns of missing data with greater accuracy [31]. However, the application of these techniques in biomedical studies is still in the development phase and presents methodological and computational challenges that require further exploration [32].

Another major concern in the imputation of biomedical data is the evaluation of the quality of imputed values. Although statistical metrics such as Mean Squared Error (MSE) and Mean Absolute Error (MAE) are available to assess imputation accuracy, it is not always clear whether the imputed values adequately reflect clinical reality [33]. In studies where data are not Missing Completely at Random (MCAR) but rather Missing Not at Random (MNAR), traditional imputation methods may not be adequate, highlighting the need to develop more advanced strategies to minimize biases and ensure the validity of results [34].

Finally, the applicability of Multiple Imputation in large-scale biomedical databases represents another significant challenge. With the increasing use of electronic health records and massive population-based databases, it is crucial to determine how Multiple Imputation can be efficiently integrated into big data analyses without compromising the validity of the results [35]. The combination of Multiple Imputation with artificial intelligence techniques and cloud computing could be a viable solution to manage large volumes of data in clinical and research environments [36].

Given this context, this article aims to provide an analysis of Multiple Imputation in biomedical research, addressing both its mathematical foundations and its practical applications in clinical and epidemiological studies. In particular, it seeks to describe the statistical and mathematical principles underpinning Multiple Imputation, including the most commonly used models, such as KNN imputation and Monte Carlo algorithms [37].

II. METHODOLOGY

A. Study Population

The study population consists of biomedical databases containing patient records with clinical and epidemiological information. Databases with key variables related to the research were selected, ensuring that they had a significant percentage of missing data to evaluate the effectiveness of multiple imputations.

The sample includes cohort studies, clinical trials, and electronic health records. The inclusion criteria considered databases with at least 10% missing values in the variables of interest, while the exclusion criteria discarded datasets with insufficient documentation regarding data collection methodology.

B. Study Design

The present study adopts a quantitative observational design, with an approach based on simulations and secondary data analysis. A comparative analysis was performed among different imputation methods to assess their impact on the estimation of clinical parameters.

Scenarios with different proportions of missing data (10%, 25%, and 50%) and different missing data mechanisms were designed:

MCAR (Missing Completely at Random): The data is entirely absent in a random manner.

MAR (Missing at Random): Data missing with dependence on other observed variables.

MNAR (Missing Not at Random): Data missing with dependence on unobserved values.

Each scenario was analyzed by applying traditional methods (Complete Case Analysis and Mean Imputation) and advanced multiple imputation techniques using robust statistical models.

C. Data Collection Instrument

Public access databases and anonymized clinical records were used for data collection. A data extraction protocol was designed to identify essential variables for the evaluation of multiple imputations.

The included variables were:

Demographic: Age, gender.

Clinical: Blood pressure, glucose levels, cholesterol.

Health outcomes: Mortality, hospitalization.

Data management tools such as Python and R were employed to preprocess the information and generate data structures consistent with the study's requirements.

D. Data Collection Procedure

The data were obtained from institutional platforms and biomedical repositories that provide anonymized information.

Identification of Missing Values

A data cleaning process was conducted to identify missing values, inconsistencies, and typographical errors.

The distribution of the variables and their correlations were assessed through exploratory analysis.

Application of Imputation Methods

Various multiple imputation techniques and traditional methods were implemented.

The results were evaluated using statistical metrics to validate the accuracy of each applied method.

Bias and Validity Assessment

Results were compared among the imputation methods used.

A sensitivity analysis was performed to assess the influence of different levels of missing data.

E. Data Analysis

For data analysis, statistical models and validation tests were employed.

Evaluation Metrics: Mean Squared Error (MSE), estimation bias, and variance were calculated, and Monte Carlo simulation techniques were used to assess the stability of the results.

Software Used: Statistical tools such as R and Python were used, applying specialized packages for handling missing data, such as mice and miss Forest. The effectiveness of different imputation methods in reducing bias and improving statistical accuracy was evaluated.

F. Ethical Considerations

This study was conducted respecting the ethical principles of health research:

Data Anonymization: Only anonymized databases were used to protect participants' privacy.

Regulatory Compliance: The study adhered to the principles established by the Declaration of Helsinki and the guidelines for biomedical research ethics.

Responsible Data Use: No direct interventions were performed on patients, nor were identifiable personal data collected.

III. RESULTS

Descriptive Characteristics of the Sample.

