



A Comparative Evaluation of Imputation Techniques for Missing Data: A Simulation-Based Analysis

Rahibu Abdalla Abassi^{1,*} Rocky Rajabu Akarro²

¹ Department of Natural Sciences, State University of Zanzibar, P.O. Box 146, Zanzibar, Tanzania

² Department of Statistics, University of Dar es Salaam, P.O. Box 35091, Dar es Salaam, Tanzania

Article History

Received: August 12, 2025

Accepted: April 03, 2026

Published: May 12, 2026

Abstract

Missing data are a common occurrence in research and, if not appropriately addressed prior to analysis, may compromise the validity of study findings. This article evaluates the effectiveness of various imputation techniques as formal approaches for handling missing covariate data. Root Mean Squared Error (RMSE) was computed for each imputation method under Missing Completely at Random (MCAR) and Missing at Random (MAR) mechanisms to identify the technique that yielded the lowest RMSE across different simulation settings. The results indicate that RMSE increased for all imputation techniques as the proportion of missing data increased under both MCAR and MAR mechanisms. Under the MCAR mechanism, both simulated and non-simulated data exhibited broadly similar trends across three multiple imputation-based techniques: Multiple Imputation by Chained Equations (MICE), Expectation-Maximization with Bootstrapping (EMB), and Predictive Mean Matching (PMM). However, the single imputation method (series mean) yielded RMSE values that differed substantially from those of the multiple imputation approaches. Among the applied multiple imputation (MI) techniques, Predictive Mean Matching (PMM) yielded the lowest RMSE values: 5.80 for the simulated dataset with a 15% missing rate under the MAR mechanism and 7.50 for the empirical dataset. The findings indicate that multiple imputation techniques are preferable for handling missing data, as they account for uncertainty and improve estimation efficiency. It is recommended that results from both imputed and original datasets be compared to assess the impact of missing data on the analysis. For clarity and transparency, researchers should also report the means and standard errors for both imputed and non-imputed datasets.

Keywords:

non-imputed data; imputation technique; missingness mechanisms, missing proportion, simulation-based analysis

1. Introduction

1.1. Background

Missing data are a common occurrence in medical research and can lead to inaccurate statistical inference if not appropriately addressed; in fact, ignoring missing values may result in biased estimates. Data may be missing for various reasons, including computational errors, random equipment malfunctions, natural events such as death, non-response to unclear or sensitive survey ques-

tions, and patients failing to attend routine clinical appointments [1]. Many researchers address missing data by removing incomplete observations prior to analysis, an approach known as complete case analysis (CCA). This method excludes any record containing at least one missing value during statistical estimation, thereby reducing statistical power and potentially introducing bias [2]. A more appropriate alternative is the use of imputation techniques, which estimate and replace missing values to generate a complete dataset for analysis [3].

* Corresponding Author:

Rahibu Abdalla Abassi, Department of Natural Sciences, State University of Zanzibar, P.O. Box 146 Zanzibar, Tanzania, r.abassi@suza.ac.tz



© 2026 Copyright by the Authors.

Licensed as an open access article using a [CC BY 4.0 license](https://creativecommons.org/licenses/by/4.0/).

Imputation methods are broadly categorized into single imputation techniques, such as mean imputation, hot-deck imputation, and k-nearest neighbor imputation, and multiple imputation approaches, including multiple imputation by chained equations (MICE), predictive mean matching (PMM), and expectation-maximization with bootstrapping (EMB), among others. Although these methods are employed to replace missing values under varying scenarios, there is no conclusive consensus regarding the most appropriate approach under specific conditions such as differing proportions of missingness, missing data patterns, and missingness mechanisms. In this context, simulation studies play a crucial role in assessing the performance and suitability of different imputation methods.

In experimental and observational studies, any variable or observation may be affected by missingness. Missing data mechanisms are commonly classified into three categories: Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR) [4]. Under the MCAR mechanism, missing values occur independently of both observed and unobserved data [2,5]. For instance, MCAR data arise when a patient's urine sample is accidentally dropped and broken, leading to missing laboratory results. Data are classified as Missing at Random (MAR) when the probability of missingness depends only on observed information and not on unobserved values [2,6,7]. The occurrence of missing values is associated with known patient characteristics or behaviors. An example of Missing at Random (MAR) is when a patient deliberately refuses to respond to an interview question, particularly when the question pertains to sensitive or private information [5]. Data are classified as Not Missing at Random (NMAR) when the probability of missingness depends on the unobserved values themselves. This implies that the mechanism driving missingness is related to factors that are not directly observable by the researcher [6]. For example, a patient with a lower educational level may avoid responding to questions concerning their educational attainment.

