AUCTORES
Globalize your Research

Journal of Women Health Care and Issues
Rahibu A. Abassi *

Open Access

Research Article

Imputation Methods on Retrospective Breast Cancer Data in Tanzania: A Comparative Study

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:** Rahibu A Abassi. Department of Natural Science, State University of Zanzibar, Zanzibar-Tanzania.

Received date: April 18, 2022; Accepted date: June 05, 2022; Published date: June 21, 2022.

Citation: Rahibu A. Abassi, Amina S. Msengwa, Rocky R. J. Akarro. (2022). Imputation Methods on Retrospective Breast Cancer Data in Tanzania: A Comparative Study. J. *Women Health Care and Issues.* 5(4); DOI:10.31579/2642-9756/118

Abstract

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

Methodology: Five imputation methods namely series mean, k-nearest neighbour, hot deck, predictive mean matching, expected maximisation via bootstrapping, and multiple imputation by chained equations 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 data values in a real breast cancer dataset under the study. It is a more effective and good approach to handle missing data. We recommend replacing missing data by using predictive mean matching since it is a plausible approach toward multiple imputations for numerical variables. 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

Introduction

Cancer can be described as a disease that occurs when abnormal cells of a certain part of the body divide in uncontrollable trend. Globally, over 19 million new cases and around 10 million deaths happened due to cancer in 2020. Breast cancer is a disease caused by uncontrollable growth of breast cells. Among the types of cancer, female's breast cancer is the most common diagnosed with an about 2.3(11.7%) million new cases, and then lung cancer with an approximate (11.4%), followed by colorectal cancer (10%), prostate cancer (7.3%) and stomach cancer with 5.6% [1]. Africa had the biggest mortality rate of breast cancer globally whilst sub-Sahara Africa records the largest incidence rate [2]. Like in other countries, Tanzania's breast cancer database is facing some challenges including the presence of missing data for some patients' records.

The issue of missing data is common scenario in most clinical studies. In these studies, missing values are not avoidable and impose great challenge, resulting to inaccurate statistical inference due to biasedness. While some investigators consider missing data a minor problem, in fact ignoring them may substantially bias estimates [3].

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 (too small or too large standard errors) 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].

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 efficiency of several imputation methods on a real retrospective breast cancer dataset 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 was evaluated by the 'Root Mean Squared Error' and 'Mean Absolute Error' [9]. The imputation techniques for treating datasets in this article are: (series) mean, hot deck, k-nearest neighbour, predictive mean matching, expectation-maximisation via bootstrapping, and multiple imputations by chained equations. The article compares the efficiency of these imputation methods in replacing numerical missing data values and classifying breast cancer cases to either 'recurrence' or 'non-recurrence' from real breast cancer data set.

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: $Y_1, Y_2, \dots, Y_k$. 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 $Y_{i+1}, Y_{i+2}, \dots, Y_P$ are also missed. An arbitrary pattern arises 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].

Materials and 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.

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%.

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

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 (like age of patient, and BMI among others) 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. These variables were chosen as we focus only on imputing numerical covariates.

Methods Of Imputation:

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.

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)

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" [8] 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).

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 [9] and [10]. 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 [12] obtains the posterior distribution of θ by (iteratively) sampling from conditional distribution $P(Y_1|Y_{-1},\theta_1),\dots,P(Y_q|Y_{-q},\theta_q)$. 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.

K-Nearest Neighbour (KNN):

A non-parametric approach used to impute missing data by averaging its neighbouring observed data [13].

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 neighbors 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.

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 [16].

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, of the two common classifiers will yield similar or different results. in the breast cancer dataset.