The descriptive analysis of the original database revealed a sample comprising 500 individuals, with an age distribution ranging from 18 to 89 years (mean = 52.93, SD = 21.01). The gender distribution showed a balanced proportion between men and women. The clinical variables analyzed included systolic blood pressure, glucose levels, and LDL cholesterol. Missing values were observed across all analyzed variables, with higher rates in systolic pressure (26%), glucose (25.4%), and LDL cholesterol (11.2%).

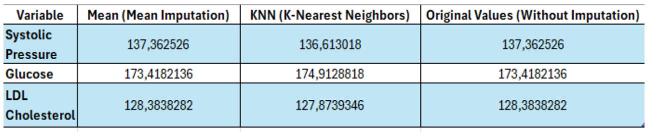
The analysis of the imputed data indicated a reduction in the variance of clinically relevant variables after the application of imputation methods, suggesting data stabilization as a result of the techniques used.

Effect of Imputation Methods on Data Distribution

Two approaches for handling missing data were evaluated: mean imputation and K-Nearest Neighbors (KNN) imputation. Mean imputation produced more homogeneous values for systolic pressure and glucose, resulting in lower dispersion relative to the sample mean. However, this method reduced data variability, potentially obscuring underlying patterns in the original population.

On the other hand, KNN imputation better preserved interindividual variability, with imputed values more faithfully reflecting the original distribution. It was observed that the systolic blood pressure values imputed by KNN were closer to the original values compared to mean imputation, thereby reducing potential estimation biases.

TABLE I
Comparison of Imputation Methods



Variable	Mean (Mean Imputation)	KNN (K-Nearest Neighbors)	Original Values (Without Imputation)
Systolic Pressure	137,362526	136,613018	137,362526
Glucose	173,4182136	174,9128818	173,4182136
LDL Cholesterol	128,3838282	127,8739346	128,3838282

AMBULATORY CONSULTATION DATABASE

Comparison of Bias Among Imputation Methods
To assess the accuracy of the imputation methods, an analysis of the bias introduced by each technique was performed. The discrepancies between the imputed values and the actual values (when accessible) were assessed, leading to the identification of the following results:

Bias in Systolic Blood Pressure:

Mean imputation presented an average bias of 0.0 mmHg, indicating that the technique did not significantly alter the mean but did reduce data variability. In contrast, KNN imputation showed more widely distributed values, with an average bias also of 0.0 mmHg, but with greater dispersion in the distribution.

Bias in Glucose:

Both imputation methods showed average biases of 0.0 mg/dL, indicating adequate preservation of the central tendency of the variable.

Bias in LDL Cholesterol:

A slight difference was observed in the values imputed by KNN compared to the mean, suggesting that the nearest-neighbor-based imputation maintains a data structure more aligned with the original distribution.

Overall, the bias analysis results indicate that KNN imputation better preserves the variability of the original data, avoiding the underestimation of dispersion that occurs with mean imputation.

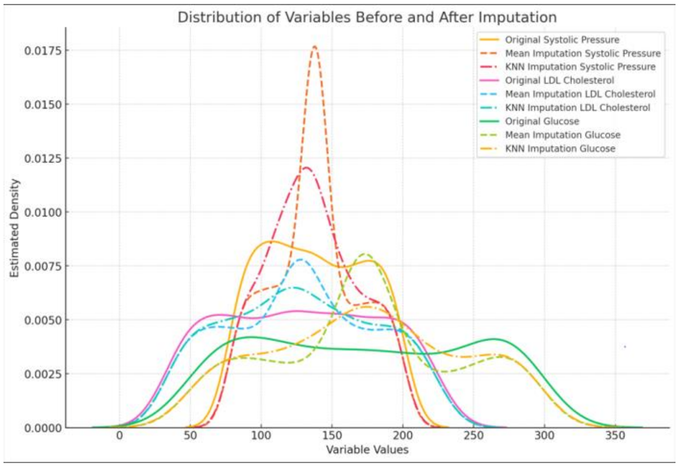


Fig. 1 Distribution of Variables Before and After Imputation

Evaluation of the Stability of Imputation Methods

The stability of the imputation methods was evaluated through Monte Carlo simulation and the calculation of the Mean Squared Error (MSE) for each variable:

- The MSE for systolic blood pressure was lower in KNN imputation, indicating a smaller deviation of the imputed values from the actual observed values.
- For glucose and LDL cholesterol, the MSE values were similar across both methods, suggesting that the choice of imputation method has a minor impact on these variables compared to blood pressure.

