Comparison of Imputation Methods on Retrospective Breast Cancer Data in Tanzania: A Case Study of Muhimbili and Ocean Road Hospitals

Rahibu A. Abassi (✉ rahibuabassi@yahoo.com)
State University of Zanzibar
Amina S. Msengwa
University of Dar es Salaam
Rocky R. J. Akarro
University of Dar es Salaam

Research Article

Keywords: Breast cancer dataset, Classification methods, Imputation methods, Missing data

Posted Date: September 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-820770/v1

Comparison of imputation methods on retrospective breast cancer data in Tanzania: A Case study of Muhimbili and Ocean Road hospitals

Rahibu A. Abassi[1*], Amina S. Msengwa[2], Rocky R. J. Akarro[2]

[1]Department of Natural Science, State University of Zanzibar, Zanzibar-Tanzania

[2]Department of Statistics, University of Dar es Salaam, Dar es Salaam-Tanzania

[*]Corresponding author: rahibuabassi@yahoo.com

Abstract

Background: Clinical data are at risk of having missing or incomplete values for several reasons including patients' failure to attend clinical measurements, wrong interpretations of measurements, and measurement recorder's defects. Missing data can significantly affect the analysis and results might be doubtful due to bias caused by omission of missed observation during statistical analysis especially if a dataset is considerably small. The objective of this study is to compare several imputation methods in terms of efficiency in filling-in the missing data so as to increase the prediction and classification accuracy in breast cancer dataset.

Methods: Five imputation methods namely series mean, k-nearest neighbour, hot deck, predictive mean matching, and multiple imputations were applied to replace the missing values to the real breast cancer dataset. The efficiency of imputation methods was compared by using the Root Mean Square Errors and Mean Absolute Errors to obtain a suitable complete dataset. Binary logistic regression and linear discrimination classifiers were applied to the imputed dataset to compare their efficacy on classification and discrimination.

Results: The evaluation of imputation methods revealed that the predictive mean matching method was better off compared to other imputation methods. In addition, the binary logistic regression and linear discriminant analyses yield almost similar values on overall classification rates, sensitivity and specificity.

Conclusion: The predictive mean matching imputation showed higher accuracy in estimating and replacing missing/incomplete data values in a real breast cancer dataset under the study. It is a more effective and good method to handle missing data in this scenario. We recommend to replace missing data by using predictive mean matching since it is a plausible approach toward multiple imputations for numerical variables, as it improves estimation and prediction accuracy over the use complete-case analysis especially when percentage of missing data is not very small.

Keywords: Breast cancer dataset, Classification methods, Imputation methods, Missing data

Background

Missing or incomplete data is a common challenge encountered by almost every researcher in both clinical and non-clinical settings. Data might be missing due to a variety of reasons, for example in a clinical context, missing data may arise because of random errors with measuring equipment or computations, attrition due to social or natural processes for instance death, non-response to some sensitive or unclear questions that the patients do not feel comfortable to answer, and study subjects failing to report to a routine clinic [1].

Cases containing missing values produce different results due to loss of power, precision and increased bias caused by analysis of incomplete datasets especially when the datasets are small. This situation necessitates the researchers to find appropriate ways used to attempt utilization of all available data so that the results of their works can be more desirable in terms of precision and overall study power. The process that is used to fill in or replace with missing data is called imputation [2].

Breast cancer databases are at risk of having missing values due to: human and equipment errors, patients' missing measurements, wrong interpretations of measurements and recorder's faults [3]. Several breast cancer studies conducted in Tanzania including did not indicate how missing values were handled before making statistical analyses and inferences [4], [5], [6], [7], and [8] among others.

The current study aims to fill this knowledge gap by applying imputation techniques to breast cancer dataset. The study focused on comparing different imputation methods on a real retrospective breast cancer datasets of female patients admitted to two largest breast cancer clinics namely Muhimbili National Hospital and Ocean Road Cancer Institute in Tanzania. The efficiency of each imputation method under the study was evaluated by the 'Root Mean Squared Error' and 'Mean Absolute Error' [9].

Imputation techniques for treating datasets with missing values are vast; however, the common imputation methods compared in this paper are: (series) mean, hot deck, predictive mean matching, k-nearest neighbour, and multiple imputations (via Markov Chain Monte Carlo (MCMC) algorithm as well as Amelia II program).

Missing Data Mechanism and Pattern

Every observation in a dataset has the probability to be missed, this probability is described by 'missing data mechanism'. Missing data mechanisms are categorized into; Missing Completely At Random (MCAR), Missing At Random (MAR) and Not Missing At Random (NMAR).

Assumptions behind these mechanisms can affect imputation methods and their results if they are not properly checked [10]. The patterns of missing data show how the missing values are distributed over variables containing missing data.

There are three types of missing data patterns, namely; univariate, monotone, and arbitrary missingness patterns. For a dataset with k variables: $Y_1, Y_2, \ldots, Y_k$. A univariate pattern is when missing data are found on at least one of the k variables for the same participant. A monotone pattern of missing data arises such that if Y_i is missed then the subsequent data $Y_{i+1}, Y_{i+2}, \ldots, Y_P$ are also missed. An arbitrary pattern arise when missing data is found in any of k variables randomly for any study participant [11].