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

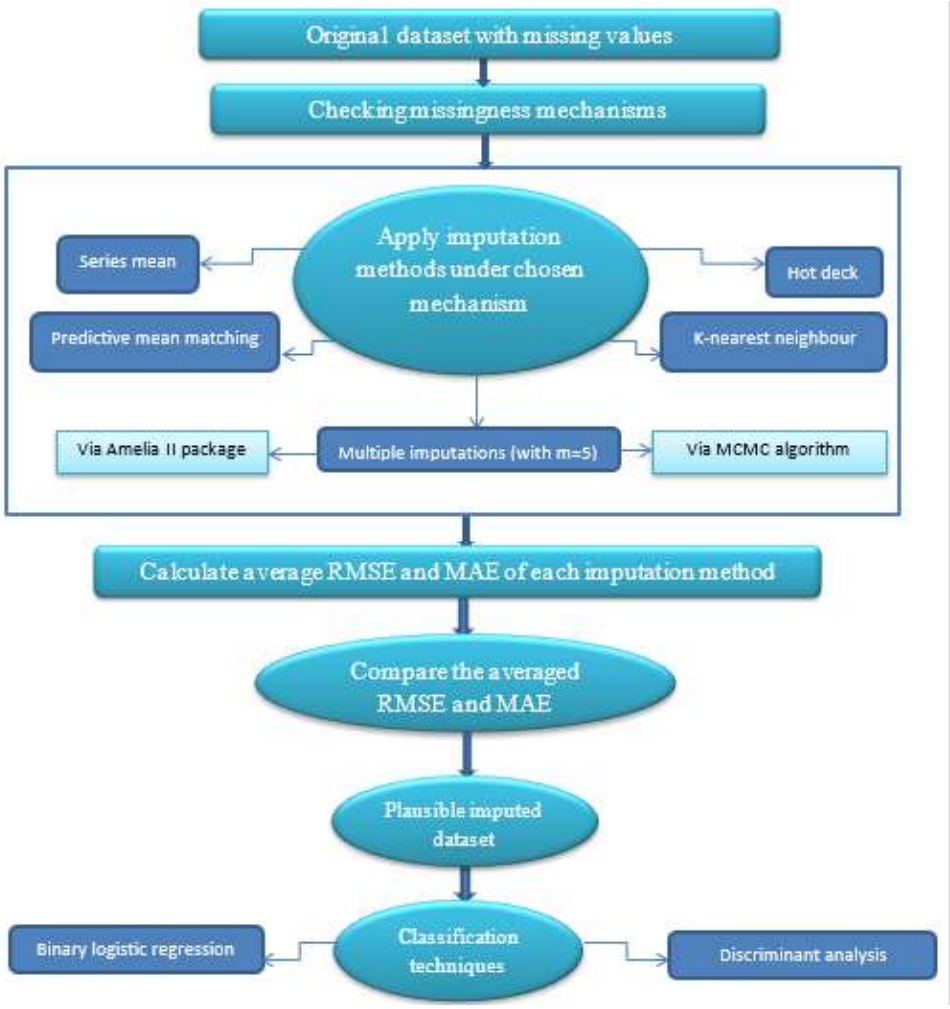

Figure 1: Process of imputation analysis, modified from [9]

Result:

Through exploratory data analysis (Table 1), displaying 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.

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

Table 1: Numerical descriptive measures

Missing Data Patterns and Mechanisms

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 (Table 2) were conducted to gain insight about missing data mechanisms. Two-sample independent t-tests (Table 2) for approximately normally distributed variables 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.

Variables tested	Variance assumptions	Levene's Test for Equality of Variances		Two-samples t - test for Equality of Means	
		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: Two-sample Independent T-test Between Numerical Independent Variables

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.

The findings from two samples t-test with p-values > 0.05 (under hypothesis of 'no significant difference' between the missing and non-missing) are in line with Little's MCAR test (Chi-Square = 129.973, p-value < 0.001) under the null hypothesis that 'data is Missing Completely at Random (MCAR)'. The test result signifies 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.

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

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

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.

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

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

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.