1.2. Analysis Gap

One of the main challenges in analyzing real-world datasets is the presence of missing values in covariates used to predict or explain an outcome. Several imputation techniques for handling missing data are available; however, no single method is universally superior to others, owing to variations in missing data proportions, patterns, and underlying mechanisms. Several studies have examined the performance of imputation methods across diverse datasets that include both numerical and categor-

ical variables. In some cases, simulation studies have been employed to assess the performance of imputation models under a range of conditions [8]. Several studies have integrated imputation techniques within simulation frameworks applied to a wide range of datasets, including clinical datasets with both numerical and categorical covariates [9–11]. In general, these studies conclude that imputation of missing data is essential and appropriate for biomedical applications in which accurate prediction is critical, particularly when missingness is unavoidable. They further indicate that no single imputation method is universally optimal, and that the use of appropriate model performance metrics is necessary to ensure reliable and unbiased results.

However, to the best of current understanding, no study has specifically focused on comparing the efficiency of imputation techniques for numerical covariates using both original (non-simulated) and simulated datasets across varying proportions of missing data and different missingness mechanisms. Building on the aforementioned studies, this work aims to compare statistical imputation approaches using both real and simulated datasets with missing data proportions of 0.15, 0.30, 0.45, and 0.60 under Missing Completely at Random (MCAR) and Missing at Random (MAR) mechanisms, in order to assess the consistency of the applied imputation techniques across different data scenarios. The remainder of this paper is organized as follows: **Section 2** describes the methodological approach adopted in the study. Sections 3 and 4 present the results and discussion, respectively, while **Section 5** provides the study's conclusions.

2. Methodology

2.1. Study Design, Study Area, and Data Description

The study utilized both real and simulated breast cancer datasets, each comprising 693 observations, with covariates containing varying proportions of missing values. The real-world female breast cancer data were obtained from medical records at Muhimbili National Hospital and the Ocean Road Cancer Institute (ORCI) in Tanzania, covering the period from January 2015 to December 2020. The inclusion criteria required patients to be female, diagnosed with breast cancer, and to have received treatment at least once.

The covariates included were respiratory rate (breaths per minute), Body Mass Index (BMI, in kg/m^2), patient age (years), and Body Surface Area (BSA, m^2).

2.2. Imputation Techniques

Various imputation methods are available for handling missing values across different types of datasets; however, given the characteristics of the present data (non-longitudinal with numerical and continuous covariates) and the objectives of the study, this work employed statistically oriented techniques, including mean imputation, multiple imputation via expectation-maximization with bootstrapping (EMB), multiple imputation by chained equations (MICE), and predictive mean matching (PMM), in order to achieve the study's aim. A brief overview of each technique is provided below, as a detailed discussion of their mathematical formulations falls outside the scope of this article.

(i) Imputation by Mean

In this approach, missing values are replaced with the mean of the observed values for that variable [2,9,12]. This method may be used when the proportion of missing values is small, although it may underestimate variability and affect associations among variables. A lower proportion of missing data may have less impact on the overall variance estimate, thereby yielding a more accurate representation of the true association between response and predictor variables [5]. The mean, also referred to as the 'SERIES MEAN (SMEAN)', is computed as $\sum_{i=1}^n x_i/n$ where x_i stands for the numerical covariate for patients $i = 1, 2, \dots, n$; with non-missing data points.

(ii) Multiple Imputations by Chained Equations (MICE)

The approach imputes every missing data value using Q plausible values [13]. It is applicable under both MCAR and MAR mechanisms. Multiple Imputation by Chained Equations (MICE) mitigates potential selection bias that may arise when cases with missing data are excluded from the dataset and reduces the likelihood of obtaining biased standard error estimates [5,6]. The modular implementation comprises three main steps: imputing the missing data q times, analyzing the q imputed datasets, and pooling the results. In the R statistical software environment, MICE implements this workflow by storing outputs in three specialized object classes: "*mids*" (multiple imputed datasets), "*mira*" (multiple imputed repeated analyses), and "*mipo*" (multiple imputed pooled results). The multiple imputation procedure follows this structured framework [4], which consists of the following three steps:

