Computing&AI Connect
ISSN: 3104-4719
2026, Article ID. x, Cite as: https://www.doi.org/10.69709/xxx

SCIFINITI PUBLISHING

OPEN ACCESS

Research Article

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

Rahibu Abdalla Abassi[1*] and Rocky Rajabu Akarro[2]

[1]Department of Natural Sciences, State University of Zanzibar, P.O. Box 146 Zanzibar-Tanzania
Email: r.bassi@suza.ac.tz
ORCID: https://orcid.org/0000-0001-7813-4947
[2]Department of Statistics, University of Dar es Salaam, P.O.Box 35091, Dar es Salaam, Tanzania
Email: akarror@udsm.ac.tz ; akarror@gmail.com
ORCID: https://orcid.org/0009-0002-2144-173X

[*]Corresponding author.

Abstract

Missing data frequently occur in research and if not properly addressed before analysis, can adversely affect the validity of findings. This article evaluates the efficiency of various imputation techniques as formal methods for replacing missing covariates' data. Root Mean Squared Error (RMSE) were calculated for each missing data under mechanisms of MCAR and MAR to ascertain the method of imputation that yield lower values of RMSE under various simulation conditions. The results show that the RMSE for every applied imputation technique increased as proportional of missing data got increased under both MAR and MCAR mechanisms. Under MCAR mechanism, both simulated and non-simulated data provided quite similar trends for three multiple imputation-based techniques; Multiple Imputations Chained Equations (MICE), Expected Maximisation via Bootstrapping (EMB), and Predictive Mean Matching (PMM) except for single-based technique (Series MEAN) that yield RMSE values that substantially different. Amongst applied Multiple Imputations (MI) techniques, the PMM techniques yields the least values of RMSE 5.80 and 7.50, respectively for imputed simulated data with 15% missing rate under MAR mechanism and non-simulated data. The study indicates that when treating missing data, the utilization of multiple imputation techniques is preferable, as they address uncertainty and improve efficiency. It is recommended to compare findings from both imputed and original datasets to evaluate how missing data influences the analysis. For clarity, researchers should also present the means and standard errors for both imputed and non-imputed data.

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

1. Introduction

1.1 Background

Missing data-values is common scenario in medical research, and they can impose inaccurate statistical inference when are not properly handled; in fact, ignoring missing values may result

biased estimates. Data can be missed owing to various reasons including computation, random errors with equipment or natural events such as death, non-response to unclear or sensitive questions in a survey and patients failing to attend clinical routine[1]. Most researchers often remove incomplete data values prior to analysis, an approach called complete case analysis (CCA). The approach excludes any data row that contains at least one missing value during statistical estimation and thus reducing statistical power and hence biasing the results[2]. A plausible way to treat this problem is to apply imputation technique. The techniques estimate and fill-in missing values to create a complete dataset [3].

Imputation methods are mainly categorized as single (such as mean imputation, hot deck, k-nearest neighbour imputations) or multiple imputations, like multiple imputations by chained equations, predictive mean matching, and expected-maximization via bootstrapping among others. Albeit these methods are used to replace missing values under different scenario, there is conclusive evidence to what approach is best under set of conditions such as missingness proportions, patterns, and mechanisms. Under these circumstances, simulation analysis plays a crucial role to check the suitability of imputation methods.

Every data value in experiment or observation has a chance to be missing. The missing mechanisms are categorized as Missing Completely At Random (MCAR), Missing At Random (MAR), and Not Missing At Random (NMAR) [4]. Within MCAR mechanism, missing values neither depend on observed nor on unobserved values [5], [2]. For instance, MCAR data arises when a patient's urine sample is accidentally dropped and broken, leading to missing laboratory results. Data are classified as MAR when the missing values depend solely on information that is already observed [2] [6], [7]. The reason for missing values is associated with patient behaviour that are known. E.g. of MAR data value is when a patient intentionally rejects to respond to an interview question, especially if a question is about personal privacy[5]. The NMAR data-value obtained when a distribution of data that has values rely on missing data. This indicates that missingness probability is related to behaviours that a researcher cannot be aware of[6]. Example: when a low-levelled education patient avoids answering questions that concerns his/her level educational status.

