Available online at www.isas.org.in/jisas
Journal of the Indian Society of Agricultural Statistics 71(2) 2017 147–153

Comparative Performance of Imputation Methods for Different Proportions of Missing Data in Classification of Crop Genotypes

Samarendra Das, Amrit Kumar Paul, S.D. Wahi and U.K. Pradhan
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Received 10 October 2016; Revised 15 April 2017; Accepted 18 April 2017

SUMMARY

Most crop datasets contain missing values, a fact which can cause severe problems in the analysis and limit the utility of resulting inference. Classification techniques for grouping of crop genotypes are used when the data is complete. However, the presence of missing values limits the utility of these techniques and creates bias in the resulting inferences. In majority of the cases, missing values are handled by deleting the genotype or traits which contain missing values there by losing information on these genotypes. An interesting approach to handle this problem is to impute the missing values. In this paper, we provided some solutions to handle missing data in crop breeding experiments for classification of crop genotypes. The performance of the imputation techniques is assessed by using the hit ratio criteria computed through four different classifiers by using extensive simulation procedure. This paper has also attempted to provide a description of missing data mechanism in agricultural experiments and various imputation techniques for missing data analysis in classification problems. For lower proportions of missing data, all four of the imputation techniques provided satisfactory results for classification of crop genotypes. For moderate level of missingness in the data, regression and multiple imputation techniques provided same levels of precision for classification of crop genotypes. When there is a high proportion of missing data, multiple imputation technique outperformed all imputation techniques for classification of crop genotypes. Among the classifiers, *k*-th nearest neighbor is the best classification technique in missing data situations.

Keywords: Missing values, Genotypes, Classification, Mean imputation, Regression imputation, Multiple imputation, Hit ratio.

1. INTRODUCTION

Missing data are major concern during conduct of the agricultural experiments (Jo *et al.* 2010). Dropout from the study before the specified completion time has been one of the major reason of the missingness in data (Mishra and Khare 2014). The missingness in the data leads to the problem of incomplete data, which limits the statistical analysis like classification, modelling, *etc.* to a greater extent (Mishra and Khare 2014). Incomplete data is of two types-missing units and missing items. Missing units are the result of non-response for a genotype due to unfavourable conditions (Das 2011). Missing items refer to those units which have missing values for some of the measurements on the genotype (Das 2011). For example, the crop genotype does not respond well due to drought or disease-pest attack, hence some of the measurements on the genotype are missing. Missing values are common when working with large agricultural dataset. Mostly missing data are not handled properly during final analysis which considerably bring biasness in the results, subsequently reduce the power of the study and lead to misleading conclusions (Mishra and Khare 2014). Even the fairest statistical analysis of a study may not be helpful if missingness is related to the outcome measure and leads to unfair results (Mallinckrodt *et al.* 2003). One common strategy to handle missing data during analyses is to include only the complete observations, *i.e.*, the genotypes whose complete data are available on variable/trait of analysis. However, interpretation from such analysis may not be satisfactory. An interesting approach to handle such type of problems is to impute the missing values. Such approaches are quite useful to deal with data sets with missing values and the resulting each completed data set is analyzed by usual statistical techniques.

Corresponding author: Samarendra Das
E-mail address: samarendra4849@gmail.com

During the last few decades, researchers have applied several methods for imputing missing values, including various ad hoc methods as well as advanced model-based approaches. One of the most primitive techniques is to fill in the missing value with the mean of non-missing values. In literature, limited procedures are available for classification in situations involving missing values. Till 1970's, missing data has been considered as a hurdle and were normally deleted, leading to loss of information on certain genotypes. There are different ways of filling these missing data sets, like expectation maximisation algorithm (Dempster *et al*. 1977) and multiple imputation techniques (Little and Rubin 2002). Troyanskaya *et al*. (2001) showed that *k*-nearest neighbors impute method for imputing missing values is more robust than the singular Value Decomposition (SVD) impute method and KNN method performs better than the commonly used row average method (Troyanskaya *et al*. 2001). Bo *et al*. (2004) concluded that least squares imputing method produce the estimates that are consistently more accurate than those obtained by using KNN impute method and are as least accurate as EM impute algorithm (Bo *et al*. 2004). Mishra and Khare (2014) used multiple imputation technique to deal with various proportions of missing data in longitudinal clinical trials by using simulation techniques (Mishra and Khare 2014). Further, Das *et al*. (2015) compared the performance of various classification techniques under multivariate normal and skewed normal set up through computer simulation (Das *et al*. 2015). There is a limited systematic study available in literature to compare various imputation techniques against different proportions of missing data for the classification of crop genotypes in case of crop breeding experiments.

In this paper, we compared four methods to treat missing values in classification of crop genotypes. We chose the simulation procedure based on random deletion technique to create various proportions of missing data and they are imputed by using zero, mean, regression and multiple imputation techniques. The criterion to compare the performance of imputation techniques is the misclassification rate computed through four different classifiers, *viz*. linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), *k-th* nearest neighbor (KNN) and oblique axes method (OAM).

2. MATERIALS AND METHODS

The secondary data on 131 genotypes of ricebean (*Vigna umbellate L*.) grown at Bhubaneswar, Odisha are used for this study. The data are available in the Annual Report for the year 2007-08 of All India Coordinated Research Network on Underutilized Crops, NBPGR, New Delhi. The dataset consists of 9 morphological quantitative characters such as days to 50% flowering, days to maturity, plant height (cm), pods per plant, pod length (cm), seeds per pod, 1000 seed wt. (gm), seed yield/plant (gm), plot yield (gm). Here, we considered the ricebean genotypes with nine morphological characters comprising three groups. First group consists of 52 genotypes, second group consists of 38 genotypes and third group consists of 41 genotypes.

2.1 Creation of Missing Data

Initially, the raw data on 131 genotypes with 9 morphological characters are considered to create missing different datasets, which is subsequently used in this study. For this purpose, we used a simulation procedure based on random deletion technique to create various proportions of missing values in the original data. The simulation procedure can be described as: to create data with α% of missing values, α% observations is deleted randomly through random number generation between one and total sample size from the data set. This constitutes with single sample with α% missing data. This procedure is repeated 500 times to generate 500 random samples with α% missing observations. Here, we took α =1, 5, 10 and 20 to create 1%, 5%, 10% and 20% missing data sets (incomplete data sets). Further, for this purpose a program was written in SAS/IML module to create various proportions of missing data.

2.2 Single Imputation Techniques

Depending on the nature of missing observations in the data, they are imputed by using different procedures. Mean imputation is quite simple and popularly used methods for treatment of missing values. This method simply replaces the missing values with mean value of the non-missing variables. It is only useful for column or attributes imputation but not for the row or case imputation. Another simple technique, *i.e.* zero imputation is quite simple and commonly used methods for treatment of missing

values. This method simply replaces the missing values with zero for a particular variable.

2.3 Mean Imputation

Let x_{ij} be the phenotypic value of *i-th* genotype for *j-th* trait, of the *k-th* group. The x_{ij} value is missing for the *k-th* group and is denoted as C_k. Through the mean imputation this missing value can be imputed as

$$\hat{x}_{ij} = \sum_{x_{ij} \in C_k} \frac{x_{ij}}{n_k}$$

where, n_k represents the number of non-missing values in the *j-th* feature of the *k-th* class.

2.4 Regression Imputation

In the regression method, a regression model is fitted for each variable with missing values. Based on the resulting model, a new regression model is then drawn and is used to impute the missing values for the variable (Rubin 1987). That is, for a variable Y_j with missing values, a model of the form

$$Y_j = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

It is fitted using observations with observed values for the variable Y_j and its covariates $X_1, X_2, ..., X_k$. This regression equation is then used to impute (predict) the values for the missing responses.

2.5 Multiple Imputation

Multiple imputation (MI) was originally proposed by Rubin as a three-step process (Rubin 1996, Little and Rubin 2002, Rubin 2006). First, a set of plausible values is estimated for each missing value, which reflects uncertainty about the non-response model. By filling the missing values with these imputations, complete data sets are created. Second, each complete data set can be analyzed using standard statistical analyses. Finally, the results are combined such that the uncertainties of imputations have been taken into account (Horton and Lipsitz 2001). Two major assumptions are made regarding the data. First, it is assumed that the missingness is missing at random (MAR), that is, the probability that an observation is missing may depend on the observed values but not the missing values. Second, multivariate normality is assumed for the data. A Markov Chain Monte Carlo method is used to impute the missing values. Mean vector and the covariance matrix for the data that do not have missing values are computed as starting values. These estimates are considered as the prior distribution. Filling missing values with the random numbers, which are drawn from the available distribution, creates a complete data set. The mean vector and covariance matrix are recomputed for the complete data set. This is the posterior distribution. Then the missing values are imputed again by generating random numbers from the posterior distribution. This procedure is iterated until the mean vector and covariance matrix are unchanging as we iterate. Imputations from the final iteration are taken to form a data set with no missing values.

2.6 Classification Techniques

LDA is one of the most popular methods of supervised classification. This procedure can be conceptualized as a nonparametric method (*i.e.* distributional assumptions are not explicitly made) because it maximizes between group variability relative to within-group variability (Pohar *et al*. 2004, Erimafa *et al*. 2009, Rausch and Kelly 2009, Glele 2010). However, it can also be conceptualized as a parametric procedure for classification. In particular, LDA is optimal (*i.e*., it maximizes classification accuracy) under the assumptions that the within-group predictors follow multivariate normal distributions and that the population covariance matrices are equal across groups (Raman *et al*. 2015).

Quadratic Discriminant Analysis (QDA) is closely related to LDA and commonly used techniques for multi-group classification. Unlike LDA however, in QDA there is no assumption that the covariance matrices of each groups are identical (Johnson and Wichern 2007).

The non-parametric (or distribution free) method, *i.e.* KNN (Kiang 2003, Wu *et al*. 2010) is used for classifying observations into multiple groups based on a set of quantitative variables. It relaxes the normality assumption and does not require a functional form as required in LDA and QDA. The distance, *d(x, y)*, between any two observations is usually defined by Mahalanobis distance between *x* and *y*. Using the nearest neighbor rule, an observation is classified to one of the groups to which a majority of its *k*-th nearest neighbors belong. The sample distribution approximation is accomplished by dividing the variable space in to arbitrary number of decision regions.

The OAM is considered to be a non-parametric (or distribution free) method as it does not assume the distributional form of the population. The method classifies the observations into one of the several groups based on the square of distances between points corresponding to observation vectors using the oblique co-ordinate system (Das 1998). Some weight factors are associated with the distances are known as compounding values (Rao 1946). These weights can be calculated by maximizing the ratio between average squared distances of all possible pairs of the group mean vectors to the pooled average squared distances within groups. Further, the performance of OAM along with LDA, QDA and KNN for classification of ricebean genotypes was studied under multivariate normal and skewed normal set up (Das *et al*., 2015).

2.7 Criteria for Performance Analysis of Imputation Techniques

Criteria called hit ratio (HR) (Erimafa *et al*. 2009) was used to compare the performance of imputation techniques based on the effect of misclassification rate of four classifiers: LDA, QDA, KNN and OAM. The performance of LDA, QDA, KNN and OAM by different imputation techniques under different levels of missing observations (1%, 5%, 10% and 20%) are assessed on the basis of HR. The HR criteria could be calculated from the confusion matrix, which shows actual *vs.* predicted group membership and is given below.

Table 1. Schematic representation of confusion matrix

	Π_1	Π_2	. . .	Π_{g-1}	Π_g	total
Π_1	N_{11}	N_{12}	. . .	N_{1g-1}	N_{1g}	N_1
Π_2	N_{21}	N_{22}	. . .	N_{2g-1}	N_{2g}	N_2
. . .						
Π_g	N_{g1}	N_{g2}	. . .	$N_{g,g-1}$	$N_{g,g}$	N_g

Where, Π_i are groups for $i=1, 2, \ldots, g$

N_{ii} = number of Π_i items correctly classified as Π_i items

N_{ij} = number of Π_i items misclassified as Π_j items

$$N = N_1 + N_2 + \ldots + N_g$$

The criteria HR can be calculated as

$$\text{HR} = \frac{N_{11} + N_{22} + \ldots + N_{gg}}{N}$$

The classification method which has high hit ratio is considered as best method and *vice-versa.*

3. THE ALGORITHM

Step 1: Generate various proportions of missing values in original data (*i.e.* incomplete datasets).

Step 2: Impute the missing values with a suitable imputation technique.

Step 3: Classification of genotypes using the imputed datasets using a proper classifier.

Step 4: Computation of Hit Ratio using Table 1.

Step 5: Assess the performance of imputation and classification techniques based on the computed Hit Ratio.

4. RESULTS AND DISCUSSION

The simulated (incomplete) datasets with randomly created missing values, to the extent of 1%, 5%, 10% and 20% in the original datasets, were imputed by using four different methods namely, zero imputation, mean imputation, regression imputation and multiple imputation. These imputed datasets were further used as inputs for classification purpose. The Tables 2-5 summarised the results obtained from 500 simulations in imputed datasets (for missing datasets) as well as original dataset (without missing values). The findings of the imputed techniques are interpreted in light of classification accuracy of four different classifiers. For the application of OAM classification method, a SAS/IML code was developed based on Das's algorithm (Das, 1998). Further, for the application of LDA, QDA and KNN method, we used the SAS (9.2) and SPSS (16.0) program. The accuracy of the classifiers was inferred by using HR criteria for different proportions of missing data as well as original data and given in Tables 2-5.

Table 2. HR of classifiers under different proportions of missing data for mean imputation.

Methods	0%	1%	5%	10%	20%
LDA	0.8396	0.8285	0.8171	0.7508	0.7106
QDA	0.6783	0.6612	0.6429	0.6078	0.5756
KNN	0.9313	0.9189	0.9121	0.8965	0.8645
OAM	0.8779	0.8479	0.8268	0.7408	0.7067

Methods, represents classification techniques; 0%, represents original data; 1, 5, 10 and 20%, represents missing data proportions

Table 3. HR of classifiers under different proportions of missing data for zero imputation.

Methods	0%	1%	5%	10%	20%
LDA	0.8396	0.8226	0.7977	0.6559	0.5255
QDA	0.6783	0.6532	0.5977	0.5597	0.4994
KNN	0.9313	0.9054	0.8484	0.8133	0.7673
OAM	0.8779	0.8245	0.7967	0.7389	0.6846

0%, represents original data; 1, 5, 10 and 20% represents missing data proportions in original data

Table 4. HR of classifiers under different proportions of missing data for regression imputation.

Methods	0%	1%	5%	10%	20%
LDA	0.8396	0.8366	0.8064	0.7967	0.7612
QDA	0.6783	0.6639	0.6365	0.6165	0.5938
KNN	0.9313	0.9202	0.9189	0.9034	0.8972
OAM	0.8558	0.8367	0.8304	0.8245	0.8246

0%, represents original data; 1, 5, 10 and 20% represents missing data proportions in original data

Table 5. HR of classifiers under different proportions of missing data for multiple imputation.

Methods	0%	1%	5%	10%	20%
LDA	0.8396	0.8378	0.8147	0.7989	0.7813
QDA	0.6783	0.6687	0.6504	0.6188	0.5978
KNN	0.9313	0.9206	0.9227	0.9168	0.9078
OAM	0.8558	0.8579	0.8374	0.8316	0.8277

0%, represents original data; 1, 5, 10 and 20% represents missing data proportions in original data

From Tables 2 and 3, it was observed that the HR for LDA at 1% of missing observations under mean and zero imputation are 0.8285 and 0.8226 respectively. The result indicated the performance of both these imputation techniques are at par for LDA, when missingness level in data is 1%. The HR for LDA at 1% of missing observations under regression imputation (0.8366) and multiple imputation (0.8378) is nearly same (Tables 4 and 5) and higher than that of mean and zero imputation techniques. Further, for the same classifier the HR obtained for regression and multiple imputation techniques are nearly equal to HR obtained for original data (complete data). This indicated that for lower missing levels in data, both regression and multiple imputation techniques provided satisfactory results as of complete data case. Further, similar interpretations can be made for other classifiers *viz.* QDA, KNN and OAM (Tables 2-5). The HR for different classification techniques decreases as the proportions of missing data increases from 1% to 5% but the rate of decrease in HR is gradual (Tables 1-4). Further, it was observed that the HR of different classification techniques decreases at higher rate when the number of missing observations increases from 5% to 20%. These findings indicated that the missing data has significant effect on performance of classifiers and largely depend on the intensity of missingness in the data. This is true for all the four imputation methods irrespective of classifiers.

For regression and multiple imputation techniques, the performance of LDA, QDA, KNN and OAM in terms of HR is consistent for all the proportions of missing data. The rate of decrease in HR of different classification techniques is higher in case of zero imputation followed by mean, regression and multiple imputation techniques. Further, to get a clear idea about the performance of imputation techniques against all proportions of missing data, we calculated the weighted average of HRs of classifiers for each imputation technique. The assigned weights are relative to the proportions of missingness in the data, *i.e.* higher the proportions of missing data, higher will be the value of weights and *vice versa*. The weighted average of HRs of different classification methods are represented in Table 6.

Table 6. Weighted average HRs of classification methods under different imputation techniques.

Classification methods	Imputation		techniques	
	Mean	Zero	Regression	Multiple
LDA	0.7398	0.6071	0.7794	0.7924
	(0.7768)	(0.7004)	(0.8002)	(0.8082)
QDA	0.5934	0.5341	0.6080	0.6129
	(0.6219)	(0.5775)	(0.6277)	(0.6339)
KNN	0.8815*	0.7915*	0.9026*	0.9127*
	(0.8980)	(0.8336)	(0.9099)	(0.9170)
OAM	0.7368	0.7191	0.8257	0.8310
	(0.7806)	(0.7612)	(0.8291)	(0.8387)

Values in () represent simple averages of Hit Ratios and * for highest value

The results showed that LDA classifier performed well under multiple imputation followed by regression imputation (Table 6). The performance of LDA was not so satisfactory when the missing data are imputed by zero and mean imputation techniques. Similar interpretations can be made for other classifiers like QDA, KNN and OAM. Among the Classifiers, KNN outperformed all other classification methods as it has highest weighted average HRs for each imputation

technique followed by OAM (Table 6). These findings might be attributed due to the fact that KNN does not require any stringent assumptions of the data like normality and equality of dispersion matrices (Kiang 2003). These assumptions of data *i.e.* normality and equality of dispersion matrices are violated by real crop data scenarios (Wahi and Bhatia 2005) as well as the characteristics of missingness in the data. Further, the classifier QDA performed poor among all other classification methods irrespective of imputation techniques followed by LDA, due to the violation of normality assumptions in data due to the presence of missing values.

For a better comparison among the four imputation techniques for various proportions missing data in classification of genotypes with respect to the complete original data scenarios (no missing data), we calculated the Pearson's correlation co-efficient between the results obtained from the various imputation techniques with that of complete data scenarios. The results are given in Table 7.

Table 7. Correlation between the results obtained from original dataset with different imputed data sets.

Methods of Imputation	1 %	5 %	10 %	20 %
Zero	0.965*	0.945*	0.789	0.645
Mean	0.989**	0.986**	0.809*	0.787
Regression	0.988**	0.994**	0.989**	0.907*
Multiple	0.987**	0.994**	0.989**	0.912*

*, ** indicate significance at 5% and 1% levels of significance; 1%, 5%, 10% and 20% represents the missing data proportions; the values represents the Pearson's correlation co-efficient between the results obtained from original data with that of imputed datasets though various imputation technique

Table 7 summarised the correlation among the classification accuracies achieved by each imputation technique for various proportions of missing data with the results from complete data situations through Pearson's correlation analysis. The value of correlation for zero substitution, mean, regression and multiple imputations against 1% and 5% missing data are nearly same (Table 7). This indicated that the results obtained from original (complete) data sets are nearly same with the results when the missing data are imputed by various imputation techniques when the missing data proportions are low, as there is no statistically detectable significance different results obtained from these two cases. In case of moderate (10%) missing data situations, the zero imputation technique performed poor. For mean imputation, the correlation co-efficient is 0.809, significant at 5% level of significance, showed moderately well performance as compared to the results from complete data. The highest value of correlation co-efficient (0.989) indicated that regression and multiple imputation performed equally well in the situation of moderate (10%) proportions of missing data.

The performance of zero and mean imputations deteriorated significantly when the proportions of missing data is higher (20% or higher). The highest value of correlation co-efficient (0.912) for multiple imputation followed by regression imputation technique showed their efficiency to handle missing data, when its proportion is high in the data when classification is concerned. Further, the results also stated the robustness and efficiency of multiple and regression imputations over mean and zero imputations for the classification of crop genotypes when there are high proportions of missing data.

5. CONCLUSION

Missing data invariably occur during conduct of a crop breeding trial and considered as a major concern for classification, regression modelling, *etc*. Unfortunately, the occurrence of missingness is unavoidable despite the carefulness in experimental design, conduct and preventive strategies. Due to complexity of data analysis while dealing with missing data, researchers and plant breeders exclude the subjects or genotypes with partial information (missing values) on response variables. In the present investigation, we explored the efficiency and appropriateness of various imputation methods with varying size of missingness in data for classification of genotypes. The efficiency of imputation techniques were assessed through various classifiers. The study is concerned with comparison of four imputation techniques applied to incomplete ricebean datasets with MAR drop-outs, randomly created using random deletion technique. The results of the study indicated that multiple and regression imputations are the most appropriate method of missing data imputation for classification of crop genotypes. It was also demonstrated that varying proportions of missing data in the study affect the performance of imputation techniques as well classification methods. The findings of the study may

have important implications, particularly for multi group classification of crop genotypes where varying proportions of missing data were encountered. As evident from the findings, single imputation like zero or mean imputation may not be a suitable approach for imputing missing data for classification because as it does not incorporate the uncertainty of missingness in imputed value. For the classification of genotypes in to multi groups, the non-parametric technique like KNN outperformed all parametric approaches of classification. The present study will surely serve as a practical guide for researchers and plant breeders to choose proper imputation and classification methods according to their object of analysis.

ACKNOWLEDGEMENTS

The help obtained from Indian Council of Agricultural Research, New Delhi and ICAR-Indian Agricultural Statistics Research Institute, New Delhi is duly acknowledged.

REFERENCES

Bo, T.H., Dysvik, B. and Jonassen, I. (2004). LS impute: Accurate estimation of missing values microarray data with least square method. *Nucleic Acids Res*., **32(3)**, 23-27.

Das, M.N. (1998). Classification of observations using distances using distances in oblique axes system. *J. Ind. Soc. Agril. Statist.*, **51**, 379-384.

Das, S. (2011). *Some investigations on different classification techniques in agriculture*. M. Sc. Thesis, ICAR-IARI, New Delhi, http://krishikosh.egranth.ac.in/handle/1/87869.

Das, S., Paul, A.K., Wahi, S.D. and Raman, R.K. (2015). A comparative study of various classification techniques in multivariate skew-normal data. *J. Ind. Soc. Agril. Statist.,* **69(3)**, 271-280.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Stat. Soc. Series B Stat. Methodol.,* **39(1)**, 1-38.

Erimafa, J.T., Iduseri, A. and Edokpa, I.W. (2009). Application of discriminant analysis to predict the class of degree for graduating students in a university system. *Int. J. Phys. Sci.,* **4(1)**, 16-21.

Glele, K.R., Pelz, D. and Palm, R. (2010). The efficiency of linear classification rule in multi-group discriminant analysis. *Afr. Jour. Math. Comp. Sci. Res*., **3(1)**, 19-25.

Horton, N.J. and Lipsitz, S.R. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing variables. *J. Amer. Statist. Assoc.*, **55**, 244-254.

Jo, B., Ginexi, E.M. and Ialongo, N.S. (2010). Handling missing data in randomized experiments with noncompliance. *Prevention Sci.,* **11(4)**, 384-396.

Johnson, R.A. and Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis*. 6th ed., Prentice Hall, Inc., New York.

Kiang, M.Y. (2003). A comparative assessment of classification methods. *Decision Support Sys.*, **35**, 441-454.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. 2nd edition. Wiley, Hoboken, NJ, USA.

Mallinckrodt, C.H., Sanger, T.M., Dube, S., DeBrota, D.J., Molenberghs, G., Carroll, R.J., Potter, W.Z. and Tollefson, G.D. (2003). Assessing and interpreting treatment effects in longitudinal clinical trials with missing data. *Biol. Psychiatry,* **53**, 754-760.

Mishra, S. and Khare, D. (2014). On comparative performance of multiple imputation methods for moderate to large proportions of missing data in clinical trials: a simulation study. *J. Med. Statist. Inform.,* **2**, doi: 10.7243/2053-7662-2-9.

Pohar, M., Blas, M. and Turk, S. (2004). Comparison of logistic regression and linear discriminant analysis: A simulation study. *Metodoloskizvezki*, **1**, 143-161.

Raman, R.K., Paul, A.K., Das, S. and Wahi, S.D. (2015). Empirical comparison of the performance of linear discriminant function under multivariate non-normal and normal data. *Inter. J. Agric. Statist. Sci.,* **11(2)**, 403-409.

Rao, C.R. (1946). Tests with discriminant function in multivariate analysis. *Sankhya*, **7**, 407-410.

Rausch, R.J. and Kelly, A. (2009). A comparison of linear and non-linear models for discriminant analysis under non-normality. *Behav. Res. Methods*, **41(1)**, 85-98.

Rubin, D.B. (1987). Estimating causal effects from large data sets using propensity scores. *Ann. Internal Medicine*, **127**, 757-763.

Rubin, D.B. (1996). Multiple imputation after 18 years. *J. Amer. Statist. Assoc*., **91**, 473-489.

Rubin, D.B. (2006). Discussion on multiple imputation. *Inter. Statist. Rev.,* **71**, 619-625.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Has, T., Tibshirani, R., Botstein, D., Altman, R.B. (2001). Missing value estimation for DNA microarrays. *Bioinformatics*, **17(6)**, 520-525.

Wahi, S.D. and Bhatia, V.K. (1995). Use of bootstrap method in comparing the performance of linear discriminant function. *J. Ind. Soc. Agric. Statist.*, **47(1)**, 12-20.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Liu, B., Yu, P.S. and Hand, D.J. (2008). Top 10 algorithims in data mining. *Know. Info. Sys.,* **14**, 1-37.