In MCAR, missing values do not depend on the values from both observed and unobserved ones in a dataset. In a clinical context, an example of MCAR data is when a patient is un-deliberately fails to provide an answer to a question that is used in the analysis. MCAR assumption is checked by Little's MCAR test under the null hypothesis that 'data are MCAR'. The MAR mechanism is when missing values depend only on observed data. That is, under the MAR distribution of dataset containing missing values depends on observed values, but not on the missing ones. An example of MAR data is when respondents deliberatively decide not to answer a question, especially if the question is about his or her privacy issues. The NMAR occurs if the distribution of dataset containing missing values depends on missing values. No approximation of the missing values can be made in NMAR by a researcher since other variables' values are not observed as well [12].

Methods

The major aim of this paper is to compare the efficiency of several imputation methods in replacing values missing data on real breast cancer dataset. The prediction and classification algorithms were applied to both datasets (the original one with missing data points and the one resulted from plausible imputation, according to minimum values of Root Mean Squared Errors and Mean Absolute Errors) to see whether the imputation-based results gain or/and improve accuracy over complete-case analysis. The summary of methodological aspects applied in the study are found in sections A, B, C and D.

A. Study design, site and data description

The study design was retrospective cross-sectional whereby the past breast cancer patients' records were used. Dataset was extracted from available patients' breast cancer medical records at Muhimbili National Hospital (MNH) and Ocean Road Cancer Institute (ORCI) in Tanzania. These hospitals were chosen because they are the only major health centres that diagnose and treat breast cancer, among all other types of cancers in Tanzania. The list of all registered female-breast cancer patients at MNH and ORCI from January 2015 to December 2020 was used as the study population. The total number of female breast cancer patients registered was 4390 (Database-MNH, 2021). About 2461 were then included in the sampling frame. The study used 345 sample units from MNH. This number was calculated using formula by Yamane, (1967) with a population size of 2461 and a margin error (e) of 5%.

$$n = \frac{N}{1 + Ne^2} \quad = \frac{2461}{1 + 2461(0.05)^2} \approx 345$$

A similar sampling procedure was repeated for ORCI to get a sample of size 348 patients from a sampling frame of 2658 females from their medical database. The final sample size consisted of 693 (345 from MNH and 348 from ORCI). A simple random sampling was then applied to identify the patients' file numbers as sample units from both clinics. The study variables were extracted from several previous related studies concerning breast cancer [6], [13], and [14]. The dependent variable is a 'cancer recurrence', with two response values; 'yes and no'. The response 'yes' means cancer comes back after recommended treatment, 'no' indicates that cancer does not come back after got a respective treatment. The independent variables were: Age of patient (in years), Body Mass Index (BMI) in kg per squared metres, Respiratory rate (in breaths per minutes), and Body Surface Area (BSA) in squared metres.

B. Methods of imputation

1. **Mean imputation**: The idea based on this approach is to use a mean value of each non-missing variable to fill in missed values for all observations [13]. The mean imputation technique is more appropriate when the amount of missingness is small whilst the size of the sample is large. The lesser the degree of missingness, the smaller impact on the overall estimate of variance, and hence, the good reflection of the true association between the response and predictor variables [12]. The mean or sometimes, called 'series mean' is calculated as $\sum_{i=1}^{n} x_i/n$ where x_i is a numerical variable

and $i = 1,2,\dots,n$; number of subjects with observed data values. In this study, the 'series mean' in command SPSS (version 25) was used to replace missing values of each numerical variable under the study.

2. **Hot deck imputation:** Each missing value is replaced by the observed value from 'identical unit'. The application of hot deck imputation techniques has been common in both epidemiological as well as in medical research settings. The method replaces missing data values of at least one variable for a subject with no response, known as 'recipient' with observed data values from a subject with the response, known as 'donor' [15]. The method needs the data with MCAR or MAR mechanism [12]. Consider the values $x_i = (x_{i1},\dots,x_{ip})$ for subject i of p covariates. For a matching recipient i and a donor j, the proximity of potential candidate donors to recipients is defined by maximum deviation given by: $D_{(i,j)} = \max_k|x_{ik} - x_{jk}|$, for nicely scaled x_k so that the comparability of difference (through ranks and standardization) can be made [15]. In this study, the hot deck imputation was employed by using function 'hot deck' from the 'VIM' (Visualization and Imputation of Missing Values) package [16] in R statistical software (version 3.6.3)
3. **The multiple imputations (MI):** Method is based on the idea of replacing each of the missed values in the dataset with a set of P acceptable values. These values are drawn from the distribution of the data at hand, and they represent the values that are more likely to be right for imputation. The Bayesian approach is used to draw the P acceptable values from 'conditional predictive distribution' containing missing values [17]. The algorithm for MI involves the three steps according to [10].

a) Missing data are filled-in P times to yield the P completed datasets.
b) The P completed datasets are then analysed by standard statistical methods.
c) The results from analysis of P completed datasets are pooled into one multiple imputations to draw inference.