(a). Missing values are replaced Q times to produce the Q completed datasets. (b). Then the Q datasets are analyzed by usual statistical techniques. The results obtained from analyzed Q datasets are combined into one M.I for the aim of making an inference. In this article,

the "*mids*" object contains five multiply imputed datasets generated via chained equations. The "*mira*" object stores the results obtained from the repeated analysis of these five datasets using a logistic regression model, while the "*mipo*" object pools the results from "*mira*" to produce combined estimates for statistical inference. This process is applied under simulation settings with missing proportions of 0.15, 0.30, 0.45, and 0.60 under both MAR and MCAR mechanisms.

(iii) Expectation-Maximization via Bootstrapping (EMB)

In this approach, the multiple imputation process is performed using a bootstrapping algorithm. It uses existing data having n observations to create new Q samples of size n with replacement. The expectation-maximization (EM) algorithm can be summarized as follows: first, it assumes a multivariate normal distribution and initializes starting values for the mean vector (μ) and the variance-covariance matrix (Σ), which are subsequently used to compute the expected value of the likelihood for the imputation model. The likelihood is then maximized, and the model parameters are estimated and updated accordingly. These steps are iteratively repeated until convergence is achieved [13,14]. The practical implementation of EMB in the R package (Amelia II) begins with bootstrapping the incomplete dataset to generate multiple re-sampled datasets. This is followed by the expectation-maximization step applied to each bootstrapped sample to obtain imputed datasets. The resulting datasets are then analyzed separately using standard statistical methods, and the results are subsequently pooled to produce a single consolidated estimate.

(iv) Predictive Mean Matching (PMM)

This multiple imputation method incorporates both parametric and non-parametric strategies. In the parametric phase, Predictive Mean Matching (PMM) generates predicted mean values for each observation in the dataset, which are subsequently used to match complete and incomplete cases. In the non-parametric phase, the nearest-neighbor donor approach is applied, whereby missing values are imputed using observed data points whose predicted means are closest to those of the incomplete cases [15] and [16]. The R program package 'mice' [17] was used to perform five (5) PMM imputations.

2.3. Simulation Experiment

The experiment aimed to evaluate the effectiveness of various methods of imputing covariates and classifiers under different simulation. Simulations were generated from a real breast cancer dataset of 693 cases, which included both observed and missing variable values:

$X_1, X_2, \dots, X_6$, representing the covariates, namely, body mass index, age, respiratory rate, heart rate, body surface area, and recurrence of breast cancer. The data were generated using a fixed mean vector and covariance matrix for each covariate, while varying the proportion of missing values across the datasets ($i = 1 : 4$), varying missing mechanisms ($j = 1 : 2$), and imputation methods ($k = 1 : 4$) producing a total of $4 \times 2 \times 4 = 32$ simulation conditions. The experimental design was implemented in four steps: (i) generation of a complete dataset, (ii) introduction of missingness (amputation), (iii) application of imputation methods, and (iv) evaluation of the effectiveness of the imputation techniques. In step (1), complete datasets of each size $N = 693$ observations were generated from a multivariate normal distribution [18,19] with a mean vector (μ), and positive definite covariance (Σ) matrix; archived by the application of the ‘*mvrnorm*’ function in the package ‘*MASS*’ [20] of the R statistical program. Step 2 involved amputation, that is, introducing missing values into the complete dataset generated in Step 1. This involved creating incomplete data sets with varying pro-

portions of missing values (0.15, 0.30, 0.45, and 0.60) and two missingness mechanisms (MAR and MCAR). The R function, ‘*ampute*’ [21] was used to perform the analysis. Step 3 involved imputing the amputated datasets using the distinct imputation approaches. The function ‘*ampute*’ uses input values including the type of missingness (‘*type* = MAR’ and ‘*type* = MCAR’) and missing probability (example, *prop* = 0.15).

2.4. Evaluation of Imputation Methods

The evaluation of the applied imputation methods focused on two parameters—means and variances—using Root Mean Squared Error (RMSE) across various simulation conditions.

3. Results

Table 1 presents the frequency of missing values, the number of patients corresponding to each count of missing entries, and the proportion of patients with missing values in their records.

Table 1: Distribution of missing data values for all patients.