1.2 Analysis gap

Among most challenging issue prior to analysing many real-life datasets is the occurrence of missing values in covariates that are employed to predict or explain a particular outcome. Several imputation techniques for replacing missing data are available: however, there is no particular imputation method that is superior to the others due to various missing data probabilities and mechanisms surrounding the occurrence of missing data. Several studies have examined the performance of imputation methods across diverse datasets that include both numerical and categorical variables. In some situations, simulation studies have been applied to determine how sufficiently an imputation model performs under various conditions[8]. Examples of studies combining the imputation techniques with simulation settings from wide ranges of datasets including clinical aspects using both numerical and categorical covariates [9], [10], and [11] among others. In common, these studies conclude that imputation of missing data is essential and suitable for biomedical applications where accurate predictions are crucial especially when missing data is unavoidable, no single imputation method is generally best, and a model performance metric is needed to yield unbiased results.

However, according to our current understanding, none of them has specifically focused on comparing the efficiency of imputation techniques on numerical covariates based on original (non-simulated) and simulated datasets under various percentages of missing data and missing

mechanisms. Based on mentioned works, this study intends to compare statistical imputation approaches using real and simulated datasets with 0.15, 0.30, 0.45, and 0.60 proportions of missing values under missingness mechanisms of MAR and MCAR to ascertain how consistent each applied imputation techniques are across various scenario of datasets. For the rest of the work, Section 2 desribes methodological approach of the paper, Sections 3 and 4 present the findings and discussion of the study respectively, while Section 5 provides conclusions of the study.

2 Methodology

2.1 Study design, Study area, and data description

The study uses both real and simulated (with covariate containing varying percentages of missing values) breast cancer datasets, each having 693 observations. Actual female breast cancer data were obtained from medical records at Muhimbili National Hospital and Ocean Road Cancer Institute (ORCI) in Tanzania, duration from January 2015 to December 2020. Inclusion criteria for study patients were female gender diagnosed with breast cancer, and received treatement at least once.

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

2.2 Imputation techniques

Various imputation methods exist for handling missing values across different types of datasets; however, considering the characteristics of the current data (non-longitudinal with numerical and continuous covariates) and study purpose, this article used the statistical-oriented techniques include mean imputation, multiple imputations by expectation-maximisation via bootstrapping, multiple imputations by chained equations, and predictive mean matching to meet the study's aim. A brief overview of each technique is provided below, as an in-depth discussion of their mathematical details falls outside the scope of this article.

(i). Imputation by Mean: Via this approach, missing values are substituted by the mean score of known values for that variable[2], [9], [12]. This method is plausible when the percentage of missing values is small, and the sample sizes are not large. A lower amount of missing data results in a smaller effect on the overall variance estimate, thereby providing a more accurate reflection of the true association between the 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 is stands 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. MICE minimize potential selection bias that could arise whenever the cases containing missing data were not included from the dataset and reduces the probability of obtaining biased standard errors [5], [6]. The modular implementation involves three main procedures: imputing the missing data q times, analysing the q imputed datasets, and combining the results. In R statistical software, MICE execute these steps by storing outputs in three specialized classes: "*mids* (multiple imputed datasets)", "*mira* (multiple imputed repeated analyses)", and "*mipo* (multiple imputed pooled results)". The process of multiple imputations follows this structured framework[4] involves the following three steps, namely:
(a). Missing values are replaced Q times to produce the Q completed datasets. (b). Then the Q datasets are analysed by usual statistical techniques. The results obtained from analysed Q datasets are combined into one M.I for the aim of making inference. In this article, the class

'mipo' contains 5 multiply imputed data sets via chained equations, 'mira' stores the results of repeatedly analysed 5 datasets via logistic regression model, and 'mipo' combines or pools the obtained results in 'mira' to yield an average result to be used for inferences under simulation case for 0.15, 0.30. 0.45, and 0.60 of missing values with MAR and MCAR missing mechanisms.

(iii). Expectation-Maximization via Bootstrapping (EMB): In this approach, multiple imputation process is performed based on bootstrapping algorithm. It uses existing data having n observations to create new Q samples of size n with replacement. The expectation-maximization (EM) algorithm is summarized as follows: Firstly, it utilises multivariate normal distribution and create starting values for mean, μ and variance-covariance matrix, $\sum$ which are later used to compute expected value for likelihood of imputation model. The likelihood is then maximised, and model's parameters are estimated and updated. The steps are repeated several times to reach convergence of the values [13], [14]. The practical implementation of EMB in R program (Amelia II) began with bootstrapping an incomplete dataset to generate several bootstrapped datasets, followed by expected-maximisation stage of these data which were then the imputed and analysed separately by standard statistical method, and the results were pooled to yield single results.