The MI method works MAR missingness mechanisms. In this work, we use both; the Amelia II, "a complete R package for MI of missing data" (Honaker et al., 2011) and the MCMC (Markov Chain Monte Carlo) algorithm in SPSS to impute original data 5 times. The pooled dataset was obtained from both programs (Amelia II and SPSS).

4. **Predictive Mean Matching (PMM):** The approach utilizes both parametric and non-parametric approaches in the imputation process. At the parametric phase, PMM establishes a predictive mean value corresponding to each observation in data. These predictive means are then used to match complete and incomplete observations. The non-parametric stage applies the method of Nearest Neighbour Donor to produce original data value from non-missing observation having nearest predictive mean distance close to missing one so as to impute a missing data value [18] and [19]. The PMM is robust to model miss-specification and ensures to yield more plausible imputed values than the regression method when the assumption of normality is violated [20]. Assume Y is partially observed sample obtained randomly from q variate multivariate distribution $P(Y|\theta)$, and that the distribution of Y is specified by a vector of unknown parameters, θ. The MICE (Multivariate Imputation by Chained Equations) algorithm obtains the posterior distribution of θ by (iteratively) sampling from conditional distribution $P(Y_1|Y_{-1},\theta_1),\ldots,P(Y_q|Y_{-q},\theta_q)$. The function and package 'mice' in R statistical software [21] was used to perform the PMM imputation five times and the average values were calculated to form a final dataset.

5. **K-Nearest Neighbour (KNN):** A non-parametric approach used to impute missing data by averaging its neighbouring observed data (9). The approach is donor-based in which imputed values are either measured as a single records in the dataset (1-NN) or as an average value obtained from k records (k-NN) [22]. The distance two between observations, and that is used to define the nearest neighbours is defined as $D_{ij} = \frac{\sum_{k=1}^{P} w_k \tau_{i,j,k}}{\sum_{k=1}^{P} w_k}$, where w_k is the weight and $\tau_{i,j,k}$ is the contribution of k^{th}

variable. The ratio of absolute distance to range is used for $\tau_{i,j,k}$ of continuous variables; $\tau_{i,j,k} = \frac{|x_{i,k}-x_{j,k}|}{r_k}$, whereas $x_{i,k}$ is a value of k^{th} variable of i^{th} observation and r_k is the range of k^{th} variable [16]. This study uses the PMM imputation with 5 nearest neighbours by using the R function 'kNN' in the package 'VIM' package.

C. **Evaluation of imputation methods**

The efficiency of five imputation techniques was evaluated by 'Root Mean Squared Error (RMSE)' and 'Mean Absolute Error (MAE). The definition and computation of these measures are based on [9]. RMSE describes the sample standard deviation between observed and imputed values expressed whereas; MAE is a measure of error's average magnitude.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i^{obs}-x_i^{imputed})^2}{n}}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_i^{obs} - x_i^{imputed}\right|$$

Whereas, n stands for the number of samples in a dataset, x_i^{obs} denotes the i^{th} target value, and $x_i^{imputed}$ represents the i^{th} sample's predicted value. Generally speaking, the more effective and good method would have a lower RMSE and MAE [9] and [23].

Classification methods

The binary classification methods namely; logistic regression and linear discriminant analyses were applied on the plausible imputed datasets to see if the classification rates, sensitivity and specificity of the two common classifiers will yield similar or different results. The same procedure was repeated for (complete-case) analysis that discards missing values in original data (un-imputed one) to compare the classification results and investigate if the analysis is done from imputed and un-imputed will differ or not in the breast cancer dataset.

The overall process of analysis, from methods of imputation to classification techniques, is summarized in Figure 1.