Missing Data Frequency	Patients Affected by Missing Values	Percentage of Patients Affected by Missing Values
0	261	37.7
1	40	5.8
2	319	46.0
3	29	4.2
4	43	6.2
5	1	0.1
Total	693	100.0

It can be observed that 261 patients (37.7%) have no missing values in their records, and only one subject (0.1%) has five missing values across different variables. Forty subjects (5.8%) have only one missing value from different variables. The number of patients with only two missing values is 319 (46%). The number of patients with 3 and 4 missing values are 29 (4.2%) and 43 (6.2%), respectively.

In the absence of simulation (Table 2), the means and their corresponding standard errors (SE) show minimal variation across the imputed covariates. These values exhibit only slight deviations from those obtained using complete-case analysis, in which all observations containing at least one missing value were excluded.

The RMSE values (Table 3) for each imputation technique increase as the proportion of missing data increases under both MCAR and MAR mechanisms. This pattern reflects a reduction in imputation efficiency as the extent of missingness grows within the datasets. Among

the methods compared, Predictive Mean Matching (PMM) yields the lowest RMSE values, recorded as 5.80 for simulated data with a 15% missing rate under the MAR mechanism and 7.50 for the non-simulated dataset.

Figure 1 indicates that, under the MAR mechanism, both simulated and non-simulated data exhibit similar trends in RMSE values across the applied imputation techniques at a 15% missing rate. The results further demonstrate that Predictive Mean Matching (PMM) outperforms the other methods, as it achieves the lowest RMSE values. For the MCAR mechanism, both simulated and non-simulated data exhibit highly similar trends across the three multiple imputation-based techniques (MICE, EMB, and PMM), with the exception of the series mean (SMEAN) method, which produces substantially different RMSE values. This finding indicates that SMEAN, as a single imputation technique, is less effective compared with multiple imputation-based approaches.

Table 2: Mean and Standard Errors (SE) from imputation techniques as compared to the ones obtained without imputation, i.e., complete-case (C-C) analysis.

Technique	MEAN $\pm$ SE for Each Covariate			
	Age	Respiratory Rate	Body Mass Index	Body Surface Area
C-C	50.46 $\pm$ 0.50	20.84 $\pm$ 0.30	27.71 $\pm$ 0.30	1.69 $\pm$ 0.01
SMEAN	50.45 $\pm$ 0.49	20.84 $\pm$ 0.18	27.71 $\pm$ 0.21	1.69 $\pm$ 0.01
EMB	50.45 $\pm$ 0.49	20.47 $\pm$ 0.23	27.55 $\pm$ 0.27	1.70 $\pm$ 0.01
MICE	50.45 $\pm$ 0.49	20.79 $\pm$ 0.19	27.69 $\pm$ 0.21	1.69 $\pm$ 0.01
PMM	50.46 $\pm$ 0.49	19.17 $\pm$ 0.21	27.85 $\pm$ 0.24	3.91 $\pm$ 0.04

Table 3: RMSE for simulated versus non-simulated imputed data.

Imputation Technique	Missing Mechanism	Percentages of Simulated Missing Data				15% Non-Simulated Data
		15%	30%	45%	60%	
SMEAN	MAR	6.4	10.4	11.2	12.8	8.4
	MCAR	6.9	9.3	11.0	12.7	8.0
MICE	MAR	6.0	10.9	11.6	12.8	8.00
	MCAR	8.0	9.7	11.1	12.8	7.9
EMB	MAR	6.0	11.1	10.8	13.0	8.1
	MCAR	7.1	10.1	11.2	13.1	8.2
PMM	MAR	5.8	10.8	11.3	12.7	7.5
	MCAR	7.0	9.8	11.2	12.5	8.5