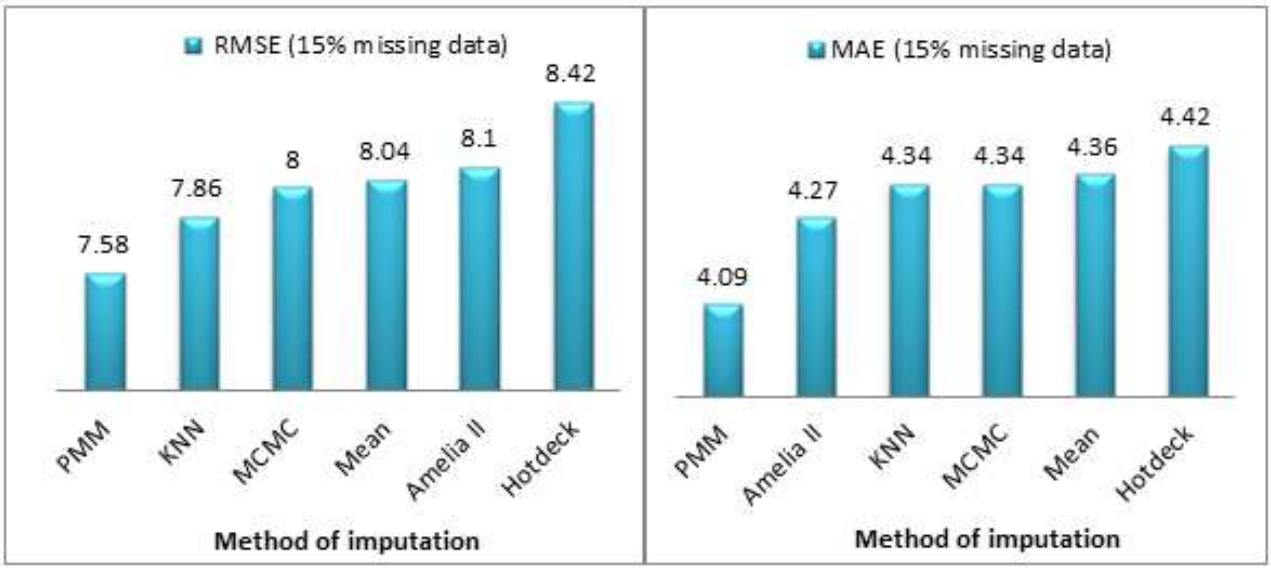

Figure 2: Averaged RMSE and MAE values for different imputation methods

The data resulted from PMM imputation was then used in the classification of observations. The results from classification algorithms (Table 5) for binary logistic regression and linear discriminant analyses. With cut value for of 0.5 for classification of breast cancer recurrence. The two classifiers provide almost similar results in terms of classification accuracy (69.4% and 68.7% respectively for logistic regression and linear discriminant analysis). The ROC curves in Figure 3 and area under the curves (Table 6) shows that the predictor 'respiratory rates' provides best discrimination or reparation of recurrence from non-recurrence cases amongst all predictors in the binary logistic regression.

