Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 26 (2023), No. 4, pp. 915–927
DOI : 10.47974/JSMS-971

Comparing the performance of eight imputation methods for propensity score matching in missing data problem

Imran Kurt Omurlu *
Division of Biostatistics
Faculty of Medicine
Adnan Menderes University
Aydin
Turkey

Bugra Varol
Division of Biostatistics
Institute of Health Sciences
Adnan Menderes University
Aydin
Turkey

Mevlut Ture
Division of Biostatistics
Faculty of Medicine
Adnan Menderes University
Aydin
Turkey

Abstract

Propensity score (PS) is a popular method to control for covariates in observational studies. A challenge in PS analyses is missing values in covariates. This study aims to investigate how different imputation methods of handling missing values of covariates in a PS analysis can affect average treatment on the treated (ATT) estimates. In this study, missing data imputation methods were evaluated using different data sets, whose covariates were low, medium, and high (r=0.10, 0.50, 0.85) correlated with each other, for n=200 units and 1000 times running simulation. Missing data structures were created according to the missing at random (MAR) mechanism and different missing rates. Different datasets were obtained after having imputed the missing values separately by eight imputation methods including mean, median, mode, hot deck, last observation carried forward (LOCF), next observation carried backward (NOCB), regression and predictive mean matching (PMM).

*E-mail: ikurtomurlu@gmail.com

Then the PS nearest neighbor matching was implemented and ATT scores were obtained using the imputed data sets. The predictive performance of imputation methods was compared according to ATT scores by hierarchical cluster analysis with Euclidean distance complete linkage. ATT scores of regression and PMM methods were closer to each other and these methods showed the best predictive performance. Additionally, when there were larger amounts of missing data, the PMM was the best method of choice. Ignoring missing values on covariates for PS analyses causes information loss significantly and this information loss becomes greater as the rate of missing data increases. PS analyses might be biased if missing data on covariates are also ignored. To prevent this information loss and bias, PS analyses should be performed after solving the problem of missing data with MAR mechanism on covariates by regression and PMM methods, which showed statistical superiority compared to other methods in this study.

Subject Classification: *62B15, 62D10, 62H30.*

Keywords: *Missing data, Imputation, Simulation, Propensity score, Hierarchical clustering.*

1. Introduction

Missing data is a big problem in the application of statistical analysis. In statistics, missing values occur when observation values for a variable cannot be obtained for various reasons. The presence of missing data that can significantly affect the data analysis process is a common fact in clinical research. Missing data cause biased results as well as a loss of statistical power and precision [12, 28].

The propensity score (PS) is used to balance the treatment and control groups for covariates and to reduce the systematic error in estimating treatment effects. For estimating the PS values covariates are generally included in the model. For this reason the covariates with missing values could be in the data. One of the essential unresolved trouble in PS estimation is how to handle missing values in covariates [6, 8, 18, 22]. Therefore, it is important to deal with the missing data problem. In order to overcome this problem, many approaches have been proposed in the literature. Missing data imputation methods are often used to deal with missing data [5, 28].

The principal focus of this study was on comparing the performance of eight missing data imputation approaches including mean, median, mode, hot deck, last observation carried forward (LOCF), next observation carried backward (NOCB), regression and predictive mean matching (PMM) to implementing in a PS matching for incompletely observed covariates with missing at random (MAR) mechanism to estimate average treatment effect on the treated (ATT) following missing value imputation. Further, hierarchical clustering with Euclidean distance complete linkage was used to identify imputation methods with similar performance in our

study. Clustering is one of the most successful methods that group similar objects into different groups according to some defined distance measure [20].

2. Material and methods

2.1. *Imputation methods for dealing with missing data*

Imputation is the process of substituting missing data by optimal values. Some data analysis techniques are not robust to missingness, and require to impute the missing data. Depending on why the values are missing, imputation methods can deliver reliable results at reasonable level. Various missing data imputation methods are used to deal with missing data [16, 27]. The imputation methods used in our study are explained below.

Mean/Median/Mode imputation

These imputations are such methods that the mean/median/mode of the observed values for each variable is obtained and the missing data for that variable is imputed by this mean/median/mode. They can be preferred because of their fast and practicality [14, 21].