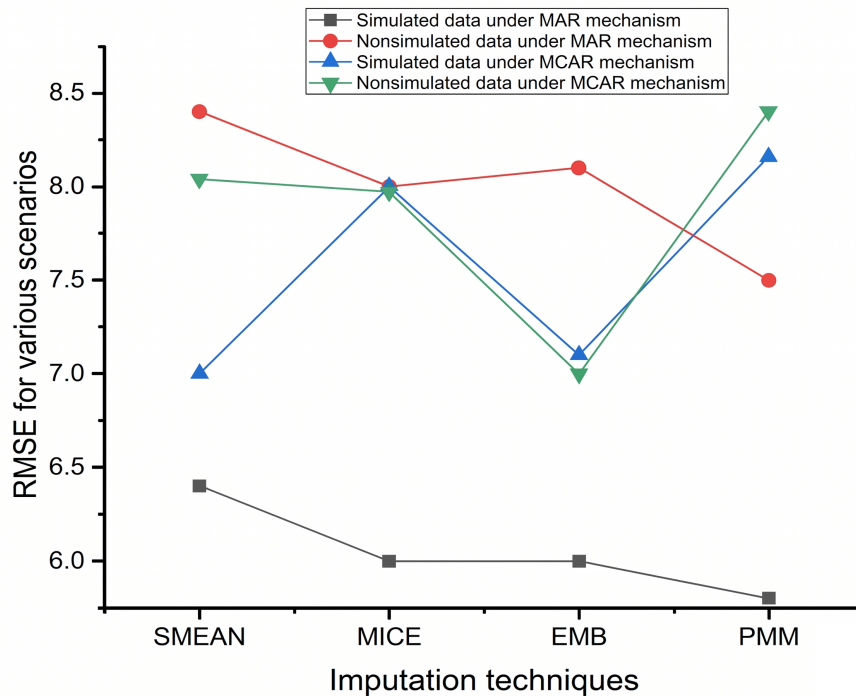


Figure 1: Average RMSE values for simulated versus non-simulated imputed data.

4. Discussion

The study aimed to compare the performance of various imputation techniques using both real and simulated breast cancer datasets containing missing values. Prior to

evaluating the accuracy of the imputation methods, the average RMSE was used to assess their performance in handling missing data generated from simulated breast cancer datasets, considering various proportions of missingness (0.15, 0.30, 0.45, and 0.60) under both MAR and

MCAR mechanisms. Overall, it was observed that, in the absence of simulation settings, the means and their corresponding standard errors for each imputed covariate exhibited minimal variation. These values show slight deviations from those obtained through complete-case analysis. A similar observation was reported in [22], which emphasized that imputation methods should demonstrate improved performance relative to complete-case (CC) analysis. However, [23] examined the impact of selecting imputation methods, including PMM, in clinical datasets and found that, in certain settings, complete-case analysis and PMM may yield comparable performance.

As the proportion of missing values increases under both MAR and MCAR mechanisms, the RMSE values for all applied imputation methods also increase, indicating a decline in imputation efficiency with higher levels of missingness. Notably, Predictive Mean Matching (PMM) emerges as the most efficient imputation technique, as it yields the lowest RMSE values for simulated data with a 15% missing rate under the MAR mechanism, as well as for the non-simulated dataset. This finding is consistent with [24], which also identified Predictive Mean Matching (PMM) as a highly effective imputation technique. It is further supported by [25], which compared multiple imputation approaches and reported that PMM performed reliably across various modelling frameworks, including linear, logistic, and Cox regression models.

The Root Mean Square Error (RMSE) trends indicate that Predictive Mean Matching (PMM) consistently yields the lowest RMSE values, thereby outperforming the other imputation techniques. Under the Missing Completely at Random (MCAR) mechanism, both simulated and non-simulated data exhibit similar RMSE trends across multiple imputation methods—such as Multiple Imputation by Chained Equations (MICE), Expectation-Maximization with Bootstrapping (EMB), and PMM—while the Series Mean Imputation (SMEAN) technique results in substantially higher RMSE values. This suggests that the series mean (SMEAN) method, as a single-imputation approach, is less effective and may introduce bias compared with multiple imputation-based techniques [26]. This observation is further supported by [27], which reported that SMEAN imputation can degrade the performance of statistical estimation methods.

In summary, imputation techniques play a crucial role in health research, particularly when missing values arise from incomplete clinical records and loss to follow-up. This is especially important for key missing covariates such as tumor stage, among others. Accordingly, effective imputation enhances the accuracy of statistical inference and supports more reliable and informed clinical decision-making.

5. Conclusions

Based on the study objectives and findings, the conclusions are summarized in four key points. First, analysis of real datasets demonstrates the impact of imputation on statistical inference, underscoring the importance of appropriate handling of missing data. Second, multiple imputation methods are generally preferred over single imputation approaches, as they provide valid estimates of uncertainty and yield more reliable statistical inferences. Third, simulated datasets with known missing-data mechanisms enable the evaluation of how effectively different imputation methods recover true means and standard errors. Finally, Predictive Mean Matching (PMM) is identified as one of the most effective imputation techniques for numerical variables. Future research should integrate a broader range of missing-data mechanisms in the prediction of survival outcomes for time-to-event datasets with incomplete observations.