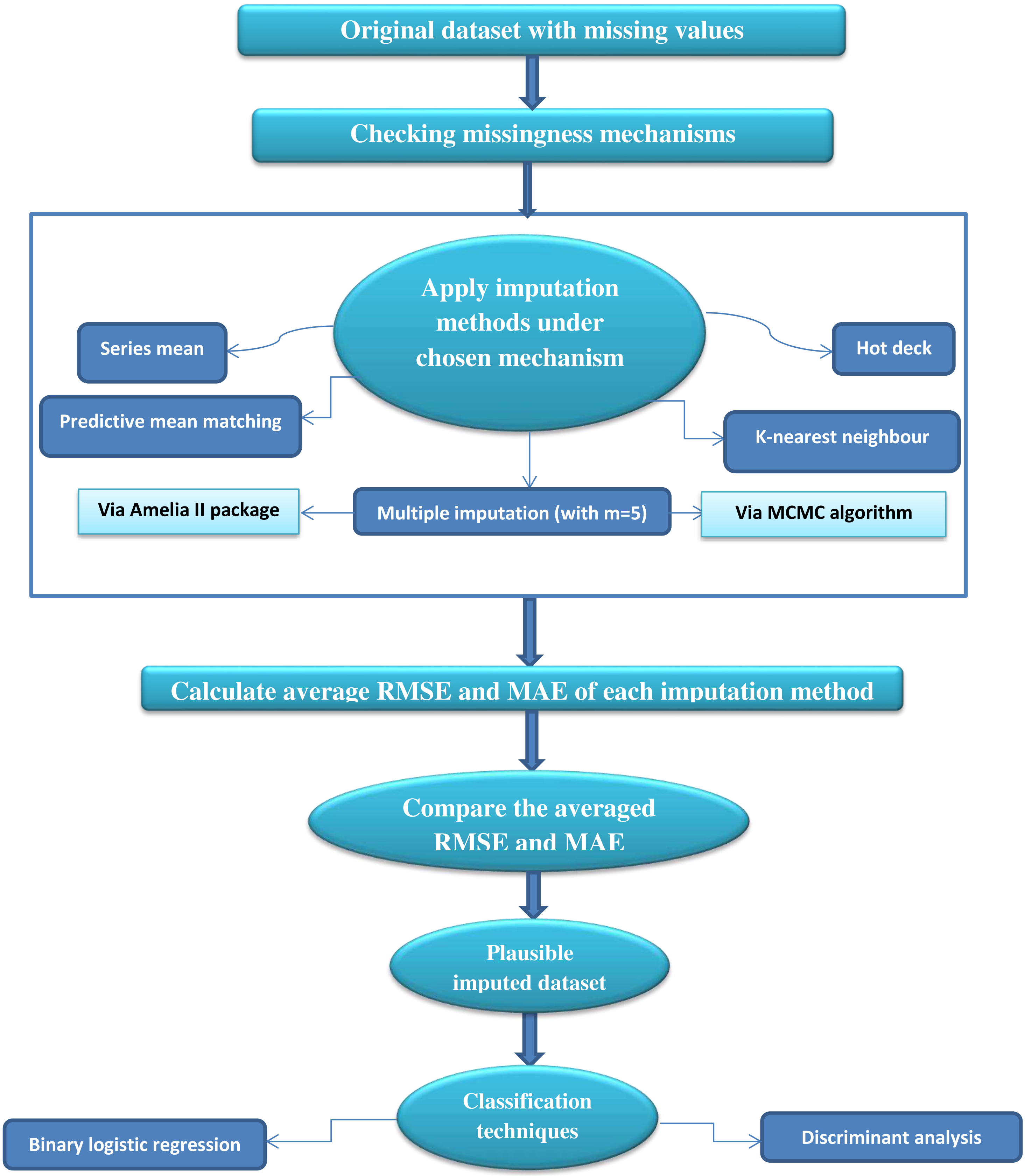
Original dataset with missing values
Checking missingness mechanisms
Apply imputation methods under chosen mechanism
Series mean
Hot deck
Predictive mean matching
K-nearest neighbour
Via Amelia II package
Multiple imputation (with m=5)
Via MCMC algorithm
Calculate average RMSE and MAE of each imputation method
Compare the averaged RMSE and MAE
Plausible imputed dataset
Classification techniques
Binary logistic regression
Discriminant analysis

Figure 1: Process of imputation analysis, modified from [9]

Result

The preliminary analysis was done through exploratory data analysis to get insights into the dataset. Table 1 shows the number of observed and missing observations recorded from (numerical variables of) the sample of 693 female breast cancer patients. It can be seen that the variable 'age of the patient' has the lowest percentage (0.29%) of missing values while the 'number of breaths per minute' appears to have highest percentage (37.95%) of missingness.

Table 1: Numerical descriptive measures

Descriptive measures		Numerical variables under the study			
		Age	Breath rate	BMI	BSA
Number of total observations (N=693)	Observed (%)	691 (99.71)	430 (62.05)	472 68.11)	467 (67.39)
	Missing (%)	2 (0.29)	263 (37.95)	221 (31.89)	226 (32.61)
Mean		50.46	20.84	27.71	1.69
Median		49.00	20.00	26.96	1.70
Mode		38	20	22	2
Std. Deviation		13.010	6.125	6.6	0.223
Range		82	86	47	2
Minimum		18	14	13	1
Maximum		100	100	61	3

Missingness pattern and missing data mechanism

The missing data in the study have an 'arbitral' pattern since missing values appear in the dataset for many variables in a non-systematic manner, they are located randomly in different variables for distinct study participants). The independent-samples t-test and the Little's MCAR test were conducted to gain insight about missingness mechanisms. Two-sample independent t-tests (Table 2) were presented to check if there is 'no significant difference' between observed and missing values of the outcome variable using coded values (1 for observed and 0 for missing values) at 5% level of significance.