Hot deck imputation

A widely-used method is that the missing value is imputed from a randomly selected similar record. All observations in the data set are divided into groups according to similar characteristics. A random observation is chosen from the group containing the data to be imputed missing data. The value in this selected observation is imputed to the missing data [16, 21].

Last observation carried forward (LOCF)

Another popular approach handling missing data is the LOCF. This method has been widely used in clinically experimental studies. The last observed non-missing value is used to fill in missing values at a later point in the data set [9, 24].

Next observation carried backward (NOCB)

NOCB is a similar method to LOCF. However this method processes opposite way from LOCF and it takes the first observation after the missing values and carries it backward [9, 15].

Regression imputation

It is based on replacing missing values by predicted scores from a linear regression by using data from the complete variables. Suppose

there is a variable x_1 that has some cases with missing data and a variable x_2 (with no missing data) that are used to impute x_1 [12, 16].

$x_1 = (x_{11}, \ldots, x_{1i}, \ldots, x_{1j}, \ldots, x_{1n})$ where x_{1i} refers to missing subjects and x_{1j} refers to no missing subjects.

$x_2 = (x_{21}, \ldots, x_{2i}, \ldots, x_{2j}, \ldots, x_{2n})$. All subjects of x_2 are complete.

Using no missing subjects on x_1, a linear regression is applied for calculate and and x_{1i} is estimated by $\hat{x}_{1i} = b_0 + b_1 x_{2i}$ [12, 16].

Predictive mean matching (PMM)

PMM is a robust method to do imputation for missing data, especially for imputing quantitative variables. PMM assumes that the missing data are MAR. Following steps are applied for PMM method [1, 25, 29]:

1. For no missing subjects on x_1 (x_{1j}), estimate $\hat{x}_{1j}$ from a linear regression $\hat{x}_{1j} = b_0 + b_1 x_{2j}$.
2. Draw randomly from bivariate normal distribution as being $\boldsymbol{b}^* = (b_0^*, b_1^*) \sim N(\boldsymbol{b}, \boldsymbol{\Sigma}_{\boldsymbol{b}})$ where $\boldsymbol{b} = (b_0, b_1)$ and $\boldsymbol{\Sigma}_{\boldsymbol{b}}$ covariance matrix of $\boldsymbol{b}$.
3. Calculate $\hat{x}_{1i} = b_0^* + b_1^* x_{2i}$.
4. For each $\hat{x}_{1i}$ find $\Delta^i = |\hat{x}_{1j} - \hat{x}_{1i}|$.
5. Randomly sample one value from $(\Delta^1, \Delta^2, \Delta^3)$ where Δ^1, Δ^2 and Δ^3 are the three smallest elements in Δ^i, respectively, and take the corresponding x_{1j} as the imputation.

2.2. *Propensity score*

PS is a method used especially in observational studies to correct or eliminate the systematic error in the variables studied. The PS method is used to balance the treatment and control groups for covariates and to reduce the systematic error in estimating treatment effects. In a data set of n units with no missing observations, PS value of the subject i is the conditional probability of assigning a specific treatment ($T_i = 1$) to the control group ($T_i = 0$) according to the observed covariate vector x_i and, expressed as following [2, 7, 27]:

$$e(x_i) = P(T_i = 1 \mid X_i = x_i)$$

Under the assumption that X_i and T_i are independent with respect to $e(x_i)$, expressed as follows [2]:

$$P(T_i = t_1, \ldots, T_n = t_n \mid X_i = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} e(x_i)^{t_i} \{1 - e(x_i)\}^{1-(t_i)}$$

Since the prediction scores obtained from logistic regression show the probability of subjects to obtain the relevant treatment, it varies between 0 and 1. Subjects whose PS values are equal or close to each other have a similar chance of being assigned to the treatment group [7, 10].

The PS matching method is expressed as 1:1 matching if a subject in the treatment group is matched with only one subject from the control group. Using nearest neighbor matching (NNM) approach, each treated subject is matched to the untreated subject whose PS value is closest to that of the treated subject [3]. Each untreated subject is matched at most once. At the end of the process, there are no mismatched subjects in the treatment group [7, 10].

Many statistical methods have been proposed to estimate ATT. The ATT is an estimation of the average treatment effect for the group who actually participated in the treating program. To estimate the ATT, PS matching is first performed using a 1:1 NNM without replacement algorithm [4, 11, 30].

$$ATT = E[(Y_i(1) - Y_i(0) \mid T_i = 1)]$$

Here $Y_i(1)$ is the expected score for subject *i* that treated, $Y_i(0)$ is the expected score for subject *i* that untreated.