(iv). Predictive Mean Matching (PMM)
This multiple imputation method incorporates both parametric and non-parametric strategies. In parametric phase, PMM generates a predicted mean value for every observation in the dataset, which is later applied to pair complete and incomplete cases. In non-parametric phase, the Nearest Neighbour Donor techniques is applied, where the missing value is imputed using the observed data point whose predictive mean is closest to that of the missing case [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 numerous simulation scenarios. Simulations were generated from a real breast cancer dataset of 693 cases, which included both observed and missing variable values: $X_1, X_2, \ldots, X_6$, representing the covariates, namely; body mass index, age, respiratory rate, heart rate, body surface area, and recurrence of breast cancer. Data were created using fixed mean vector and covariance matrix for each covariate while changing the proportion of missing values in data sets ($i = 1{:}\,4$), 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 built in implemented 4 steps, namely, generate complete data dataset, Amputation, Imputations, and 'Evaluate the effectiveness of method of imputation. In step (1) create complete datasets each of size $N = 693$ rows of data from multivariate normal distribution[18], [19] with means vector (μ), and positive definite covariance ($\sum$) matrix given below; archived by the application of '*mvrnorm*' function in the package '*MASS*'[20] of R statistical program. Step (2) 'Amputation' (i.e., making missing values from the complete data set, in step 1). This involved creating incomplete data sets with varying proportion of missing values: 0.15, 0.30, 0.45, and 0.60; and two missingness mechanisms (MAR and MCAR). The R function, '*amput*'[21] was used to perform the analysis. Third step (3) was to impute the amputee data sets using the distinct imputation approaches. The function 'amput' uses inputs values including type of missingness ('$type$ =MAR' and '$type$ =MCAR') and missing probability (example, $prop = 0.15$).

2.4 Evaluation of imputation methods

The evaluation of applied imputation methods was focused on two parameters; means and variances using Root Mean Squared Errors (RMSE) under various simulations conditions.

3. Results

The Table 1 shows frequency of missing values, number of patients with the count of missing values at each frequency and the proportion of patients having missing values in their records.

Table 1: Distribution of missing data values for all patients

Missing data frequency	Patients affected by with 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 (0.1%) subject has five missing values from 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.

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

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

Under absence of simulation (Table 2), the means and their respective standard errors (SE) are not differing much for each imputed covariate. Their values are slightly deviate from the ones obtained when complete-case analysis (discarding all rows containing at least one missing values) were carried out.

The values of RMSE (Table 3) for each imputation technique tend to increase as percentages of missing data are increased under both MAR and MCAR missing mechanisms. This reflects the reduction of efficiency of imputations as missing proportions increase in datasets. The

PMM techniques yields the least values of RMSE 5.80 and 7.50, respectively for imputed simulated data with 15% missing rate under MAR mechanism and non-simulated data.

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 indicates that, under MAR mechanism both simulated and no-simulated data yield similar trends of RMSE values over imputation techniques performed at 15% missing rate. The trend shows that PMM outperformed the remaining technique as it attains lowest values of RMSE. For the MCAR mechanism, both simulated and non-simulated data provided very similar trends for three multiple imputation-based techniques (MICE, EMB and PMM) except for the SMEAN that results in RMSE values that substantially different. This indicates that the SMEAN (a single imputation-based) technique is less effective compared to the multiple-imputation based techniques.