Table 2: Two-sample Independent T-test Between Numerical Independent Variables

		Levene's Test for Equality of Variances		Two-samples t - test for Equality of Means	
Variables tested	Variance assumptions	F - value	P - value	T – value	P - value
Age of patient in years	Equal variances assumed	0.079	0.778	-1.678	0.094
	Equal variances not assumed			-1.741	0.091
Respiratory rate in breaths per minute	Equal variances assumed	0.158	0.691	-0.144	0.886
	Equal variances not assumed			-0.296	0.769
Body Mass Index in kg/m^2	Equal variances assumed	0.931	0.335	-1.788	0.074
	Equal variances not assumed			-2.060	0.051
Body surface area in m^2	Equal variances assumed	0.476	0.490	0.012	0.991
	Equal variances not assumed			0.013	0.990

Table 2 reveals that the Levene’s test for equality of variances under the null hypothesis of ‘population variances are equal’ reveals that there is equality of variances (p-values > 0.05) between missing and non-missing values from breast cancer recurrence. All p-values >0.05 from the two-sample independent to t-tests do not allow the rejection of the hypothesis of ‘no significant difference’ between breast cancer recurrence and the missing and non-missing of numerical independent variables. This finding is in line with Little’s MCAR test result under the null hypothesis that ‘data is Missing Completely At Random (MCAR)’. The Little’s MCAR (Chi-Square = 129.973, p-value < 0.001) implies that the data was not under MCAR mechanisms since the p-value is much smaller than 0.05 significance level. The test result indicates the presence of a significant relationship between missing and non-missing values, thus MAR assumption is valid in the data.

Tables 3 and 4 summarize the results from each imputation method based on RMSE and MAE respectively. The Predictive Mean Matching (PMM) attained the lowest averaged values of RMSE and MAE suggesting that PMM imputes the dataset more effectively.

Table 3: Root Mean Square Error (RMSE) from Imputation Methods

Method of imputation	RMSE for each imputed numerical variable				Average RMSE
	Age of patient	Respiratory Rate	Body Mass Index	Body Surface Area	
Hot deck	3.03	13.78	15.78	1.11	8.42
Series mean	2.70	12.84	15.65	0.97	7.86
MI via MCMC	2.67	12.76	15.61	0.96	8.00
MI via Amelia II	2.52	12.81	16.07	0.98	8.09
K Nearest Neighbors	2.58	12.49	15.19	1.16	7.88
Predictive Mean Matching	2.75	10.27	16.19	1.11	7.58

Comparing the imputation techniques (PMM-Predictive Mean Matching, KNN-k Nearest Neighbour, MCMC-Markov Chain Monte Carlo, Mean-series mean, Amelia II- multiple imputations via Amelia package, and Hot deck – imputation) used to fill-in missing values in real breast cancer dataset.

Table 4: Mean Absolute Error (MAE) from Imputation Methods

Method of imputation	MAE for each imputed numerical variable				Average MAE
	Age of patient	Respiratory rate	Body Mass Index	Body Surface Area	
Hot deck	0.16	7.86	8.86	0.79	4.42
Series mean	0.15	7.91	8.84	0.55	4.36
MI via MCMC	0.14	7.85	8.81	0.55	4.34
MI via Amelia II	0.14	7.54	8.82	0.56	4.27
K Nearest Neighbor	0.14	7.68	8.70	0.83	4.34
Predictive Mean Matching	0.15	6.23	9.15	0.81	4.09

Figure 2 reveals that the imputation technique from PMM is more plausible based on lowest values of RMSE and MAE of numerical data with 15% of missing values; and hence in this scenario, PMM method is more effective and good for handling missing data.

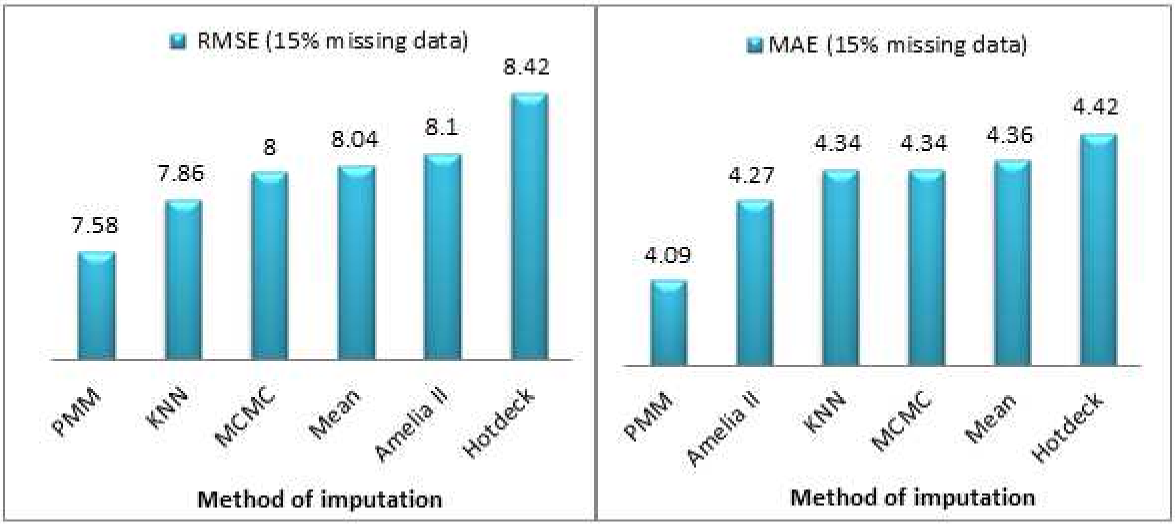

Figure 2: Averaged RMSE and MAE values for different imputation methods

The data resulted from PMM imputation was then used in the classification of observations. The results from classification algorithms are presented in Table 5 for binary logistic regression and linear discriminant analyses. The cut value for classification is 0.5.