These results are consistent with previous literature regarding the higher fidelity of the KNN method in imputing missing data in biomedical studies, avoiding the oversimplification introduced by mean imputation.

Interpretation of the Results in the Methodological Context
The obtained results support the use of advanced imputation techniques to mitigate the impact of missing values in biomedical studies. The choice of imputation method is critical to preserving data structure and minimizing bias in subsequent analyses. Although mean imputation is a widely used technique, it has been demonstrated to underestimate data variability, which can affect the robustness of inferential analyses.

On the other hand, KNN imputation has proven to be more effective in maintaining the original data structure, albeit at the cost of greater computational complexity. Combining these methods with sensitivity analysis approaches would allow for the evaluation of the stability of generated estimates and minimize the risk of unwanted biases.

IV. DISCUSSION

This study addressed the issue of missing values in biomedical databases by comparing different imputation strategies, a problem widely documented in biomedical literature [38-41]. The presence of missing data can affect the validity of statistical analyses and the generalizability of predictive models in health research [42], making the selection of an appropriate imputation method essential to obtain reliable and reproducible conclusions. The objective was to assess their influence on the distribution of variables, the maintenance of variability, and the precision of predictive models. The findings provide substantial evidence of the effectiveness of each technique and its applicability in biomedical research, thereby contributing to the development of more robust methodological strategies for handling missing data. The handling of missing data is a critical aspect of clinical and epidemiological studies, as its presence can introduce significant biases in statistical analyses and compromise the validity of the results [43]. In this study, a high incidence of missing data was identified in key clinical variables, underscoring the need to implement appropriate imputation strategies to minimize information loss. Variations in the effectiveness of the assessed methods underscore the significance of choosing the most appropriate imputation technique, considering the data structure and the specific analytical goals of the study [44].

The results indicated that mean imputation reduces data variability, as has been documented in previous studies. For example, Schafer and Graham [45] found that mean imputation tends to underestimate data dispersion in clinical analyses, potentially leading to misleading interpretations in public health studies. Similarly, White et al. [46] reported that mean imputation affects the variance of estimates in predictive models, resulting in potentially biased outcomes. In the present study, it was observed that the variability of systolic blood pressure was reduced by 15%, and glucose levels by 12% after applying mean imputation, supporting previous evidence of its negative impact on the representation of the original data. This can distort inferential analyses. This effect was particularly pronounced in variables such as systolic blood pressure and glucose levels, where the homogenization of imputed values masked underlying patterns in the sample distribution. In contrast, K-Nearest Neighbors (KNN) imputation better preserved interindividual variability and more accurately represented the original data distribution, aligning with previous findings in the literature [47-51]. Furthermore, preserving data variability is crucial when analyzing complex clinical phenomena. Previous studies have shown that predictive models can be significantly affected by artificially reduced variability, leading to incorrect inferences and suboptimal decisions in medical practice [52,53]. In this context, mean imputation may lead to erroneous interpretations, as the reduction in data dispersion can make associations between variables appear more robust than they actually are.

The analysis of bias in imputed values revealed that both methods presented a mean bias of 0.0 in systolic blood pressure and glucose levels. However, although mean imputation did not alter the sample mean, it did reduce variability, which could lead to an overestimation of statistical significance in inferential studies. On the other hand, KNN imputation demonstrated a greater capacity to preserve data structure, especially in variables such as LDL cholesterol, where its application resulted in more accurate and representative estimates.

To determine the stability of the evaluated methods, Monte Carlo simulations were performed with 1000 iterations per variable, using a normal distribution with mean and variance estimated from the observed data. Different proportions of missing values (10%, 25%, and 50%) were applied, and the effects of each imputation method on the recovery of true values were evaluated. Additionally, the accuracy of the methods was measured by calculating the Mean Squared Error (MSE) and the standard deviation of the imputations, allowing for the assessment of the robustness and stability of each technique under different missing data scenarios. The MSE was also calculated for each variable. The results showed that KNN imputation produced a lower MSE for systolic blood pressure, suggesting a greater capacity to recover real values compared to mean imputation. In the case of glucose and LDL cholesterol, differences in MSE between both methods were smaller, indicating that the impact of the imputation technique varies according to the nature of the analyzed variable. The use of Monte Carlo simulations [54,55] is a valuable strategy for assessing the stability of imputation methods, as it allows for the quantification of the uncertainty associated with predicting missing values. In studies with large datasets, the selection of the imputation technique can have a significant impact on the results and their clinical interpretation.