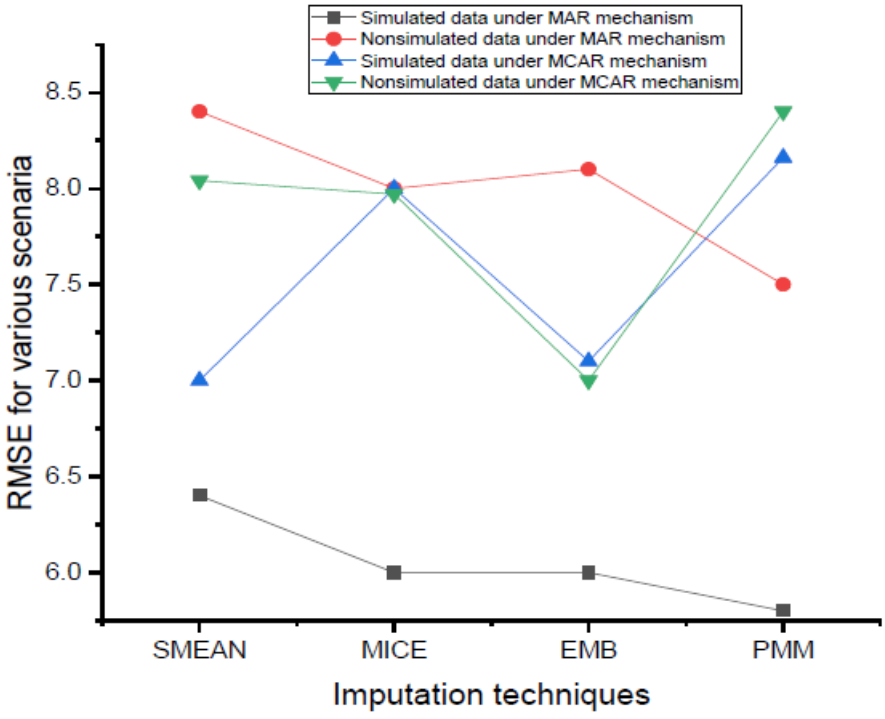

Figure 1: Average RMSE values for simulated versus non-simulated imputed data

4. Discussion

The study intended to compare performance of various imputation techniques based on 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 missing values (0.15, 0.3, 0.45, and 0.60) and mechanisms values (MAR and MCAR). Overall, it was observed that, without the influence of simulation settings, the means and their associated standard errors for each imputed covariate showed minimal variation. These values show slight deviations from those obtained through complete-case analysis. Similar observation was obtained from [22] which emphasized that imputations should improve over C-C technique; however, [23] assessed the impact of selecting imputation methods including PMM in clinical data and found that in some settings the methods (CCA and PMM) are performing similarly.

With increasing percentages of missing values under mechanisms of MAR and MCAR, the RMSE values for all applied methods of imputation also increase, reflecting a decline in imputation efficiency as the proportion of missing data grows. Notably, the Predictive Mean Matching (PMM) most efficient imputation technique as it yields the lowest RMSE values for imputed simulated data with a 15% missing rate under the MAR mechanism and for non-simulated missing data. This finding is in line with [24] which also found PMM as most plausible imputation technique and also supported by [25] who compared imputation techniques and found that PMM was plausible in models including linear, logistic, and Cox regression.

The Root Mean Square Error (RMSE) trends indicate that Predictive Mean Matching (PMM) consistently achieves the lowest RMSE values, outperforming 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 Single Mean Imputation (SMEAN) technique results in substantially higher RMSE values. This suggests that SMEAN, being a single imputation method, is less effective and can produce bias or compared to multiple imputation-based techniques [26]. This observation is also in line with [27] who claimed that series mean (SMEAN) imputation degrades the performance of estimation methods.

To sum up the discussion, imputation techniques play important role in health research, when missing values obtained from incomplete clinical records or/and loss to follow-up. This is crucial for missing covariates like tumor stage among others. Therefore, a successful imputation process improves accuracy of statistical inferences and results in informed clinical decisions'

5. Conclusions

Based on study under objectives and findings, the article's conclusions are focused on four points: first, analysing real datasets demonstrates how imputation affects statistical inferences, highlighting the importance of appropriate handling of missing data; second, multiple imputation methods are generally favoured over single imputation approaches, as they provide valid estimates of uncertainty, leading to more reliable statistical inferences; third, simulated datasets with known missing data mechanisms enable the assessment of how well different imputation methods recover true means and standard errors; and, predictive mean matching

has been identified as one of the most effective imputation techniques for numerical variables. With respect to the current study, further research should focus on combination of a range of missing mechanisms on predicting survival outcomes for missing time-to-event datasets.

List of Abbreviations

BMI: Body Mass Index
BSA: Body Surface Area
CCA: Complete Case Analysis
EMB: Expected Maximisation 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 Science
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

Ethical Approval and Consent to Participate

University of Dar es Salaam Research Ethics Committee (UDSM-REC) issued ethical approval for the study at 02 March 2021. The approval was registered under number 2018-07-00106. Review Board of MUHAS and ORCI waived the requirement for informed consent, as the data were obtained from patient records.

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.

Conflicts of Interest: Authors of this paper declare no conflicts of interest.

Funding: No funding was acquired for this research.

Availability of Data and Materials

Data supporting the results of this study are not publicly available but can be requested from the author on need.

Acknowledgment

We thank the UDSM-REC for providing the research facilities.