Table 5: Classification tables from imputed dataset

Binary Logistic Regression					
Observed		Predicted group membership			
		Breast cancer recurrence			% correct classification
		Yes	No	Total	
Breast cancer recurrence	Yes	8	202	210	3.8
	No	10	473	483	97.9
Total		18	675	693	69.4
Linear Discriminant Analysis					
Observed		Predicted group membership			
		Breast cancer recurrence			% correct classification
		Yes	No	Total	
Breast cancer recurrence	Yes	24	186	210	11.4
	No	31	452	483	93.6
Total		55	638	693	68.7

As a classifier, the logistic regression yields an **overall classification rate** (which is the sum of main diagonal numbers divided by sample size) of 69.4% [= (8+473)/693], for linear discriminant analysis we have 68.7 per cent. These values indicate the probability of correctly classifying breast cancer patients into their respective groups ('recurrence' and 'non-recurrence). Other important probabilistic information obtained from classifiers are:

Sensitivity - the probability of correctly classifying the patients having a breast cancer recurrence. Alternatively, this is the percentage of cases with the observed characteristic 'yes' for breast cancer recurrence which was correctly predicted by a classifier. In this study, sensitivity is equal to 44.4% or 0.444 ($= 8/18$) for the logistic regression classifier. It is often referred to as 'true positives. For linear discriminative analysis, this is about 44% or 0.436 ($= 24/55$).

Specificity - the probability of correctly classifying the patients having non-recurrence breast cancer. Alternatively, this is the percentage of cases with observed characteristic 'no' for breast cancer recurrence which were correctly predicted by the fitted model. In this study, specificity is about 70 % or 0.701 ($= 473/675$). It is often referred to as 'true negatives'.

False-positive rate – probability or ratio between the number of negative events (non-recurrence breast cancer cases) wrongly classified as positive (recurrence breast cancer cases) and a total number of actual negative events. In this study, the false positive rate is about 30% or 0.299 ($= 202/675$) for binary logistic classifier and 29.2% or 0.292 ($= 186/638$) for linear discriminant analysis. It indicates the probability that a 'cancer recurrence' is present while in fact, it is absent.

False-negative rate – the probability or ratio between the number of positive events (recurrence breast cancer cases) wrongly classified as negative (non-recurrence breast cancer cases) and a total number of actual positive events. In this study, the false positive rate is 55.6% ($= 10/18$) for binary logistic classifier and 56.4%($= 31/55$) linear discriminant analysis. It indicates the probability that a 'cancer recurrence' is absent while in fact, it is present. **Prevalence** – the ratio or probability that any breast cancer patient has a 'cancer recurrence' in the population. In this study, a prevalence is about 2.6% or 0.026 ($= 18/693$) predicted by binary logistic classifier, and 7.9% or 0.079

(= 55/693) linear discriminant analysis. This suggests that every female having breast cancer in Tanzania has a chance not more than 8% to get cancer recurrence after getting a recommended treatment. The Receiver Operating Characteristic (ROC) curves were plotted to explore on how individual variables have the ability to discriminate between patients or observations that fell into 'yes' category and 'no' categories regarding breast cancer recurrence. Figure 3 displays the ROC for variables under the study. Table 6 displays the areas under the ROC curves for each variable used in classification under the null hypothesis of 'true area =0.5'.