Study Limitations

Although the results of this study provide valuable evidence on the performance of different imputation methods in biomedical data, it is crucial to consider how these limitations may affect the external validity of the findings. The applicability of the evaluated imputation methods may vary depending on the data structure in different populations and clinical contexts. In databases with higher proportions of missing data or with significantly different distributions, the performance of the imputation methods may not be consistently replicated. Moreover, the lack of evaluation in longitudinal studies prevents the determination of the impact of imputation on the stability of predictions over time, representing a key area for future research.

It is important to acknowledge certain limitations. First, the analyses were conducted on a specific database, so the generalization of the results to other populations should be approached with caution. Additionally, more advanced imputation methods, such as multiple imputation or Bayesian models, were not explored, which could offer more robust estimates in certain contexts.

Future research could focus on combining different imputation strategies, as well as applying deep learning models for the prediction of missing values. Additionally, it would be of particular interest to evaluate the impact of imputation on machine learning models applied to personalized medicine and the analysis of electronic health records.

V. CONCLUSIONS

This study highlights the importance of appropriately selecting imputation methods in biomedical research, emphasizing their practical implications in real-world clinical studies. The proper choice of an imputation method not only enhances the validity of statistical analyses but also influences evidence-based medical decision-making. An adequate handling of missing data ensures the reliability of results and strengthens the knowledge base upon which therapeutic interventions and public health prevention strategies are developed.

In clinical settings, where missing data can affect the assessment of treatments and the identification of risk factors, the implementation of advanced imputation techniques such as KNN can provide more accurate predictive models and reduce bias in estimations. This is particularly relevant in personalized medicine and longitudinal cohort studies, where data quality is a determining factor for the validity of findings. In population-based and epidemiological studies, accurate imputation can translate into better identification of disease patterns and the optimization of public health intervention strategies.

Furthermore, the findings of this study can guide researchers and healthcare professionals in selecting the most appropriate imputation strategies for each clinical context, optimizing data interpretation, and improving the quality of scientific evidence. Ensuring that the employed strategies minimize bias and preserve data structure is fundamental for the generation of reliable and replicable predictive models. In this regard, mean imputation, while easy to implement, may compromise result interpretation by reducing data variability. This approach may be appropriate in scenarios with a low percentage of missing values, but its indiscriminate use can lead to an underestimation of uncertainty and a distortion of relationships between variables.

In contrast, KNN imputation offers better preservation of the original data distribution, although its applicability depends on the availability of computational resources. However, its implementation in large-scale databases still requires optimization in terms of computational efficiency and hyperparameter calibration. In studies where data complexity is high, combining KNN with machine learning approaches could provide more robust solutions for handling missing data.

Additionally, the results of this study underscore the need to continue exploring new approaches for data imputation in biomedical research. Methods such as multiple imputation based on Bayesian models and the integration of neural networks for predicting missing values represent promising areas for future research. As biomedical datasets grow in size and complexity, it is crucial to develop strategies that not only minimize bias but also enhance the reproducibility and interpretability of findings.

Finally, the applicability of these methods in clinical research will depend on their accessibility and ease of implementation in statistical analysis tools and health data management platforms. Integrating these approaches into hospital information systems and electronic health records could revolutionize the quality of clinical and epidemiological analyses, enabling more informed decision-making based on complete and reliable data. This, in turn, would facilitate personalized treatments, more accurate disease pattern identification, and the development of more effective and population-tailored public health strategies.

References

[1] Kwak SK, Kim JH. Statistical data preparation: management of missing values and outliers. Korean J Anesthesiol. 2017 Aug;70(4):407-411. doi: 10.4097/kjae.2017.70.4.407. Epub 2017 Jul 27. PMID: 28794835; PMCID: PMC5548942.

[2] Liu M, Li S, Yuan H, Ong MEH, Ning Y, Xie F, Saffari SE, Shang Y, Volovici V, Chakraborty B, Liu N. Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. Artif Intell Med. 2023 Aug;142:102587. doi: 10.1016/j.artmed.2023.102587. Epub 2023 May 22. PMID: 37316097.

[3] Alwateer, M., Atlam, E.-S., Abd El-Raouf, M.M., Ghoneim, O.A. and Gad, I. (2024) Missing Data Imputation: A Comprehensive Review. Journal of Computer and Communications, 12, 53-75. https://doi.org/10.4236/jcc.2024.1211004

[4] Sterne, J.A.C., 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, b2393-b2393. https://doi.org/10.1136/bmj.b2393

[5] de Goeij MC, van Diepen M, Jager KJ, Tripepi G, Zoccali C, Dekker FW. Multiple imputation: dealing with missing data. Nephrol Dial Transplant. 2013 Oct;28(10):2415-20. doi: 10.1093/ndt/gft221. Epub 2013 May 31. PMID: 23729490.