List of Abbreviations

BMI	Body Mass Index
BSA	Body Surface Area
CCA	Complete Case Analysis
EMB	Expected Maximization via Bootstrapping
MAR	Missing at Random
MCAR	Missing Completely at Random
MI	Multiple Imputations
MICE	Multiple Imputations Chained Equations
MUHAS	Muhimbili University of Health and Allied Sciences
NMAR	Not Missing at Random
ORCI	Ocean Road Cancer Institute
PMM	Predictive Mean Matching
RMSE	Root Mean Squared Error
SMEAN	Series Mean
UDSM-REC	University of Dar es Salaam Research Ethics Committee

Author Contributions

Conceptualization, data curation, methodology, review and editing: R.A.A. and R.R.A.; Validation, formal analysis, investigation, writing—original draft preparation, and visualization: R.A.A. All authors have read and agreed to the published version of the manuscript.

Availability of Data and Materials

Data supporting the results of this study are available upon request.

Ethical Approval and Consent to Participate

Ethical approval for the study was granted by the University of Dar es Salaam Research Ethics Committee (UDSM-REC) on 2 March 2021, under approval number 2018-07-00106. The Review Boards of MUHAS and ORCI waived the requirement for informed consent, as the data were obtained from patient records.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

The study did not receive any external funding and was conducted using only institutional resources.

Acknowledgments

The authors thank the UDSM-REC for providing the research facilities.

AI Declaration

The authors declare that an artificial intelligence tool (ChatGPT-5) was used solely for grammar and language editing during the manuscript development stage. The text enhanced using ChatGPT-5 was critically reviewed and verified by the authors. The authors take full responsibility for the originality, accuracy, and integrity of the entire work.

References

- [1] M. Humphries, "Missing data & how to deal: an overview of missing data," *Popul. Res. Cent. (Univ. Tex.)*, 2013, Art. no. 45. Accessed: Mar. 28, 2026. [Online]. Available: <https://minio.la.utexas.edu/webeditor-files/prc/pdf/missing-data.pdf>
- [2] M. C. M. De Goeij, M. Van Diepen, K. J. Jager, G. Tripepi, and F. W. Dekker, "Multiple imputation: dealing with missing data," *Nephrol. Dial. Transplant.*, vol. 28, no. 10, pp. 2415–2420, 2013. [CrossRef]
- [3] Z. Zhang, "Missing data imputation: focusing on single imputation," *Ann. Transl. Med.*, vol. 4, no. 1, 2016, Art. no. 9. [CrossRef]
- [4] D. Rubin and R. J. A. Little, *Statistical Analysis with Missing Data*. Hoboken, NJ, USA: John Wiley & Sons, 1987. [Online]. Available: <https://www.wiley.com/en-us/Statistical+Analysis+with+Missing+Data%2C+2nd+Edition-p-9781119013563>
- [5] C. Curley, R. M. Krause, R. Feiock, and C. V. Hawkins, "Dealing with missing data: a comparative exploration of approaches using the integrated city sustainability database," *Urban Aff. Rev.*, vol. 55, no. 2, 2019. [CrossRef]
- [6] A. Z. Alruhaymi and C. J. Kim, "Study on the missing data mechanisms and imputation methods," *Open J. Stat.*, pp. 477–492, 2021. [CrossRef]
- [7] J. Luengo, S. García, and F. Herrera, "On the choice of the best imputation methods for missing values considering three groups of classification methods," *Knowl. Inf. Syst.*, vol. 32, no. 1, pp. 77–108, 2012. [CrossRef]
- [8] H. Kevin A., "Conducting simulation studies in the R programming environment," *Tutor. Quant. Methods Psychol.*, vol. 9, no. 2, pp. 43–60, 2013. [CrossRef]
- [9] J. M. Jerez et al., "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artif. Intell. Med.*, vol. 50, pp. 105–115, 2010. [CrossRef]
- [10] M. Pazhoohesh, S. Walker, and Z. Pourmirza, "A comparison of methods for missing data treatment in building sensor data," in *Proc. IEEE Int. Conf. Smart Energy Grid Eng.*, 2019. [CrossRef]
- [11] J. Hendriksen, G. J. Geersing, K. G. Moons, and J. A. H. de Groot, "Diagnostic and prognostic prediction models," *J. Thromb. Haemost.*, vol. 11, pp. 129–141, 2013. [CrossRef]
- [12] C. A. Glas, "Imputation methods," in *International Encyclopedia of Education*, 3rd ed., Amsterdam, The Netherlands: Elsevier, 2010. [Online]. Available: <https://www.sciencedirect.com/topics/mathematics/multiple-imputation>
- [13] T. Masayoshi, "Multiple ratio imputation by the EMB algorithm," in *Proc. Conf. Eur. Stat. Work. Sess. Stat. Data Ed.*, Budapest, Hungary, Sep. 14–16, 2015. Accessed: May 6, 2026. [Online]. Available: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2015/mtg1/Japan_Multiple_ratio_imputation_by_the_EMB_algorithm.pdf
- [14] C. B. Do and S. Batzoglou, "What is the expectation maximization algorithm?," *Nat. Biotechnol.*, vol. 26, no. 8, pp. 897–899, 2008.
- [15] T. Siswantining, S. M. Soemartojo, and D. Sarwinda, "Multiple imputation with predictive mean matching method for numerical missing data," in *Proc. IEEE Int. Conf. Inf. Commun. Technol. Syst.*, Semarang, Indonesia, 2019, pp. 1–6. [CrossRef]
- [16] B. E. Bailey, R. Andridge, and A. B. Shoben, "Multiple imputation by predictive mean matching in cluster-randomized trials," *BMC Med. Res. Methodol.*, vol. 20, no. 1, pp. 1–16, 2020. [CrossRef]
- [17] S. Van Buuren and K. Groothuis-oudshoorn, "mice: Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, no. 3, 2014. [CrossRef]
- [18] G. Roussas, "Chapter 9 - Some generalizations to k random variables, and three multivariate distributions," in *Introduction to Probability*. Cambridge, MA, USA: Academic Press, 2014, pp. 179–199. [CrossRef]
- [19] J. Tacq, "Multivariate normal distribution," in *International Encyclopedia of Education*. Amsterdam,