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

Table 5: Classification tables from imputed dataset

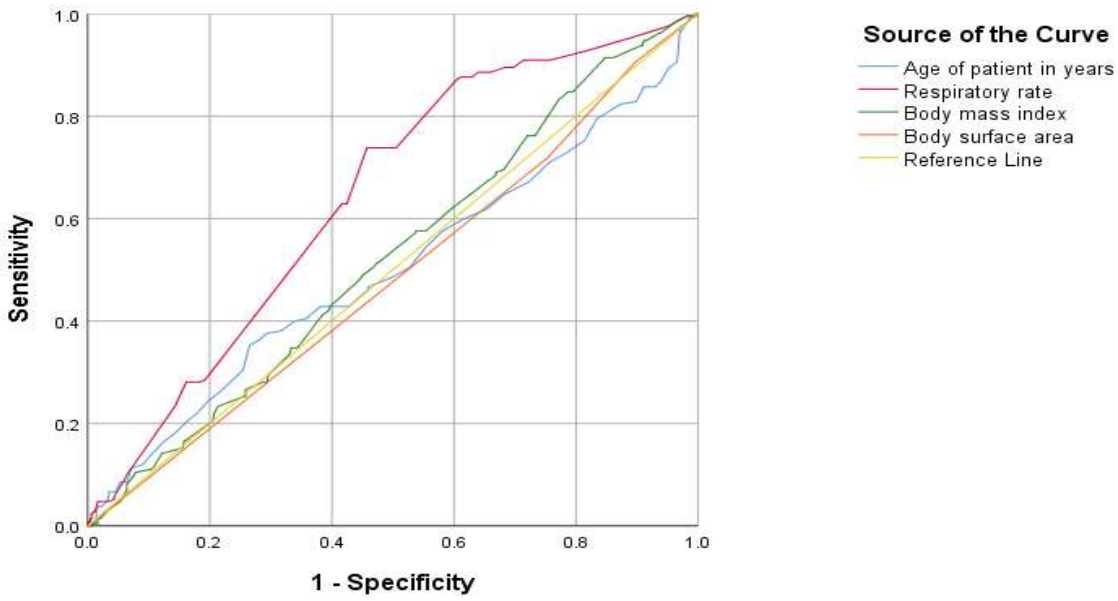

Figure 3: ROC curves for variables used in binary logistic regression classifier

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

Table 6: Areas under the ROC curves: Null the hypothesis 'true area = 0.5'according to binary logistic regression

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 [17]. In other study, the PMM yield regression parameters that are significant and just a loss of relative efficiency for about 1% [9].

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.

References

1. H. Sung et al., (2021,) "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," vol. 71, no. 3, pp. 209–249,
2. S. O. Azubuike, C. Muirhead, L. Hayes, and R. McNally, (2018) "Rising global burden of breast cancer: The case of sub-Saharan Africa (with emphasis on Nigeria) and implications for regional development: A review," World J. Surg. Oncol., vol. 16, no. 1, pp. 1–14,.
3. Nekouie and M. H. Moattar, (2018) "Missing Value Imputation for Breast Cancer Diagnosis Data Using Tensor Factorization Improved by Enhanced Reduced Adaptive Particle Swarm Optimization Atefeh Nekouie Cancer refers to a disease in which a group of cells show uncontrolled growth, invasion ," J. King Saud Univ. - Comput. Inf. Sci.,
4. M. Humphries, (2013) "Missing Data & How to Deal: An overview of missing data," Popul. Res. Cent., p. 45, [Online]. Available:
5. C. Curley, R. M. Krause, R. Feiock, and C. V Hawkins, (2019) "Dealing with Missing Data: A Comparative Exploration of Approaches Using the Integrated City Sustainability Database.
6. Molenburghs & Verbeke, (2005) Models for Discrete Longitudinal Data. Springer Series in Statistics.
7. Little and Rubin, (1987) Statistical Analysis with Missing Data. John Willey & Sons.
8. J. Honaker, G. King, and M. Blackwell, "Amelia II : A Program for Missing Data," vol. 45, no. 7,
9. T. Siswantining, S. M. Soemartojo, and D. Sarwinda, ,(2019) "Multiple Imputation with Predictive Mean Matching Method for Numerical Missing Data.
10. B. E. Bailey, R. Andridge, and A. B. Shoben, (2020) "Multiple imputation by predictive mean matching in cluster-randomized trials," BMC Med. Res. Methodol., vol. 20, no. 1, pp. 1–16.
11. N. J. Horton and S. R. Lipsitz, (2001) "Multiple imputation in practice : Comparison of software packages for regress ...," Sci. York, vol. 55, no. 3, pp. 244–254.
12. S. Van Buuren and K. Groothuis-oudshoorn, (2014) "mice: Multivariate Imputation by Chained,".
13. M. Pazhoohesh, S. Walker, and Z. Pourmirza, (2019) "A comparison of Methods for Missing data treatment in building sensor data.
14. L. Beretta and A. Santaniello, (2016) "Nearest neighbor imputation algorithms : a critical evaluation," BMC Med. Inform. Decis. Mak., vol. 16, no. Suppl 3.
15. Kowarik and M. Templ, (2016) "Imputation with the R Package VIM," vol. 74, no. 7.
16. X. Zhu, (2014) "Comparison of Four Methods for Handing Missing Data in Longitudinal Data Analysis through a Simulation Study," no. December, pp. 933–944.
17. P. Gaffert, F. Meinfelder, and V. Bosch, (2016) "Towards an MI-proper Predictive Mean Matching.

DOI: 10.31579/2642-9756/118