2.3. *Simulation methods*

It was aimed to generate three different complete data sets to examine the performance of imputation methods under MAR missing data mechanism. The only difference between these data sets were the relationship between the covariates in the data generation phase. Following simulation algorithm detailed below was applied for n = 200 subject and was conducted with *R* Statistical Software version 4.0.3.

2.3.1. Simulation algorithm

1. Two covariates (x_1 and x_2) with low, medium and high correlated (0.10, 0.50, 0.85) were derived from bivariate standard normal distribution as being $\boldsymbol{X} = (x_1, x_2) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
2. The value of T was assigned from Binomial distributions as being $T \sim B(n, p)$ where n=200 and p (the probability of treatment assignment) defined as a linear function of x_1 and x_2 from inverse logit function. The function that forms p were determined such that

approximately 20% of the sample size was assigned to the treatment group ($T = 1$) and 80% of the sample size was assigned to the control group ($T = 0$).

3. Potential outcomes $Y_0 \sim N(\mu_0, 1)$ and $Y_1 \sim N(\mu_1, 1)$ where μ_0 and μ_1 to be polynomial function of $\boldsymbol{X}$ and $\mu_1 > \mu_0$ were generated.
4. Observed outcome (Y) was determined as follows:

$$Y = \begin{cases} Y_1, & T=1 \\ Y_0, & T=0 \end{cases}$$

5. Missing data sets with MAR mechanism were constituted by deleting data on x_1 depending on x_2 in different ratios (5%, 15%, 25% and 40%).
6. Then these datasets were completed by eight different data imputation methods including mean, median, mode, hot deck, LOCF, NOCB, regression and PMM methods.
7. Matching on the logit of the PS using NNM [1:1] was performed using the complete (reference) and imputed data sets and ATT was estimated by PS matching. ATT scores obtained from complete data sets were named as reference.
8. Steps 1. – 7. were repeated 1000 times and means and standard deviations of ATT scores were obtained.

Hierarchical clustering with Euclidean distance complete linkage was applied to results separately according to each missing rate and correlation structure. These methods were clustered according to the means of ATT scores and performance evaluation was made according to the clustering status of other scores around the reference score.

Means of ATT scores were edited such that references and imputation methods were observations and missing rates were variables and hierarchical clustering was applied immediately after, then in a different way, edited again such that references and imputation methods were observations and correlation structures were variables and hierarchical clustering was applied again.

3. Results

ATT scores obtained from complete and imputed datasets for each missing rate and correlation structure were approximately normally distributed. Descriptive statistics of ATT scores were specified as mean ± standard deviation (Table 1). The clustering results were shown by dendrogram graphs (Figure 1 and 2).