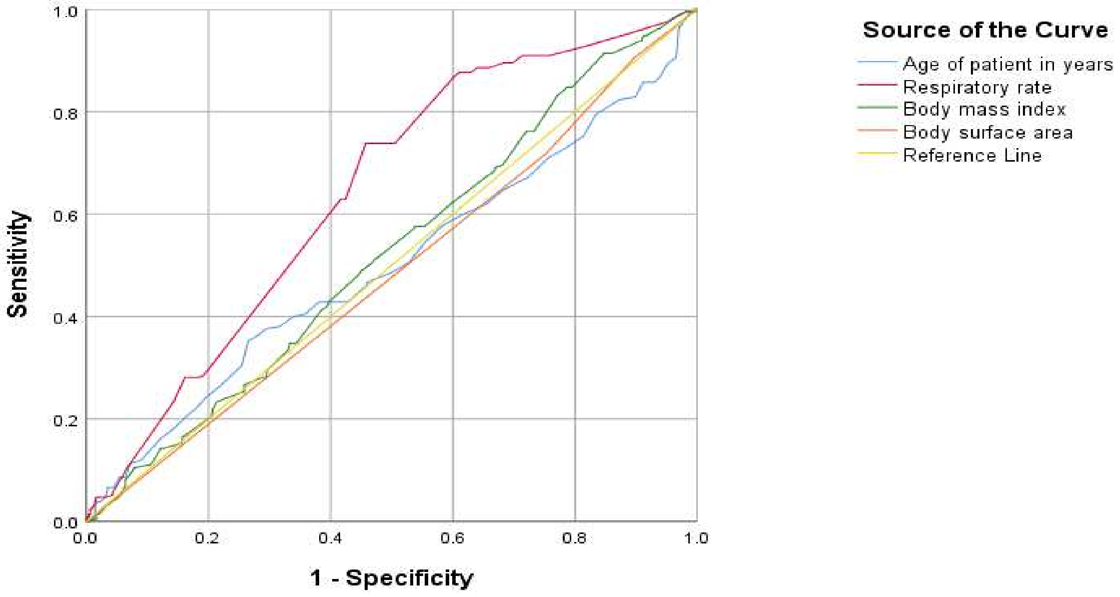

Figure 3: ROC curves for variables used in classifiers

Table 6: Areas under the ROC curves: Null the hypothesis 'true area = 0.5'

Variable used in classifiers	Area	S.E	P-value	95% C.I	
				Lower	Upper
Age of patient	0.500	0.025	0.984	0.450	0.549
Respiratory rate	0.642	0.022	0.000	0.600	0.685
Body Mass Index	0.523	0.023	0.330	0.477	0.569
Body surface area	0.485	0.024	0.532	0.438	0.532

S.E; Standard Error, C.I; Confidence Interval

The areas under the ROC reveals that all variables under the study, except 'respiratory rate' have no significant (p-values > 0.05) area under the curve estimates, implying that these variables cannot effectively discriminate the patient with recurrence of breast cancer from those without recurrence breast cancer for the observations under the study. Also, the 95% confidence interval for these two variables contains 0.5. It can be noted that the maximum area is about 64%, implies that, respiratory rate of a patient has about 64% chances to correctly discriminate a breast cancer patient with recurrence from non-recurrence events.

Discussion

The purpose of this paper was to compare several methods of imputation in replacing missing data values in real breast cancer dataset and to classify observations based on the plausible imputation method. The research found that, among five popular methods of imputation, the predictive mean matching method provided the least values of mean square errors and mean absolute errors. These findings imply that when numerical missing data points exit in a dataset, a PMM imputation technique can be used to replace them more efficiently compared to other methods like series mean, hot deck, k-nearest neighbour, and multiple imputations via both MCMC algorithm and Amelia II package for handling missing data values. This result is in line with [19] which found that PMM techniques more plausible for imputing missing data and it performed well then imputations based on random effects. It has been reported that PMM method diminishes the bias of variance estimate (24). In other study, the PMM yield regression parameters that are significant and just a loss of relative efficiency for about 1% (18).

The binary classifiers; logistic regression and linear discriminant analysis provided very similar results in terms of classification rates, sensitivity and specificity. This indicates that both logistic regression and linear discrimination classifiers are provide similar classification results, and thus any of the two procedures can applied to predict group membership of new breast cancer patients efficiently.

Through comparison of classification results from imputed data set in Table 6 and Figure 3 and un-imputed data (Appendix; Table 8, Figure 4:), it can be noticed that the overall classification rate has raised from 58% and 57% to 68% and 69% imputed data for binary logistic regression and linear discriminant analysis, respectively of original cases that were correctly classified. This is probably caused by a significant reduction of sample size from analysis done on an un-imputed dataset with only 245 cases (without missing observations) of 693(original sample size). Thus, about 65% (448) patients have at least one missing observation. This implies that that the imputation process improves prediction and estimation of missing data values in numerical variables.

Conclusions

The study conclusions are briefly summarised as follows: First, the predictive mean matching is a plausible method of imputing missing data values of numerical variables in clinical or/and breast cancer dataset in general. Secondly, the binary logistic regression and linear discriminant classifiers provide similar prediction (of group membership for breast cancer recurrence) accuracy. Lastly, analysing incomplete datasets through imputation phase is superior than using a case-complete approach towards prediction and estimation. Successful imputation process helps to avoid excessive biased prediction, classification and reduction of sample size.

List of abbreviations

BC**:** Breast cancer; BSA: Body Surface Area; BMI: Body Mass Index; MAR: Missing At Random; MI: Multiple Imputations; MCAR: Missing Completely At Random; NMAR: Not Missing At Random; MCMC: Markov Chain Monte Carlo; ROC: Receiver Operating Characteristics; KNN: K-Nearest Neighbour; VIM: Visualization and Imputation of Missing Values; S.E: Standard Error; SPSS: Statistical Package for Social Sciences. R: R statistical software.

Declarations

Ethics approval and consent to participate

University of Dar es Salaam Research Ethics Committee (UDSM-REC) issued the ethical approval for the study. The need for informed consent was waived by the Institutional review board of Muhimbili National hospital and Ocean Road Cancer Institute. All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication

Not applicable

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding: The costs of research were funded by Ministry of Education, Science and Technology, Tanzania.