- The Netherlands: Elsevier, 2010, pp. 332–338. [CrossRef]
- [20] B. Ripley, B. Venables, D. M. Bates, D. Firth, K. Hornik, and A. Gebhardt, *Support functions and datasets for Venables and Ripley's MASS*. Vienna, Austria: R Foundation for Statistical Computing, 2018, no. 169. [Online]. Available: <https://cran.r-project.org/web/packages/MASS/index.html>
- [21] R. M. Schouten, P. Lugtig, and G. Vink, “Generating missing values for simulation purposes: a multivariate amputation procedure,” *J. Stat. Comput. Simul.*, vol. 9655, 2018. [CrossRef]
- [22] T. P. Morris, I. R. White, and P. Royston, “Tuning multiple imputation by predictive mean matching and local residual draws,” *BMC Med. Res. Methodol.*, vol. 14, no. 1, 2014, Art. no. 75. [CrossRef]
- [23] S. J. Zhan, S. E. Sng, M. E. H. Ong, and F. J. Siddiqui, “Missing data in OHCA registries: how multiple imputation methods affect research conclusions—paper II,” *J. Clin. Med.*, vol. 15, no. 2, 2026, Art. no. 732. [CrossRef]
- [24] R. A. Abassi and A. S. Msengwa, “Classification of breast cancer recurrence based on imputed data: a simulation study,” *BioData Min.*, vol. 15, no. 1, Dec. 2022, Art. no. 30. [CrossRef]
- [25] J. Kampf, I. Dykun, T. Rassaf, and A. Mahabadi, “A comparison of various imputation algorithms for missing data,” *PLoS ONE*, vol. 20, no. 5, May 2025, Art. no. e0319784. [CrossRef]
- [26] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, “A survey on missing data in machine learning,” *J. Big Data*, vol. 8, no. 1, Dec. 2021, Art. no. 140. [CrossRef]
- [27] F. Aracri, M. G. Bianco, A. Quattrone, and A. Sarica, “Bridging the gap: missing data imputation methods and their effect on dementia classification performance,” *Brain Sci.*, vol. 15, no. 6, 2025, Art. no. 639. [CrossRef]

Disclaimer/Publisher’s Note: The views expressed in this article are those of the author(s) and do not necessarily reflect the views of the publisher or editors. The publisher and editors assume no responsibility for any injury or damage resulting from the use of information contained herein.