Table 1

ATT scores obtained from complete and imputed data sets

		MISSING DATA RATE			
SIMULATION	**METHODS**	**5%**	**15%**	**25%**	**40%**
Low Correlated	**Reference**	3.245 ± 0.421	3.245 ± 0.421	3.245 ± 0.421	3.245 ± 0.421
	Hot Deck	3.224 ± 0.385	3.283 ± 0.362	3.311 ± 0.350	3.356 ± 0.342
	NOCB	3.236 ± 0.372	3.276 ± 0.367	3.325 ± 0.347	3.361 ± 0.337
	Mean	3.229 ± 0.386	3.268 ± 0.375	3.280 ± 0.358	3.318 ± 0.357
	Median	3.235 ± 0.388	3.266 ± 0.378	3.277 ± 0.366	3.325 ± 0.357
	PMM	3.258 ± 0.399	3.279 ± 0.411	3.257 ± 0.425	3.278 ± 0.452
	Mode	3.229 ± 0.386	3.268 ± 0.375	3.280 ± 0.358	3.318 ± 0.357
	Regression	3.238 ± 0.390	3.263 ± 0.368	3.277 ± 0.360	3.308 ± 0.360
	LOCF	3.226 ± 0.377	3.274 ± 0.358	3.317 ± 0.347	3.366 ± 0.340
Medium Correlated	**Reference**	2.932 ± 0.397	2.932 ± 0.397	2.932 ± 0.397	2.932 ± 0.397
	Hot Deck	2.931 ± 0.406	2.956 ± 0.386	2.971 ± 0.378	2.984 ± 0.373
	NOCB	2.940 ± 0.410	2.962 ± 0.393	2.979 ± 0.385	2.996 ± 0.384
	Mean	2.922 ± 0.405	2.949 ± 0.396	2.955 ± 0.396	2.984 ± 0.386
	Median	2.920 ± 0.402	2.950 ± 0.403	2.956 ± 0.397	2.969 ± 0.379
	PMM	2.932 ± 0.409	2.923 ± 0.413	2.944 ± 0.434	2.953 ± 0.424
	Mode	2.922 ± 0.405	2.949 ± 0.396	2.955 ± 0.396	2.984 ± 0.386
	Regression	2.928 ± 0.406	2.925 ± 0.405	2.932 ± 0.388	2.965 ± 0.372
	LOCF	2.947 ± 0.402	2.966 ± 0.398	2.965 ± 0.367	2.994 ± 0.365
High Correlated	**Reference**	2.560 ± 0.413	2.560 ± 0.413	2.560 ± 0.413	2.560 ± 0.413
	Hot Deck	2.641 ± 0.414	2.626 ± 0.422	2.602 ± 0.420	2.604 ± 0.421
	NOCB	2.626 ± 0.428	2.625 ± 0.419	2.610 ± 0.408	2.605 ± 0.416
	Mean	2.647 ± 0.438	2.627 ± 0.412	2.610 ± 0.414	2.621 ± 0.405
	Median	2.645 ± 0.430	2.612 ± 0.414	2.616 ± 0.407	2.622 ± 0.385
	PMM	2.548 ± 0.416	2.574 ± 0.434	2.569 ± 0.411	2.594 ± 0.423
	Mode	2.647 ± 0.438	2.627 ± 0.412	2.610 ± 0.414	2.621 ± 0.405
	Regression	2.550 ± 0.414	2.539 ± 0.411	2.539 ± 0.400	2.560 ± 0.382
	LOCF	2.638 ± 0.424	2.615 ± 0.419	2.601 ± 0.425	2.595 ± 0.391

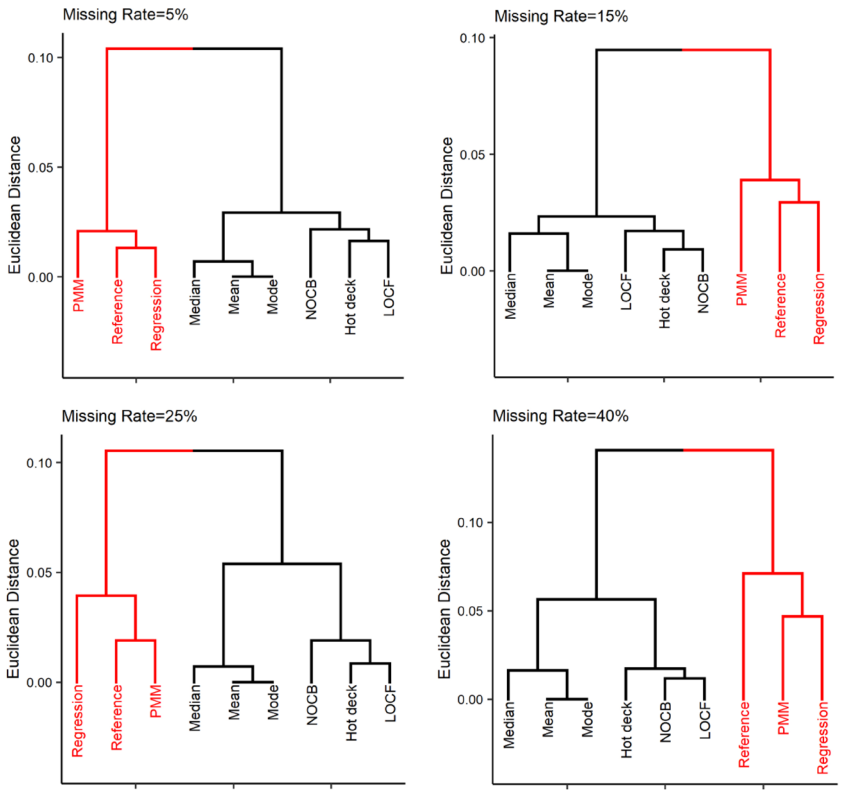

Figure 1

Dendrograms of ATT scores according to missing rates.