AI-declaration: The authors declare that the artificial intelligence tool (ChatGPT-5) was used to improve grammar and language editing only during a manuscript development stage. All ChatGPT-5 generated texts were critically reviewed and verified by the authors, and authors take responsibility for originality, accuracy, and integrity of the entire work.

References

[1] M. Humphries, “Missing Data & How to Deal: An overview of missing data,” *Population Research Center (University of Texas)*, p. 45, 2013. Retrieved from 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,” no. May, pp. 2415–2420, 2013, doi: 10.1093/ndt/gft221.

[3] Z. Zhang. “Missing Data Imputation: Focusing on Single Imputation.” Annals of Translational Medicine, vol. 4, no. 1, 2016, p. 9, doi:10.3978/j.issn.2305-5839.2015.12.38.

[4] Little and Rubin, *Statistical Analysis with Missing Data*. John Willey & Sons, 1987.

[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,” 2019, doi: 10.1177/1078087417726394.

[6] A. Z. Alruhaymi and C. J. Kim, “Study on the Missing Data Mechanisms and Imputation Methods,” pp. 477–492, 2021, doi: 10.4236/ojs.2021.114030.

[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.” *Knowledge and Information Systems*, vol. 32, no. 1, 2012, pp. 77–108, doi:10.1007/s10115-011-0424-2

[8] H, Kevin A. “Conducting Simulation Studies in the R Programming Environment.” *Tutorials in Quantitative Methods for Psychology*, vol. 9, no. 2, 2013, pp. 43–60, doi:10.20982/tqmp.09.2.p043.

[9] Jerez *et al.*, “Artificial Intelligence in Medicine Missing data imputation using statistical and machine learning methods in a real breast cancer problem,” vol. 50, pp. 105–115, 2010, doi: 10.1016/j.artmed.2010.05.002.

[10] M. Pazhoohesh, S. Walker, and Z. Pourmirza, “A comparison of Methods for Missing data treatment in building sensor data,” 2019.

[11] J. Hendriksen, G. J. Geersing, K. G. Moons, and G. A. H, “Diagnostic and prognostic prediction models,” vol. 11, pp. 129–141, 2013, doi: 10.1111/jth.12262.

[12] C. A. Glas, “Imputation Methods,” *International Encyclopedia of Education*, no. Third Edition, 2010.

[13] T. Masayoshi, “Multiple Ratio Imputation by the EMB Algorithm Assumptions of Missing Mechanisms,” 2015.

[14] C. B. Do and S. Batzoglou, “What is the expectation maximization algorithm?,” 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," 2019. doi: 10.1109/ICICoS48119.2019.8982510.

[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, doi: 10.1186/s12874-020-00948-6.

[17] S. Van Buuren and K. Groothuis-oudshoorn, "Multivariate Imputation by Chained Equation," *J Stat Softw*, vol. 45, no. 3, 2014, doi: 10.18637/jss.v045.i03.

[18] G. Roussas, "Some Generalizations to k Random Variables, and Three Multivariate Distributions," *Introduction to Probability*, pp. 179–199, 2014, doi: 10.1016/B978-0-12-800041-0.00009-2.

[19] J. Tacq, "Multivariate normal distribution," *International Encyclopedia of Education*, pp. 332–338, 2010, doi: 10.1016/B978-0-08-044894-7.01351-8.

[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," p. 169, 2018.

[21] R. M. Schouten, P. Lugtig, and G. Vink, "Generating missing values for simulation purposes: a multivariate amputation procedure," vol. 9655, 2018, doi: 10.1080/00949655.2018.1491577.

[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. 75, 2014.

[23] S. J. Zhan, Sng, S. E., Ong, M. E. H., & Siddiqui, F. J. (2026). Missing Data in OHCA Registries: How Multiple Imputation Methods Affect Research Conclusions—Paper II. *Journal of Clinical Medicine*, 15(2), 732. DOI: 10.3390/jcm15020732.

[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, doi: 10.1186/s13040-022-00316-8.

[25] J. Kampf, I. Dykun, T. Rassaf, A. Mahabadi. A comparison of various imputation algorithms for missing data. PLoS One. 2025 May 12;20(5):e0319784. doi: 10.1371/journal.pone.0319784.

[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, doi: 10.1186/s40537-021-00516-9.

[27] F.A. Siddiqui, A. Sharma. Bridging the gap: Missing data imputation methods and their effect on dementia classification performance. *Brain Sci*. 2024;14(6):639.

Computing&AI Connect
2026, Vol. 3
doi:10.x/journal.x.x.x

SCIFINITI
PUBLISHING