Authors' contributions

RAA, ASM, RRJA: study concept and design. RA: data collection, analysis and manuscript development. ASM and RRJA: active manuscript revision and editing. All authors read and approved the final manuscript.

Acknowledgements

This article is a part of the requirement for the PhD program at the University of Dar es Salaam.

Authors' information

[1]Department of Natural Science, State University of Zanzibar, Zanzibar-Tanzania

[2]Department of Statistics, University of Dar es Salaam, Dar es Salaam-Tanzania

* Corresponding author: rahibuabassi@yahoo.com

References

1. Humphries M. Missing Data & How to Deal: An overview of missing data. Popul Res Cent [Internet]. 2013;45. Available from: http://www.texaslonghornsl.com/cola/centers/prc/_files/cs/Missing-Data.pdf

2. Iren M, Tokle R. Comparison of Missing Data Imputation Methods for Improving Detection of Obstructive Sleep Apnea. 2017;

3. Nekouie A, Moattar MH. Missing value imputation for breast cancer diagnosis data using tensor factorization improved by enhanced reduced adaptive particle swarm optimization. J King Saud Univ - Comput Inf Sci [Internet]. 2019;31(3):287–94. Available from: https://doi.org/10.1016/j.jksuci.2018.01.006

4. Morse EP, Maegga B, Joseph, Joseph G, Miesfeldt S. Breast Cancer : Basic and Clinical Research. 2014;73–9.

5. Mbonde et al. NIH Public Access. 2010;31(1):33–41.

6. Burson et al. NIH Public Access. Bone [Internet]. 2014;23(1):1–7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf

7. Rambau P, Masalu N, Jackson K, Chalya P, Serra P, Bravaccini S. Triple negative breast cancer in a poor resource setting in North-Western Tanzania : a preliminary study of 52 patients. 2014;1–6.

8. Tanzania Breast Health Care. Tanzania Breast Health Care Assessment 2017: An assessment of breast cancer early detection, diagnosis and treatment in Tanzania. Available https://ww5.komen.org/breastcancertanzania [Internet]. 2017;1–62. Available from: https://ww5.komen.org/uploadedFiles/_Komen/Content/Grants_Central/International_Grants/Grantee_Resources/Full_Tanzania_Assessment_report.pdf

9. Pazhoohesh M, Walker S, Pourmirza Z. A comparison of Methods for Missing data treatment

in building sensor data. In 2019.

10. Little and Rubin. Statistical Analysis with Missing Data. John Willey & Sons; 1987.

11. Dong Y, Peng CJ. Principled missing data methods for researchers. 2013;2004:1–17.

12. Curley C, Krause RM, Feiock R, Hawkins C V. Dealing with Missing Data : A Comparative Exploration of Approaches Using the Integrated City Sustainability Database. 2019;

13. Jerez, Molina I, Garcı PJ, Alba E, Ribelles N, Franco L, et al. Artificial Intelligence in Medicine Missing data imputation using statistical and machine learning methods in a real breast cancer problem. 2010;50:105–15.

14. Song WJ, Kim K Il, Park SH, Kwon MS, Lee TH, Park HK, et al. Breast Cancer The Risk Factors Influencing between the Early and Late Recurrence in Systemic Recurrent Breast Cancer. 2012;15(2):218–23.

15. Andridge RR, Little RJ. A review of hot deck imputation for survey non-response. Int Stat Rev. 2011;78(1):40–64.

16. Kowarik A, Templ M. Imputation with the R Package VIM. 2016;74(7).

17. Molenburghs & Verbeke. Models for Discrete Longitudinal Data. 8Bickel P, Diggle P, Fienberg S, Gather U, Olkin I, Zeger S, editors. Springer Series in Statistics; 2005.

18. Siswantining T, Soemartojo SM, Sarwinda D. Multiple Imputation with Predictive Mean Matching Method for Numerical Missing Data. In 2019.

19. Bailey BE, Andridge R, Shoben AB. Multiple imputation by predictive mean matching in cluster-randomized trials. BMC Med Res Methodol. 2020;20(1):1–16.

20. Horton NJ, Lipsitz SR. Multiple imputation in practice : Comparison of software packages for regress ... Sci York. 2001;55(3):244–54.

21. Buuren S Van, Groothuis-oudshoorn K. mice : Multivariate Imputation by Chained. 2014;(December 2011).

22. Beretta L, Santaniello A. Nearest neighbor imputation algorithms : a critical evaluation. BMC Med Inform Decis Mak [Internet]. 2016;16(Suppl 3). Available from: http://dx.doi.org/10.1186/s12911-016-0318-z

23. Zhu X. Comparison of Four Methods for Handing Missing Data in Longitudinal Data Analysis through a Simulation Study. 2014;(December):933–44.

24. Gaffert P, Meinfelder F, Bosch V. Towards an MI-proper Predictive Mean Matching. 2016;

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx