Pertanika J. Sci. & Technol. 29 (2): 979 - 998 (2021)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Comparison of Single and MICE Imputation Methods for Missing Values: A Simulation Study

Nurul Azifah Mohd Pauzi[1], Yap Bee Wah[1,2]*, Sayang Mohd Deni[1,2], Siti Khatijah Nor Abdul Rahim[3] and Suhartono[4]

[1]Centre of Statistical and Decision Science Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[2]Advanced Analytics Engineering Centre, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[3]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,
40450 Shah Alam, Selangor, Malaysia
[4]Department of Statistics and Data Science, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

ABSTRACT

High quality data is essential in every field of research for valid research findings. The presence of missing data in a dataset is common and occurs for a variety of reasons such as incomplete responses, equipment malfunction and data entry error. Single and multiple data imputation methods have been developed for data imputation of missing values. This study investigated the performance of single imputation using mean and multiple imputation method using Multivariate Imputation by Chained Equations (MICE) via a simulation study. The MCAR which means missing completely at random were generated randomly for ten levels of missing rates (proportion of missing data): 5% to 50% for different sample sizes. Mean Square Error (MSE) was used to evaluate the performance of the imputation methods. Data imputation method depends on data types. Mean imputation is commonly used to impute missing values for continuous variable while MICE method can handle both continuous and categorical variables. The simulation results indicate that group mean imputation (GMI) performed better compared to overall mean imputation (OMI) and MICE with lowest value of MSE for all sample sizes and missing rates. The MSE of OMI, GMI, and MICE increases when missing rate increases. The MICE method has the lowest performance (i.e. highest

ARTICLE INFO

Article history:
Received: 24 September 2020
Accepted: 30 December 2020
Published: 30 April 2021

DOI: https://doi.org/10.47836/pjst.29.2.15

E-mail addresses:
azifahpauzi93@yahoo.com (Nurul Azifah Mohd Pauzi)
beewah@tmsk.uitm.edu.my (Yap Bee Wah)
sayang@tmsk.uitm.edu.my (Sayang Mohd Deni)
sitikhatijahnor@tmsk.uitm.edu.my
(Siti Khatijah Nor Abdul Rahim)
suhartono@statistika.its.ac.id (Suhartono)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

MSE) when percentage of missing rates is more than 15%. Overall, GMI is more superior compared to OMI and MICE for all missing rates and sample size for MCAR mechanism. An application to a real dataset confirmed the findings of the simulation results. The findings of this study can provide knowledge to researchers and practitioners on which imputation method is more suitable when the data involves missing data.

Keywords: MICE, missing data, multiple imputation, simulation, single imputation

INTRODUCTION

High quality data is important to ensure correct information and valid findings for better evidence-based decision-making. One of the key concerns related to data quality is missing data. Missing data can occur for various reasons during the data collection process, such as incomplete responses (Pampaka et al., 2016; Barnett et al., 2017), equipment malfunction (Masconi et al., 2015; Gopal et al., 2019) and manual data entry errors (Bhati & Gupta, 2016). Incomplete data is a serious data quality problem since it leads to a reduction of statistical power, bias in parameter estimates, and loss of efficiency in the analytical process (Kaiser, 2014; Ayilara et al., 2019; Hughes et al., 2019). These problems have led to extensive research on developing methods to treat missing data.

Data imputation is widely used to deal with missing data in many areas such as medical research (Pedersen et al., 2017; Sullivan et al., 2018; Turner et al., 2019; Stavseth et al., 2019), organizational research (Newman, 2003; Fichman & Cummings, 2003; Newman, 2014) and educational research (Grund et al., 2018; Shi et al., 2019). The main objective of data imputation is to replace any missing data with estimated values to obtain a complete dataset. The estimated values could be from the calculated mean, median, mode, predefined value of missing variable or values that are obtained from the predictive model (Malarvizhi &Thanamani, 2012).

One of the important considerations in addressing missing data is identifying the mechanism of missing data. There are three types of missing data mechanisms known as MCAR, missing at random (MAR) or missing not at random (MNAR). The mechanism and pattern of missing data plays an important role during the selection of imputation methods. They have greater impact on research results compared to the proportion of missing data (Tabachnick et al., 2007; Song & Shepperd, 2007). There is a need to address these issues properly to help in selecting appropriate data imputation methods. If a study employs improper imputation methods, it may lead to incorrect data analysis, false conclusions, and erroneous predictions (Chaudhry et al., 2019). Reliable data imputation methods are needed as the outcome of data analysis tasks relies upon efficient handling of missing data.

Several data imputation methods have been developed such as mean imputation, hot deck imputation, regression imputation and Multivariate Imputation by Chained

Equations (MICE). The simplest and commonly used method is by replacing the missing data with the mean of the continuous variable. Various studies have been conducted on the performance of data imputation methods. The study conducted by Le et al. (2018) compared the performance of four data imputation method, Expectation Maximization (EM), Multiple Imputation, K-Nearest Neighbor and Mean Imputation in healthcare dataset. The results showed that imputation using EM outperformed other methods with lowest RMSE value. Jadhav et al. (2019) compared seven imputation methods namely mean imputation, median imputation, k-NN imputation, predictive mean matching, Bayesian Linear Regression (norm), Linear Regression, non-Bayesian (norm.nob). They reported that k-NN imputation method is the more superior imputation method. Recently, Kamatchi and Baranidharan (2019) proposed deep learning imputation method (DNN) and compared it with statistical imputation methods (Mean, imputation with zero or constant, Stochastic Regression imputation, Extrapolation and Interpolation and Hot-Deck imputation, MICE), machine learning methods (K-nearest neighbour) using an Autism dataset. Their results showed that DNN outperformed statistical and machine learning methods. The study by Ochieng’Odhiambo (2020) investigates the three most common conventional methods (Listwise, Mean Imputation, Median Imputation) in handling missing data. Their finding showed that median imputation was a better data imputation method among the conventional methods.

The aim of this study is to evaluate the performance of single and multiple imputation methods for continuous variable via a simulation study. The selected single imputation method is mean imputation while the multiple imputation method is MICE. In this simulation study, MCAR missing data were generated for different sample size and missing rates ranging from 5% to 50%. Then, the performance of these imputation methods was evaluated based on Mean Square Error (MSE). The simulation approach allows the researcher to control the parameter in order to study the pattern and changes in the estimation of each imputation method. The simulation procedures were carried out using R programming language software. The imputation methods were applied to a real dataset taken from the Kaggle website to validate the findings of the simulation study.

MATERIALS AND METHODS

Missing Data Mechanisms

The mechanism of missing data describes the reason the values are missing. Little and Rubin (1987) classified the mechanisms of missing data into two major types known as ignorable and non-ignorable missingness. Non-ignorable means that the missing data mechanism is related to the missing values. It is referred to as non-ignorable since the missing data mechanism itself must be modelled when dealing with missing data. There is a need to include some model for why the data are missing and what the likely values are. Ignorable missingness however ignores any information about the missing data itself when dealing with missing data. Missing Completely at Random (MCAR) and Missing at Random (MAR) are classified under ignorable missingness, whereas Missing Not at Random (MNAR) is non-ignorable missingness (Aljuaid & Sasi, 2016; Chhabra et al., 2017). The mechanism of missing data is determined by the dependency of *K* on the variables in the data set where *K* represents the missing data indicator, $\mathrm{Y}_{\mathrm{obs}}$ is the observed values, and $\mathrm{Y}_{\mathrm{miss}}$ is the missing values.

The probability of missingness is related to observed data but not on the missing data itself for the MAR mechanism (Ma & Chen, 2018; van Ginkel et al., 2019; Li et al., 2019). There is a systematic relationship between missing data and observed data in the dataset. The MAR assumption is defined as Equation 1:

$$Pr(K|Y_{obs}, Y_{miss}) = Pr(K|Y_{obs}) \qquad [1]$$

while the stronger assumption of MCAR can be written as Equation 2:

$$Pr(K|Y_{obs}, Y_{miss}) = Pr(K) \qquad [2]$$

which implies the missing data indicator, *K* is completely unrelated to both observed and missing data (Aljuaid & Sasi, 2016; Abidin et al., 2018; Madley-Dowd et al., 2019). MCAR is a special case of the MAR mechanism and represents the highest level of randomness. Finally, any missing data that does not satisfy Equation 1 and 2 is classified as MNAR where the missing indicators, *K* are related to missing data itself (Dettori et al., 2018; Goretzko et al., 2019). The assumption of the MNAR mechanism can be expressed as Equation 3:

$$Pr(K|Y_{obs}, Y_{miss}) = Pr(K|Y_{miss}) \qquad [3]$$

Data Imputation Methods

Data imputation is a powerful and widely used method in treating missing data. An imputation method preserves the sample size of data by imputing with estimated values without discarding cases with missing data. There are single or multiple imputation methods (Dettori et al., 2018). Single imputation treats missing data by replacing with a single value for all missing data of that variable (Pederson et al., 2017; Papageorgiou, 2018). As the same values are used to replace each missing data, it ignores the uncertainty of the parameter estimates (Salgado et al., 2016; Yadav & Roychoudhury, 2018). Multiple imputations overcome this problem by considering the uncertainty associated with missing data, which was unaccounted for in single imputation (Pedersen et al., 2017). However, single imputation methods are popular among the researchers partly due to their simplicity and availability in many statistical software. The most basic single imputation is arithmetic mean imputation proposed by Wilks (1932). Mean imputation can be categorized into overall mean imputation (OMI) and group mean imputation (GMI). Thus, OMI and GMI are categorized under single imputation while Multivariate Imputation by Chained Equations (MICE) falls under multiple imputation methods.

Single Imputation

Overall Mean Imputation (OMI). In overall mean imputation, the mean of the observed values of the variable is used to impute missing data in the same variable. Thus, the mean of the non-missing values of that particular variable will be calculated and then used to replace each missing value. This can be expressed by the mathematical Equation 4 (Sim et al., 2015):

$$X_i^j = \sum_{k \in I(complete)} \frac{X_k^j}{n \mid I(complete) \mid} \qquad [4]$$

where X_i^j is the ith instance (or case) for the jth variable, while X_k^j is the non-missing value of the jth variable. *I(complete)* is an index set with 1 if X_i^j is non-missing and 0 otherwise, and $n \mid I(complete) \mid$ is the total number of observed values (non-missing cases) in the jth variable.

Group Mean Imputation (GMI). Group mean imputation is also known as similar case imputation. In GMI, the mean of the observed data is calculated based on the underlying group. Thus, the mean of non-missing data of that particular variable will be calculated separately. For instance, mean of group based on male and female and then replace each missing data according to the gender. The group mean can be expressed by the mathematical Equation 5:

$$X_{m,i}^j = \sum_{k \in I(complete)} \frac{X_{m,k}^j}{n \mid I(complete) \mid} \qquad [5]$$

where $X_{m,i}^j$ is the ith instance (or case) for the jth variable of the mth class. *I(complete)* is an index set with 1 if $X_{m,i}^j$ is non-missing and 0 otherwise, and $n \mid I(complete) \mid$ is the total number of observed values (non-missing cases) in the jth variable of the mth class.

Multiple Imputation

Multivariate Imputation by Chained Equations (MICE). Multivariate Imputation by Chained Equations is a particular multiple imputation technique (Van Buuren, 2007). MICE is also known as fully conditional specification (FCS) that is widely used in handling missing data. It is an extension of single imputation that gives much better results since missing data are estimated multiple times (*m* times) which reflects the uncertainty of parameter estimates of the imputed variables (Zhang, 2016). In the MICE procedure, imputation is carried out by conducting series of estimations whereby each variable with missing data is modeled according to its distribution. An introduction to the MICE method is given in the paper by Buuren and Groothuis-Oudshoorn (2010) which provided good practical resources to guide researchers in implementing this method. The MICE imputation method can be described in a three-step procedure (Ratolojanahary et al., 2019; Lo et al., 2019):

(i) Imputation phase
(ii) Analysis phase
(iii) Pooling phase

In the imputation phase, missing data are imputed for $m > 1$ times resulting in multiple imputed datasets (*m* imputed datasets). Then, in the analysis phase, each *m* imputed dataset is analyzed separately to obtain the parameter estimates and standard errors (*m* point estimates and standard errors) by using a standard statistical procedure. In the pooling phase, the *m* point estimates and standard errors for each analysis are then pooled together to get overall estimates and standard errors. The estimated parameter P_i can be donated by $\overset{\square}{P}_i$ from the i^{th} dataset. Then, the pooled estimates of the parameter P can be obtained as Equation 6 (Dong & Peng, 2013):

$$\overline{P} = \frac{1}{m}\sum_{i=1}^{m}\hat{P}_i \qquad [6]$$

while $\overline{C}$ is the within imputation variance (Equation 7)

$$\overline{C} = \frac{1}{m}\sum_{i=1}^{m}\hat{V}_i \qquad [7]$$

which is the average of the estimated variances for i^{th} imputed datasets. The variance estimates from the i^{th} dataset is denoted as $\overset{\square}{V}_i$. The between imputation variance, W is the variability of the imputed values for multiple imputed datasets and can be written as Equation 8:

$$W = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{P}_i - \overline{P})^2 \qquad [8]$$

The variance of the pooled estimate is the weighted sum of two variances (within imputation $(\overline{C})$, between imputation variance (W) and can be written as Equation 9:

$$Var(\overline{P}) = \overline{C} + (1 + \frac{1}{m})W \qquad [9]$$

where $(\frac{1}{m})$ is an adjustment for the randomness associated with a finite number of imputations.

Performance Evaluation Measure

The evaluation of the imputation methods is based on Mean Square Error (MSE). Based on the past studies (Gad & Abdelkhalek, 2017; Abidin et al., 2018; Nwakuya & Nwabueze, 2018), MSE is the most commonly used metric in *evaluate of the performance* of different *imputation algorithms* for continuous variables. The MSE measures the discrepancy between imputed and actual observed values. The best imputation method will be the one with the lowest MSE which indicates that the predicted values are close to the actual values. MSE is calculated as Equation 10:

$$MSE = \frac{\sum_{i=1}^{n}(x_i^{actual} - x_i^{imputed})^2}{n} \qquad [10]$$

where x^{actual} is the actual observed values and $x^{imputed}$ is the imputed values, and n is the number of cases (observations).

METHODS

Study 1: Simulation Study

The simulation study was carried out to provide empirical evidence on the performance of single and MICE imputation methods in handling missing data. The simulations were conducted using R programming language software. This study adopted the simulation approach by Wah et al. (2018) for generating data with a binary Y and continuous independent variables. As most data imputation methods depend heavily on the assumption of normality (Scheffer, 2002), two continuous independent variables (X_1, X_2) were simulated from standard normal distributions. Data were generated for various sample sizes, $n(30, 50, 100, 200, 500, 1000)$ and ten levels of missing rates, 5% to 50% (with 5% increment) under the MCAR mechanism. The MCAR approach is by removing data randomly from a variable. Little's MCAR Test (Little, 1988) was used to test the assumption of MCAR. The null hypothesis (H_0) is missing data are MCAR. If we reject the null hypothesis under

Little's Test of MCAR, the data may then be assumed to be MAR or MNAR. Some packages in R were used to carry out the simulation study. The 'missForest' (Stekhoven & Bühlmann, 2012) package was used to generate missing data MCAR randomly from the simulated dataset and 'mvnmle', and 'BaylorEdPsych' were used for conducting Little's MCAR Test. The 'mice' (Van Buuren et al.,1999) package was used to impute incomplete multivariate data by chained equations. Although, simulation study was carried out in previous works to compare the performance of imputation method, the findings were based on different parameters such as missing data mechanism, sample size, percentage of missing data. This simulation study, however, dealt with all these parameters simultaneously for continuous missing variables. The simulation involves 1000 replications.

The procedure of the simulation process is as follows:

1. Simulate *n* data for two continuous independent variables from standard normal distributions *N* (0,1).
2. Calculate z = (0.7+ 1.08*X_1+1.69*X_2) and $\pi(x) = \frac{1}{1+e^{-z}}$
3. Simulate the data u from a random uniform distribution, U (0,1).
4. Generate the Y dependent variable for binary logistic regression by using the rule of y=1 if u$\leq$Pr(Y), if else assign y = 0.
5. Artificially create ten levels of missing rates under the MCAR mechanism: 5% to 50% (with 5% increment)
6. Test the assumption of MCAR using Little's Test of MCAR**.**
7. Apply data imputation method for each sample size with different missing rates.
8. Obtain MSE.
9. Perform 1000 simulations and average MSE over 1000 simulations.

Steps 1 to 9 were repeated for each sample size, *n*(30, 50, 100, 200, 500, 1000).

Study 2: Application to a Real Dataset

To validate the findings of the simulation study, the imputation methods were applied to a Human Resource (HR) dataset which was taken from the Kaggle website. The Human Resource dataset consists of 15000 observations with 10 features. The outcome variable is status which represents the status of employee turnover (1 = Left (23.81%) and 0 = Stay (76.19%)). The selected variables are *sl* and *le* which represent the values of satisfaction level and last evaluation, respectively. The variables *sl* and *le* are selected as these variables are continuous variable while the other variables (number of projects, average monthly hours, time spend at company, accident at work, promotion last 5 Years, position at work, level of salary) were classified as discrete and categorical variables. The summary statistics of satisfaction level (*sl*) and last evaluation (*le*) are given in Table 1.

Table 1
Summary Statistics

Variable	Group	Mean	Standard Deviation	Skewness	Kurtosis
Satisfaction Level (*sl*)	1=Left	0.4402	0.2640	0.291	-1.034
	0= Stay	0.6668	0.2171	-0.605	-0.215
Last Evaluation (*le*)	1=Left	0.7182	0.1977	-0.014	-1.710
	0= Stay	0.7155	0.1620	-0.039	-1.014

Using the original dataset, sample sizes of 30, 50, 100, 200, 500, and 1000 were randomly selected from the 15000 observations to study the effect of small, medium and large sample size. We follow the work by Cheema (2014), who selected 10 sub-samples size (small to large sample) from ($n = 10000$) simulated dataset to study the effect of sample size on the performance of missing data treatment. Missing data in the HR dataset were generated randomly for 5% to 50% missing rates. Single imputation and MICE methods were then applied and the MSE was obtained to evaluate the performance of each imputation method.

RESULTS AND DISCUSSION

Study 1: Simulation Results

This section presents the simulation results. Table 2 displays the result of Little's Test of MCAR which tests the assumption of MCAR. The result indicates that the assumption of MCAR is satisfied since the p-value for all sample sizes and missing rates is greater than 0.05. This test confirmed that missing data were simulated under the MCAR mechanism.

Table 3 shows the simulation results of each imputation method for different levels of missing rates and sample sizes. The best imputation method is determined based on the lowest MSE.

Based on the simulation results shown in Table 3 and line chart in Figure 1, the MSE of group mean imputation (GMI) is the lowest compared to OMI and MICE regardless of missing rates and small sample size. Imputation using MICE has the highest MSE compared to OMI and GMI regardless of missing rates and small sample size.

Thus, in terms of single imputation methods, GMI is more superior compared to OMI. The line chart in Figure 1 also show that the MSE for all three methods (OMI, GMI, MICE) increases when missing rate increases with MICE having the highest increase in MSE. This simulation results support the findings by Schmitt et al. (2015), whose results showed that the RMSE for mean imputation and MICE increased with increasing missing rates.

Overall, it can be concluded that single imputation method by group mean (GMI) is more superior compared to overall mean (OMI) and MICE in treating missing data for all sample sizes regardless of missing rates. In addition, it is not recommended to perform imputation using these methods if missing rate is more than 15% due to the large MSE.

Table 2
Little's Test for Missing Completely at Random (MCAR) results

Sample Size	p-value									
	Missing Rates (%)									
	5	10	15	20	25	30	35	40	45	50
30	0.514	0.6068	0.567	0.101	0.337	0.2901	0.202	0.152	0.082	0.125
50	0.867	0.2858	0.350	0.928	0.895	0.409	0.340	0.059	0.230	0.651
100	0.398	0.8264	0.409	0.136	0.817	0.077	0.800	0.700	0.766	0.141
200	0.910	0.3677	0.431	0.287	0.141	0.333	0.914	0.689	0.494	0.494
500	0.725	0.9612	0.324	0.568	0.575	0.978	0.879	0.959	0.959	0.782
1000	0.757	0.6522	0.858	0.097	0.057	0.281	0.145	0.237	0.436	0.879

Table 3
Mean Square Error (MSE) of Imputation Methods

Sample	Missing Rates									
Size	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
30	0.02927^{a}	0.08895^{a}	0.11587^{a}	0.17729^{a}	0.20589^{a}	0.26562^{a}	0.29670^{a}	0.35766^{a}	0.38986^{a}	0.45146^{a}
	0.02125^{b}	**0.07179**^{b}	**0.10279**^{b}	**0.14545**^{b}	**0.14761**^{b}	**0.19056**^{b}	**0.21375**^{b}	**0.25976**^{b}	**0.28295**^{b}	**0.33305**^{b}
	0.03965^{c}	0.12101^{c}	0.16148^{c}	0.24238^{c}	0.28065^{c}	0.35642^{c}	0.40207^{c}	0.48558^{c}	0.52647^{c}	0.61462^{c}
50	0.03387^{a}	0.08573^{a}	0.12013^{a}	0.17208^{a}	0.20512^{a}	0.25671^{a}	0.29322^{a}	0.34744^{a}	0.38039^{a}	0.43655^{a}
	0.02852^{b}	**0.04853**^{b}	**0.06673**^{b}	**0.09590**^{b}	**0.17525**^{b}	**0.21805**^{b}	**0.25093**^{b}	**0.29541**^{b}	**0.32792**^{b}	**0.37603**^{b}
	0.05551^{c}	0.13501^{c}	0.19408^{c}	0.27735^{c}	0.33524^{c}	0.41746^{c}	0.47461^{c}	0.56411^{c}	0.62280^{c}	0.70729^{c}
100	0.03845^{a}	0.07679^{a}	0.11372^{a}	0.15324^{a}	0.19081^{a}	0.22856^{a}	0.26824^{a}	0.30717^{a}	0.34525^{a}	0.38503^{a}
	0.03170^{b}	**0.06234**^{b}	**0.09321**^{b}	**0.12628**^{b}	**0.15665**^{b}	**0.18701**^{b}	**0.21976**^{b}	**0.25217**^{b}	**0.28441**^{b}	**0.31652**^{b}
	0.05982^{c}	0.11750^{c}	0.17922^{c}	0.23702^{c}	0.29774^{c}	0.35807^{c}	0.42519^{c}	0.48188^{c}	0.55260^{c}	0.61548^{c}
200	0.03458^{a}	0.06821^{a}	0.10335^{a}	0.13819^{a}	0.17159^{a}	0.20778^{a}	0.24161^{a}	0.27560^{a}	0.31157^{a}	0.34504^{a}
	0.02916^{b}	**0.05704**^{b}	**0.08658**^{b}	**0.11575**^{b}	**0.14402**^{b}	**0.17396**^{b}	**0.20243**^{b}	**0.23167**^{b}	**0.26216**^{b}	**0.29063**^{b}
	0.05844^{c}	0.11660^{c}	0.17840^{c}	0.23530^{c}	0.29690^{c}	0.35542^{c}	0.41574^{c}	0.47448^{c}	0.53077^{c}	0.58641^{c}
500	0.03350^{a}	0.06679^{a}	0.10000^{a}	0.13384^{a}	0.16717^{a}	0.20040^{a}	0.23398^{a}	0.26686^{a}	0.30102^{a}	0.33327^{a}
	0.02774^{b}	**0.05535**^{b}	**0.08313**^{b}	**0.11145**^{b}	**0.13881**^{b}	**0.16665**^{b}	**0.19434**^{b}	**0.22233**^{b}	**0.25044**^{b}	**0.27723**^{b}
	0.05276^{c}	0.10646^{c}	0.15951^{c}	0.21291^{c}	0.26779^{c}	0.32188^{c}	0.37583^{c}	0.43280^{c}	0.48800^{c}	0.54432^{c}
1000	0.03371^{a}	0.06703^{a}	0.10093^{a}	0.13446^{a}	0.16781^{a}	0.20181^{a}	0.23529^{a}	0.26870^{a}	0.30302^{a}	0.33634^{a}
	0.02772^{b}	**0.05514**^{b}	**0.08292**^{b}	**0.11064**^{b}	**0.13803**^{b}	**0.16581**^{b}	**0.19344**^{b}	**0.22087**^{b}	**0.24896**^{b}	**0.27680**^{b}
	0.05449^{c}	0.10949^{c}	0.16325^{c}	0.21909^{c}	0.27310^{c}	0.32938^{c}	0.38511^{c}	0.43943^{c}	0.49507^{c}	0.55291^{c}

Notes: a=OMI; b=GMI; c=MICE

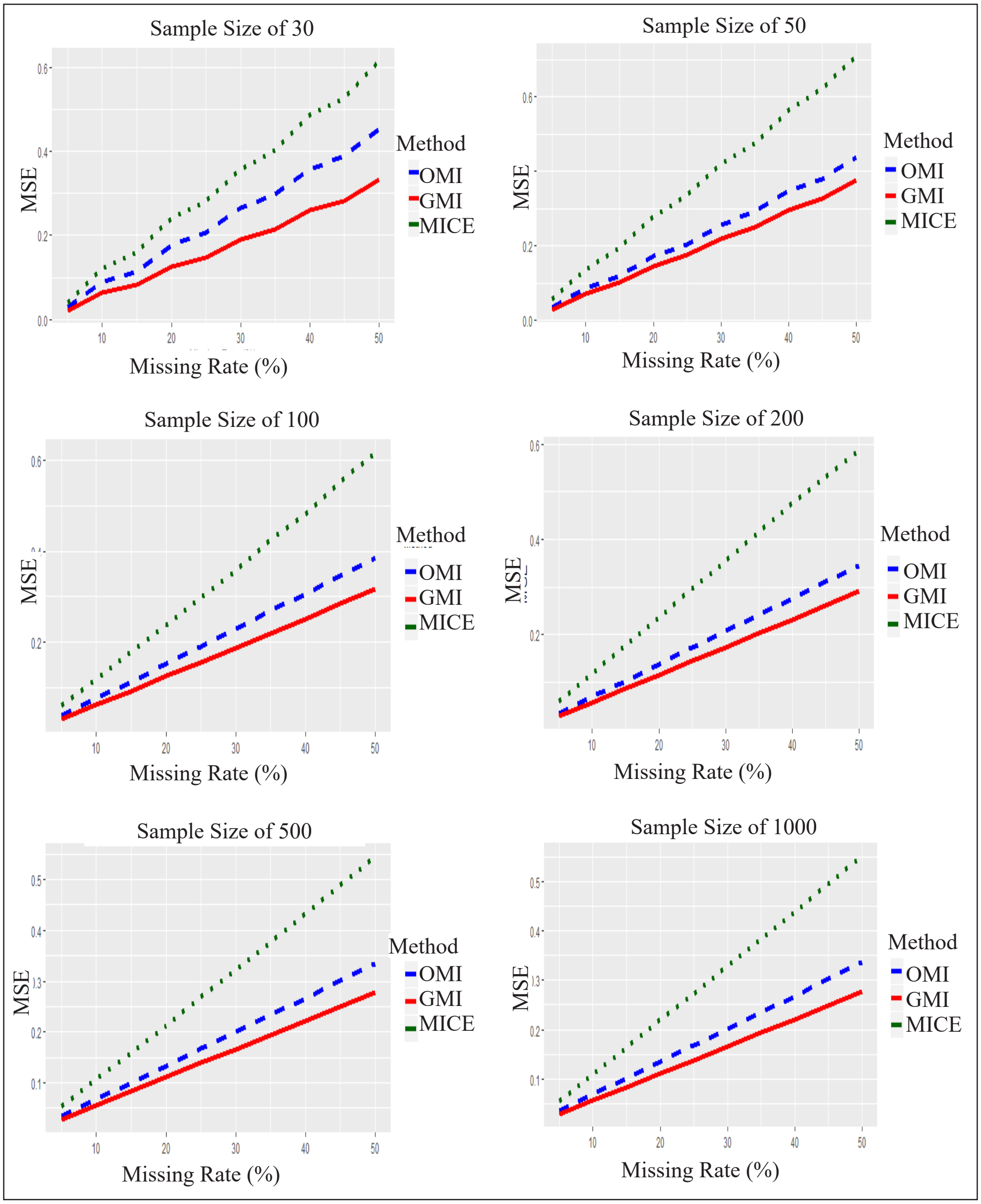

Figure 1. Line chart of MSE for different missing rates

Performance Measures of Different Imputation Methods

The pattern of data imputation methods was observed based on MSE for different missing rate and sample size using clustered boxplots. The clustered boxplots of MSE for the three imputation methods are shown in Figures 2 and 3. Figure 2 shows the pattern of MSE for different missing rate at a fixed sample size. The clustered boxplots show that MSE of single imputation (OMI and GMI) are quite similar for samples sizes ($n = 30, 50$) and missing rate (5%). For sample sizes 100 and above, GMI has the lowest MSE for all missing rates. In Figure 3, the MSE for different sample size at a fixed missing rate was observed. The clustered boxplots show that the variability of MSE decreases as sample size increases. It can be concluded that the GMI method is more superior compared to OMI and MICE for all sample sizes regardless of missing rates.

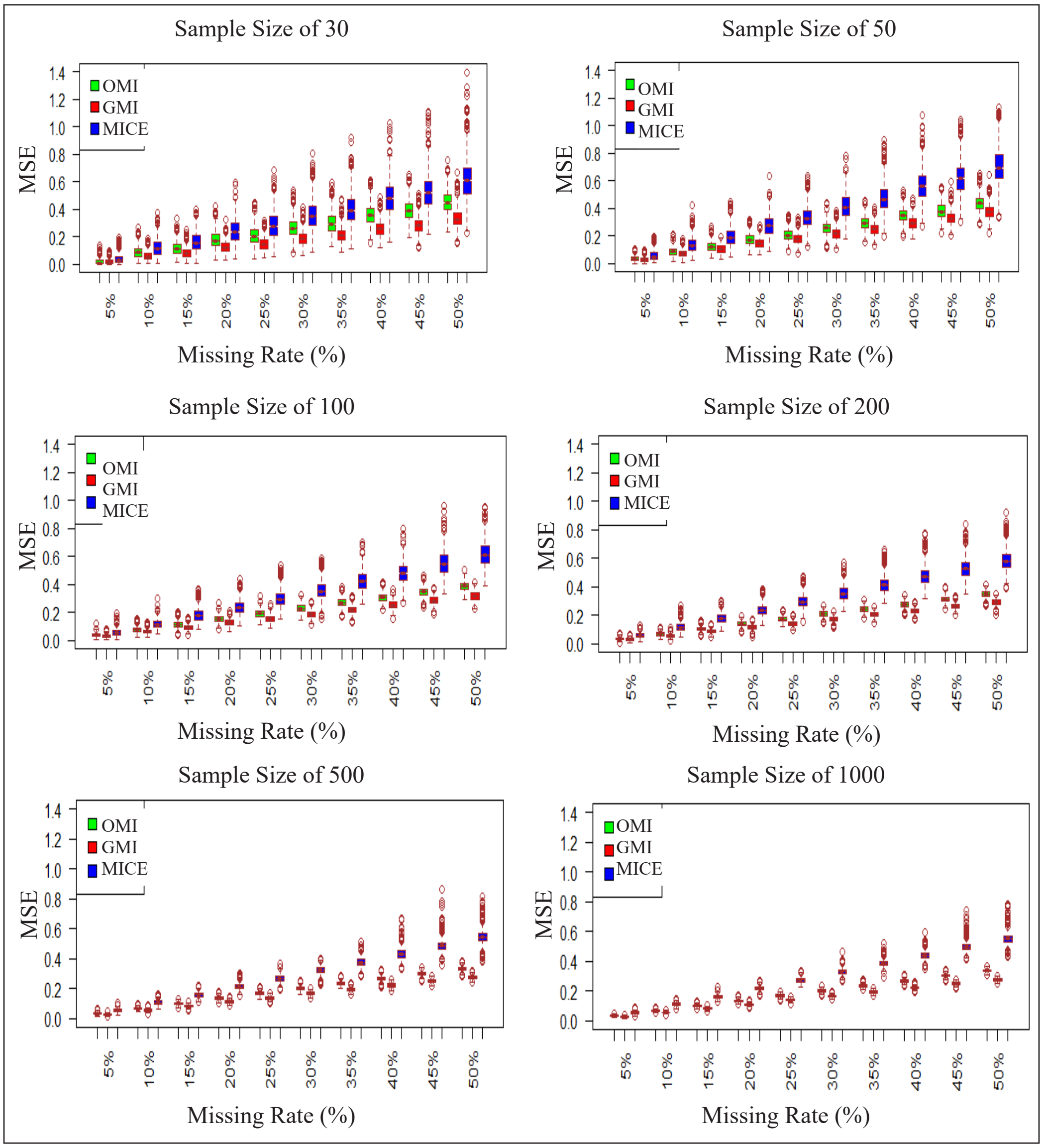

Figure 2. Clustered boxplots of MSE for different missing rates

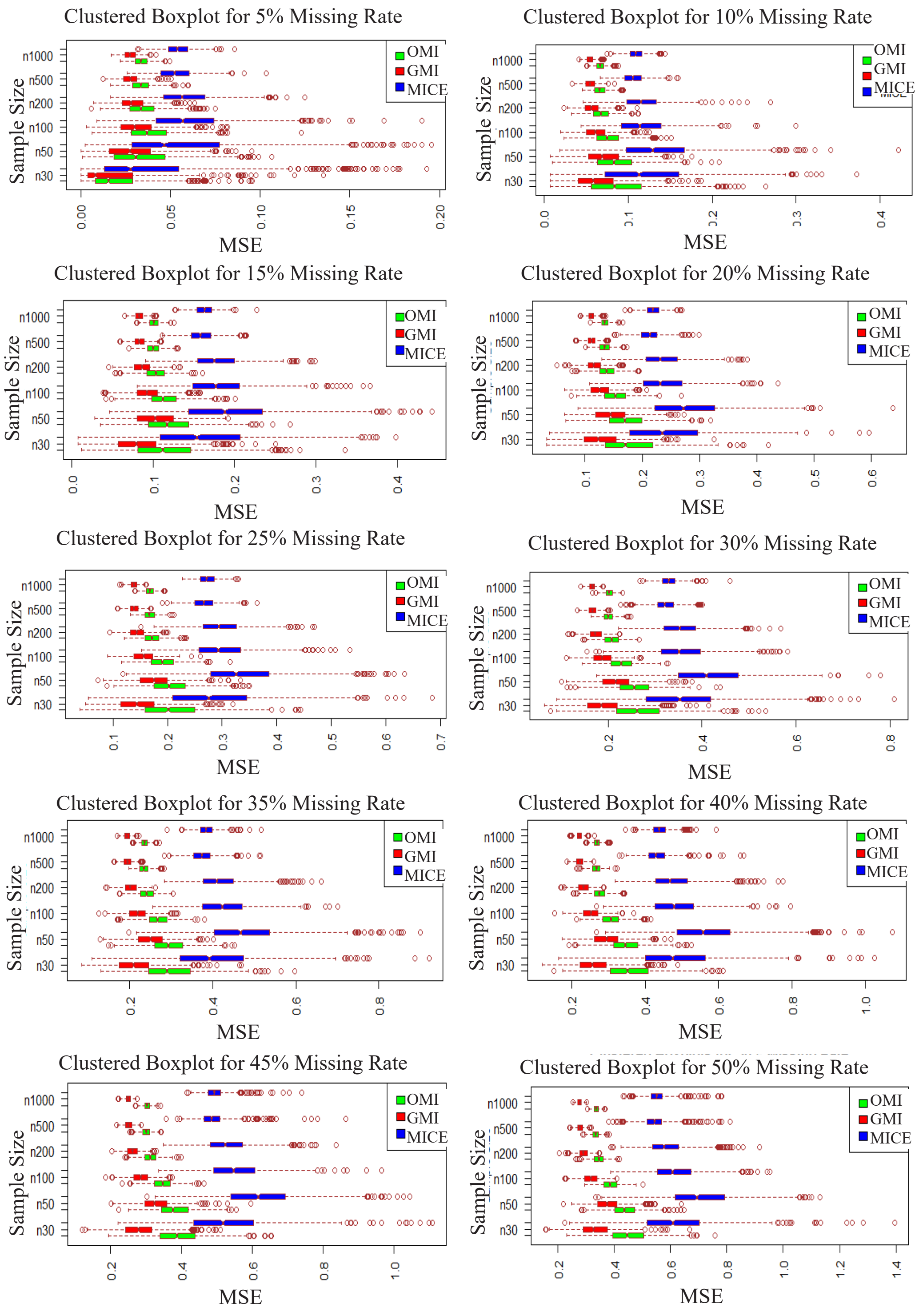

Figure 3. Clustered boxplots of MSE for different sample size

Table 4 shows the recommendation of imputation methods for each level of missing rate and sample size based on the simulation results.

Table 4
Recommended Method for Each Sample Size and Missing Rate

Sample Size	Missing Rates									
	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
30	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI
50	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI
100	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI
200	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI
500	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI
1000	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI	GMI

Study 2: Results of Application to Real Dataset

The imputation methods were then applied to the HR dataset. Table 5 displays the result of Little's Test of MCAR for the real dataset. This test confirmed that missing data in the real dataset satisfied the assumption of the MCAR mechanism. Table 6 summarizes the results of the experiment. The MSE for GMI were observed to be lowest across different levels of missing rates and sample sizes. This indicates that the single imputation method using GMI outperformed the OMI and MICE imputation methods. The MSE for both single imputation methods and MICE increased with increasing missing rates. These results also confirmed the finding of the simulation results.

CONCLUSION

The study evaluated the performance of single imputation methods using mean with MICE (Multivariate Imputation with Chained Equations). The result of Little's Test of MCAR for both simulation study and real dataset confirmed that missing data satisfied the assumption of MCAR. The simulation results showed that GMI is superior compared to OMI and MICE for all sample size, regardless of missing rates. However, OMI or GMI can be used when samples size is small $(n = 30, 50)$ and missing rate is low (5%). The MSE for OMI, GMI and MICE increased with increasing missing rates. MICE was found not to perform well especially when missing rate was high (more than 15%). The MSE of MICE appeared to be high when missing rates were more than 15% compared to when missing rates were below 15%. An application to a real dataset confirmed the findings of the simulation results that GMI performed well for all sample sizes and missing rates. There are several other advanced imputation methods such as Regression Imputation, Regression Tree Imputation, and KNN Imputation methods. The limitation of this study is that only continuous missing attribute was considered. Future work will include a simulation study involving categorical

Table 5
Little's Test for MCAR: Real Dataset

Sample Size	p-value									
	Missing Rates (%)									
	5	10	15	20	25	30	35	40	45	50
30	0.808	0.853	0.270	0.828	0.829	0.886	0.992	0.482	0.441	0.470
50	0.744	0.259	0.953	0.067	0.760	0.777	0.291	0.673	0.492	0.314
100	0.820	0.296	0.202	0.625	0.901	0.806	0.616	0.163	0.052	0.564
200	0.939	0.664	0.369	0.369	0.564	0.095	0.782	0.846	0.881	0.833
500	0.190	0.083	0.800	0.110	0.411	0.563	0.136	0.723	0.614	0.112
1000	0.724	0.732	0.984	0.472	0.711	0.203	0.695	0.887	0.745	0.615

Table 6
Mean Square Error (MSE) of Imputation Methods (Real Dataset)

Sample Size	Missing rates									
	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
30	0.00019^{a}	0.00302^{a}	0.00443^{a}	0.00470^{a}	0.00965^{a}	0.01030^{a}	0.01084^{a}	0.01324^{a}	0.01584^{a}	0.01587^{a}
	0.00009^{b}	**0.00230^{b}**	**0.00251^{b}**	**0.00264^{b}**	**0.00376^{b}**	**0.00440^{b}**	**0.00547^{b}**	**0.00590^{b}**	**0.00736^{b}**	**0.00882^{b}**
	0.00045^{c}	0.00460^{c}	0.00712^{c}	0.01012^{c}	0.01261^{c}	0.01507^{c}	0.01602^{c}	0.01870^{c}	0.01985^{c}	0.01995^{c}
50	0.00056^{a}	0.00391^{a}	0.00439^{a}	0.00603^{a}	0.00654^{a}	0.01034^{a}	0.01166^{a}	0.01312^{a}	0.01315^{a}	0.01513^{a}
	0.00006^{b}	**0.00190^{b}**	**0.00242^{b}**	**0.00421^{b}**	**0.00444^{b}**	**0.00748^{b}**	**0.00883^{b}**	**0.00998^{b}**	**0.01020^{b}**	**0.01368^{b}**
	0.00096^{c}	0.00475^{c}	0.00486^{c}	0.01011^{c}	0.01192^{c}	0.01494^{c}	0.01790^{c}	0.01894^{c}	0.01963^{c}	0.02066^{c}
100	0.00073^{a}	0.00242^{a}	0.00407^{a}	0.00544^{a}	0.00641^{a}	0.00806^{a}	0.00864^{a}	0.01020^{a}	0.01124^{a}	0.01261^{a}
	0.00071^{b}	**0.00224^{b}**	**0.00305^{b}**	**0.00494^{b}**	**0.00618^{b}**	**0.00774^{b}**	**0.00810^{b}**	**0.00903^{b}**	**0.01012^{b}**	**0.01109^{b}**
	0.00120^{c}	0.00230^{c}	0.00782^{c}	0.01297^{c}	0.01356^{c}	0.01746^{c}	0.01772^{c}	0.01782^{c}	0.02346^{c}	0.02446^{c}
200	0.00151^{a}	0.00369^{a}	0.00545^{a}	0.00575^{a}	0.00706^{a}	0.00907^{a}	0.01193^{a}	0.01349^{a}	0.01473^{a}	0.01544^{a}
	0.00113^{b}	**0.00294^{b}**	**0.00419^{b}**	**0.00444^{b}**	**0.00589^{b}**	**0.00725^{b}**	**0.00965^{b}**	**0.01058^{b}**	**0.01183^{b}**	**0.01214^{b}**
	0.00219^{c}	0.00470^{c}	0.00706^{c}	0.01256^{c}	0.01313^{c}	0.01736^{c}	0.01738^{c}	0.02068^{c}	0.02221^{c}	0.02705^{c}
500	0.00145^{a}	0.00289^{a}	0.00473^{a}	0.00629^{a}	0.00757^{a}	0.00908^{a}	0.01036^{a}	0.01161^{a}	0.01275^{a}	0.01476^{a}
	0.00136^{b}	**0.00252^{b}**	**0.00410^{b}**	**0.00565^{b}**	**0.00683^{b}**	**0.00816^{b}**	**0.00925^{b}**	**0.01043^{b}**	**0.01141^{b}**	**0.01305^{b}**
	0.00205^{c}	0.00486^{c}	0.00736^{c}	0.01142^{c}	0.01309^{c}	0.02197^{c}	0.02297^{c}	0.02383^{c}	0.02438^{c}	0.02799^{c}
1000	0.00155^{a}	0.00296^{a}	0.00427^{a}	0.00598^{a}	0.00750^{a}	0.00918^{a}	0.01073^{a}	0.01262^{a}	0.01397^{a}	0.01527^{a}
	0.00137^{b}	**0.00256^{b}**	**0.00368^{b}**	**0.00502^{b}**	**0.00651^{b}**	**0.00790^{b}**	**0.00923^{b}**	**0.00987^{b}**	**0.01214^{b}**	**0.01389^{b}**
	0.00242^{c}	0.00483^{c}	0.00858^{c}	0.01132^{c}	0.01286^{c}	0.01676^{c}	0.02029^{c}	0.02207^{c}	0.02429^{c}	0.03114^{c}

Notes: a=OMI; b=GMI; c=MICE

or mixed types of variables. The issue of missing data for auto correlated data or time series data also will be interesting for future research.

The R-syntax for the simulation study and data imputation methods using R programming can be obtained from the first author, Nurul Azifah Mohd Pauzi. The basic data imputation R commands are given in the Appendix.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to Dr Hamzah Abdul Hamid from Institute of Engineering Mathematics, Universiti Malaysia Perlis for his permission to use his simulation model as a basis to generate the data for the simulation study. We also thank the Research Management Centre, Universiti Teknologi MARA, Shah Alam, Selangor for the funding of the publication of this paper.

REFERENCES

Abidin, N. Z., Ismail, A. R., & Emran, N. A. (2018). Performance analysis of machine learning algorithms for missing value imputation. *International Journal of Advanced Computer Science and Applications*, *9*(6), 442-447.

Aljuaid, T., & Sasi, S. (2016). Proper imputation techniques for missing values in data sets. In *International Conference on Data Science and Engineering (ICDSE)* (pp. 1-5). IEEE Conference Publication. https://doi.org/10.1109/ICDSE.2016.7823957

Ayilara, O. F., Zhang, L., Sajobi, T. T., Sawatzky, R., Bohm, E., & Lix, L. M. (2019). Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health and Quality of Life Outcomes*, *17*(1), 106. https://doi.org/10.1186/s12955-019-1181-2

Barnett, A. G., McElwee, P., Nathan, A., Burton, N. W., & Turrell, G. (2017). Identifying patterns of item missing survey data using latent groups: An observational study. *BMJ Open*, *7*(10), 1-9. https://doi.org/10.1136/bmjopen-2017-017284

Bhati, S., & Gupta, M. K. (2016). Missing data imputation for medical database: Review. *International Journal of Advanced Research in Computer Science and Software Engineering*, *6*(4), 754-758.

Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 1-68.

Chaudhry, A., Li, W., Basri, A., & Patenaude, F. (2019). A method for improving imputation and prediction accuracy of highly seasonal univariate data with large periods of missingness. *Wireless Communications and Mobile Computing*, *2019*, 1-13. https://doi.org/10.1155/2019/4039758

Cheema, J. R. (2014). Some general guidelines for choosing missing data handling methods in educational research. *Journal of Modern Applied Statistical Methods*, *13*(2), 53-75. https://doi.org/10.22237/jmasm/1414814520

Chhabra, G., Vashisht, V., & Ranjan, J. (2017). A comparison of multiple imputation methods for data with missing values. *Indian Journal of Science and Technology*, *10*(19), 1-7. https://doi.org/10.17485/ijst/2017/v10i19/110646

Dettori, J. R., Norvell, D. C., & Chapman, J. R. (2018). The sin of missing data: Is all forgiven by way of imputation? *Global Spine Journal*, *8*(8), 892-894. https://doi.org/10.1177/2192568218811922

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 1-17. https://doi.org/10.1186/2193-1801-2-222

Fichman, M., & Cummings, J. N. (2003). Multiple imputation for missing data: Making the most of what you know. *Organizational Research Methods*, *6*(3), 282-308. https://doi.org/10.1177/1094428103255532

Gad, A. M., & Abdelkhalek, R. H. M. (2017). Imputation methods for longitudinal data: A comparative study. *International Journal of Statistical Distributions and Applications*, *3*(4), 72. https://doi.org/10.11648/j.ijsd.20170304.13

Gopal, K. M., Durgaprasad, N., Deepa, K. S., Sravan, R. G., & Revanth, R. D. (2019). Comparative analysis of different imputation techniques for handling missing dataset. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8*(7), 347-351.

Goretzko, D., Heumann, C., & Bühner, M. (2019). Investigating parallel analysis in the context of missing data: A simulation study comparing six missing data methods. *Educational and Psychological Measurement*, *80*(4), 756-774. https://doi.org/10.1177/0013164419893413

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data at level 2: A comparison of fully conditional and joint modeling in multilevel designs. *Journal of Educational and Behavioral Statistics*, *43*(3), 316-353. https://doi.org/10.3102/1076998617738087

Hughes, R. A., Heron, J., Sterne, J. A., & Tilling, K. (2019). Accounting for missing data in statistical analyses: Multiple imputation is not always the answer. *International Journal of Epidemiology*, *48*(4), 1294-1304. https://doi.org/10.1093/ije/dyz032

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, *33*(10), 913-933. https://doi.org/10.1080/08839514.2019.1637138

Kaiser, J. (2014). Dealing with missing values in data. *Journal of Systems Integration*, *5*(1), 42- 51. http://dx.doi.org/10.20470/jsi.v5i1.178

Kamatchi P, L., & Baranidharan, C. (2019). Missing data imputation methods for autism prediction. *International Journal of Recent Technology and Engineering*, 8(5), 940-944.

Le, T. D., Beuran, R., & Tan, Y. (2018). Comparison of the most influential missing data imputation algorithms for healthcare. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)* (pp. 247-251). IEEE Conference Publication. http://dx.doi.org/10.1109/KSE.2018.8573344

Li, Y., Ji, L., Oravecz, Z., Brick, T. R., Hunter, M. D., & Chow, S. M. (2019). dynr. mi: An R program for multiple imputation in dynamic modeling. *World Academy of Science, Engineering and Technology*, *13*(5), 302-311. https://doi.org/10.5281/zenodo.3298841

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of The American Statistical Association*, *83*(404), 1198-1202.

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley & Sons.

Lo, A. W., Siah, K. W., & Wong, C. H. (2019). Machine learning with statistical imputation for predicting drug approvals. *Harvard Data Science Review*, *1*(1), 1-25. https://doi.org/10.1162/99608f92.5c5f0525

Ma, Z., & Chen, G. (2018). Bayesian methods for dealing with missing data problems. *Journal of The Korean Statistical Society*, *47*(3), 297-313. https://doi.org/10.1016/j.jkss.2018.03.002

Madley-Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, *110*, 63-73. https://doi.org/10.1016/j.jclinepi.2019.02.016

Malarvizhi, M. R., & Thanamani, A. S. (2012). K-Nearest Neighbor in missing data imputation. *International Journal of Engineering Research and Development*, *5*(1), 5-7.

Masconi, K. L., Matsha, T. E., Echouffo-Tcheugui, J. B., Erasmus, R. T., & Kengne, A. P. (2015). Reporting and handling of missing data in predictive research for prevalent undiagnosed type 2 diabetes mellitus: A systematic review. *The EPMA Journal*, *6*(1), 1-11. https://doi.org/10.1186/s13167-015-0028-0

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*, *6*(3), 328-362. https://doi.org/10.1177/1094428103254673

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, *17*(4), 372-411. https://doi.org/10.1177/1094428114548590

Nwakuya, M. T., & Nwabueze, J. C. (2018). Comparison of shrinkage–based estimators in the presence of missing data: A multiple imputation analysis. *International Journal of Statistics and Applications, 8*(6), 305-308. https://doi.org/10.5923/j.statistics.20180806.03

Ochieng'Odhiambo, F. (2020). Comparative study of various methods of handling missing data. *Mathematical Modelling and Applications*, *5*(2), 87.

Pampaka, M., Hutcheson, G., & Williams, J. (2016). Handling missing data: Analysis of a challenging data set using multiple imputation. *International Journal of Research & Method in Education*, *39*(1), 19-37. https://doi.org/10.1080/1743727X.2014.979146

Papageorgiou, G., Grant, S. W., Takkenberg, J. J., & Mokhles, M. M. (2018). Statistical primer: How to deal with missing data in scientific research? *Interactive Cardiovascular and Thoracic Surgery*, *27*(2), 153-158. https://doi.org/10.1093/icvts/ivy102

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., & Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology*, *9*, 157-166. https://doi.org/10.2147/CLEP.S129785

Ratolojanahary, R., Ngouna, R. H., Medjaher, K., Junca-Bourié, J., Dauriac, F., & Sebilo, M. (2019). Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset. *Expert Systems with Applications*, *131*, 299-307. https://doi.org/10.1016/j.eswa.2019.04.049

Salgado C. M., Azevedo C., Proença H., & Vieira S. M. (2016) Missing data. In *Secondary analysis of electronic health records* (pp. 143-162). Springer.

Scheffer, J. (2002). Dealing with missing data. *Research Letters in the Information and Mathematical Sciences, 3*, 153-160.

Schmitt, P., Mandel, J., & Guedj, M. (2015). A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics*, *6*(1), 1-6. https://doi.org/10.472/2155-6180.1000224

Shi, D., Lee, T., Fairchild, A. J., &Maydeu-Olivares, A. (2019). Fitting ordinal factor analysis models with missing data: A comparison between pairwise deletion and multiple imputation. *Educational and Psychological Measurement*, *80*(1), 41-66. https://doi.org/10.1177/0013164419845039

Sim, J., Lee, J. S., & Kwon, O. (2015). Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications. *Mathematical Problems in Engineering*, *2015*, 1-14. https://doi.org/10.1155/2015/538613

Song, Q., & Shepperd, M. (2007). Missing data imputation techniques. *International Journal of Business Intelligence and Data Mining*, *2*(3), 261-291. https://doi.org/10.1504/IJBIDM.2007.015485

Stavseth, M. R., Clausen, T., &Røislien, J. (2019). How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. *SAGE Open Medicine*, *7*, 1-12. https://doi.org/10.1177/2050312118822912

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest - Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112-118. https://doi.org/10.1093/bioinformatics/btr597

Sullivan, T. R., White, I. R., Salter, A. B., Ryan, P., & Lee, K. J. (2018). Should multiple imputation be the method of choice for handling missing data in randomized trials? *Statistical Methods in Medical Research*, *27*(9), 2610-2626. https://doi.org/10.1177/0962280216683570

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5). Pearson.

Turner, E. L., Yao, L., Li, F., & Prague, M. (2019). Properties and pitfalls of weighting as an alternative to multilevel multiple imputation in cluster randomized trials with missing binary outcomes under covariate-dependent missingness. *Statistical Methods in Medical Research*, *29*(5), 1338-1353. https://doi.org/10.1177/0962280219859915

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219-242. https://doi.org/10.1177/0962280206074463

Van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, *18*(6), 681-694. https://doi.org/10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R

van Ginkel, J. R., Linting, M., Rippe, R. C., & van der Voort, A. (2019). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, *102*(3), 297-308. https://doi.org/10.1080/00223891.2018.1530680

Wah, Y. B., Ibrahim, N., Hamid, H. A., Abdul-Rahman, S., & Fong, S. (2018). Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Pertanika Journal of Science & Technology*, *26*, 329-340.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika, 24*(3/4), 471-494. https://doi.org/10.2307/2331979

Yadav, M. L., & Roychoudhury, B. (2018). Handling missing values: A study of popular imputation packages in R. *Knowledge-Based Systems*, *160,* 104-118. https://doi.org/10.1016/j.knosys.2018.06.012

Zhang, Z. (2016). Missing data imputation: focusing on single imputation. *Annals of Translational Medicine, 4*(1), 1-9. https://doi.org/10.3978/j.issn.2305-5839.2015.12.38