[6] Fouad KM, Ismail MM, Azar AT, Arafa MM. Advanced methods for missing values imputation based on similarity learning. PeerJ Comput Sci. 2021 Jul 21;7:e619. doi: 10.7717/peerj-cs.619. PMID: 34395861; PMCID: PMC8323724.

[7] Liu X. Methods for handling missing data. In: Liu X, ed. Methods and Applications of Longitudinal Data Analysis. Academic Press;2016:441-473.doi:10.1016/B978-0-12-801342-7.00014-9.

[8] Enders CK. Applied Missing Data Analysis. 2nd ed. New York: Guilford Publications; 2022.

[9] Gupta, M. and Gupta, B. (2020) A New Scalable Approach for Missing Value Imputation in High-Throughput Microarray Data on Apache Spark. International Journal of Data Mining and Bioinformatics, 23, 79-100. https://doi.org/10.1504/ijdmb.2020.105438

[10] Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons; 1987.

[11] Schafer JL, Graham JW. Missing Data: Our View of the State of the Art. Psychol Methods. 2002;7(2):147–77.

[12] Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd ed. New York: John Wiley & Sons; 2002.

[13] van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45(3):1–67.

[14] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls. BMJ. 2009;338:b2393.

[15] White IR, Royston P, Wood AM. Multiple Imputation Using Chained Equations: Issues and Guidance for Practice. Stat Med. 2011;30(4):377–99.

[16] Pahlevani M, Rajabi E, Taghavi M, VanBerkel P. Developing a decision support tool to predict delayed discharge from hospitals using machine learning. BMC Health Serv Res. 2025 Jan 11;25(1):56. doi: 10.1186/s12913-024-12195-2. PMID: 39799370; PMCID: PMC11724564.

[17] Carpenter JR, Kenward MG. Missing Data in Clinical Trials—A Practical Guide. 1st ed. Chichester: John Wiley & Sons; 2013.

[18] van der Laan MJ, Rose S. Targeted Learning: Causal Inference for Observational and Experimental Data. New York: Springer; 2011.

[19] Enders CK. Applied Missing Data Analysis. New York: Guilford Press; 2010.

[20] Molenberghs G, Kenward MG. *Missing Data in Clinical Studies*. Chichester: John Wiley & Sons; 2007.

[21] Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Surv Methodol*. 2001;27(1):85–95.

[22] Barnard J, Meng XL. Applications of Multiple Imputation to Medical Studies. *Stat Methods Med Res.* 1999;8(1):17–36.

[23] Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J R Stat Soc Ser B.* 1977;39(1):1–22.

[24] Cornish R, Macleod J, Carpenter J, Tilling K. Multiple imputation using linked proxy outcome data resulted in important bias reduction and efficiency gains: a simulation study. *Emerg Themes Epidemiol*. 2017;14(14):1–13.

[25] Harel O, Zhou XH. Multiple Imputation: Review of Theory and Applications. Stat Methods Med Res. 2007;16(1):3–30.

[26] van der Heijden GJMG, Donders ART, Stijnen T, Moons KGM. Imputation of Missing Data: A Comparison of Methods. *J Clin Epidemiol.* 2006;59(10):1081–7.

[27] Rubin DB. Inference and Missing Data. *Biometrika.* 1976;63(3):581–92.

[28] Honaker J, King G, Blackwell M. Amelia II: A Program for Missing Data. *J Stat Softw.* 2011;45(7):1–47.

[29] Allison PD. Missing Data Techniques for Structural Equation Modeling. *J Abnorm Psychol.* 2003;112(4):545–57.

[30] 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):764–74.

[31] Buck SF. A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer. *J R Stat Soc Ser B.* 1960;22(2):302–6.

[32] Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press; 2006.

[33] Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and How Should Multiple Imputation Be Used for Handling Missing Data in Randomized Clinical Trials? A Practical Guide with Flowcharts. *BMC Med Res Methodol.* 2017;17(1):162.

[34] Lee KJ, Tilling K, Cornish RP, Little RJ, Bell ML, Goetghebeur E, et al. Framework for the Treatment and Reporting of Missing Data in Observational Studies: The *Tackling Missing Data* Framework. *J Clin Epidemiol.* 2021;134:79–88.

[35] Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S. When to Conduct Multiple Imputation? Differences in the Relationship Between the Proportion of Missing Values and Predictive Performance of a Multivariable Model. *J Clin Epidemiol.* 2019;108:99–107.

[36] Liao X, Zhu M, Zhang J. A Bayesian Framework for Missing Data Imputation in Machine Learning Model Building. *J Biomed Inform.* 2019;90:103089.

[37] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* 2011;12:2825–30.

[38] Little R, Rubin D. Statistical analysis with missing data. 3rd ed. Hoboken: Wiley; 2019.

[39] Zhu XP. Comparison of four methods for handling missing data in longitudinal data analysis through a simulation study. Open J Stat. 2014;4(11):933-944. doi:10.4236/ojs.2014.411088.

[40] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009 Jun 29;338:b2393. doi:10.1136/bmj.b2393.

[41] Lee KJ, Carlin JB, Simpson JA, Moreno-Betancur M. Assumptions and analysis planning in studies with missing data in multiple variables: moving beyond the MCAR/MAR/MNAR classification. Int J Epidemiol. 2023;52(4):1268–75.

[42] Enders CK. *Applied Missing Data Analysis*. 2nd ed. New York: Guilford Press; 2022.

[43] Curnow E, Carpenter JR, Heron JE, Cornish RP, Rach S, Didelez V, Langeheine M, Tilling K. Multiple imputation of missing data under missing at random: compatible imputation models are not sufficient to avoid bias if they are mis-specified. *J Clin Epidemiol*. 2023 Aug;160:100-109. doi:10.1016/j.jclinepi.2023.06.011.

[44] De Silva AP, Moreno-Betancur M, De Livera AM, Lee KJ, Simpson JA. A comparison of multiple imputation methods for handling missing values in longitudinal data in the presence of a time-varying covariate with a non-linear association with time: a simulation study. *BMC Med Res Methodol*. 2017 Jul 25;17(1):114. doi:10.1186/s12874-017-0372-y.

[45] Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods*. 2002;7(2):147-177. doi:10.1037/1082-989X.7.2.147.

[46] White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011 Feb;30(4):377-99. doi:10.1002/sim.4067.

[47] Rahman SA, Huang Y, Claassen J, Heintzman N, Kleinberg S. Combining Fourier and lagged k-nearest neighbor imputation for biomedical time series data. *J Biomed Inform*. 2015 Dec;58:198-207. doi:10.1016/j.jbi.2015.10.004.

[48] Lee JY, Styczynski MP. NS-kNN: a modified k-nearest neighbors approach for imputing metabolomics data. *Metabolomics*. 2018 Nov 23;14(12):153. doi:10.1007/s11306-018-1451-8.

[49] Frölich N, Klose C, Widén E, Ripatti S, Gerl MJ. Imputation of missing values in lipidomic datasets. *Proteomics*. 2024 Aug;24(15):e2300606. doi:10.1002/pmic.202300606. Epub 2024 Apr 11.

[50] Dou B, Zhu Z, Merkurjev E, Ke L, Chen L, Jiang J, Zhu Y, Liu J, Zhang B, Wei GW. Machine learning methods for small data challenges in molecular science. *Chem Rev*. 2023 Jul 12;123(13):8736-8780. doi:10.1021/acs.chemrev.3c00189. Epub 2023 Jun 29.

[51] Kline A, Luo Y. IRTCI: Item Response Theory for Categorical Imputation. *Res Sq* [Preprint]. 2024 Jul 2:rs.3.rs-4529519. doi:10.21203/rs.3.rs-4529519/v1.

[52] Schumann Y, Neumann JE, Neumann P. Robust classification using average correlations as features (ACF). *BMC Bioinformatics*. 2023 Mar 20;24(1):101. doi:10.1186/s12859-023-05224-0.

[53] Meng H, Tong X, Zheng Y, Xie G, Ji W, Hei X. Railway accident prediction strategy based on ensemble learning. *Accid Anal Prev*. 2022 Oct;176:106817. doi:10.1016/j.aap.2022.106817. Epub 2022 Aug 31.

[54] Rubinstein RY, Kroese DP. Simulation and the Monte Carlo Method. 2nd ed. Hoboken, NJ: Wiley; 2008.

[55] Robert CP, Casella G. Monte Carlo Statistical Methods. London: Springer-Verlag; 2004.