As a general inference, it can be explained that the methods were basically divided into 2 clusters. Regression and PMM methods tended to cluster around the reference category for first cluster (red line). Other methods that form the second cluster (black line) were also clustered within themselves and these were further away from the reference. Second cluster divided into two groups also. First group contained mean, median and mode methods, second one contained hot deck, NOCB and LOCF methods. For 5% and 15% missing rate, regression method was closer to reference than PMM method. As the rate of missing increased, the PMM method got closer to the reference category than the regression method (Figure 1).

As shown in Figure 2, the methods were clustered similar to inference that was made according the missing rate. There were clearly two very distinct groups. First group (red line) contained reference, regression

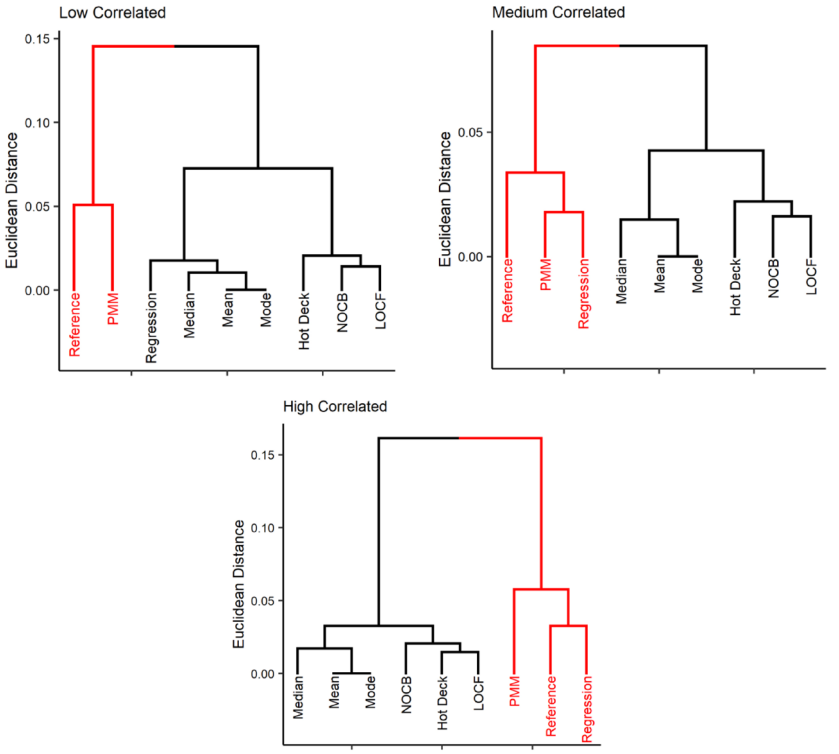

Figure 2

Dendrograms of ATT scores according to correlation structures.

and PMM methods, while the other one (black line) contained remaining methods. When the correlation was low, the PMM method was closer to the reference than the regression method. However, as the correlation relationship increased, the regression model became closer to the reference.

4. Discussion

Although many techniques are used for imputation of missing values, it is difficult to decide which technique should be preferred and in which case. The optimal strategy may vary depending on the missing rate and data structure [6, 26]. Therefore, in this study, it was investigated how different imputation methods dealing with missing values with MAR mechanism in a PS analysis might affect ATT estimates and how these estimates change according to the correlation structure of covariates and missing rates.

On Result of examination of literature, it is seen that some studies to deal with missing data are noteworthy. Mitra and Reiter (2016) reported that the choice of imputation model affected treatment effect estimates for PS analyses with missing values. Mayer and Puschner (2015) conducted a simulation study with different missing rates (5%, 20%, 50%) and MAR mechanism based on an observational study data. Their analysis model that comprised a PS-adjusted treatment effect estimation and missing values were evaluated using different imputation approaches, as well as mean and regression imputation. The regression imputation approach showed slightly better results than other methods. But they didn't suggest a general superiority of any method in the considered analysis model. However, they reported that all point estimators of the applied methods fell within the 95% CI of the original treatment effect, which arose from the complete data set. Jochen et al. (2013) conducted a simulation study and reported that the more missing data were introduced, the lower the precision of the estimates became. They recommended substituting no more than 40% of the missing data in studies consisting of $n = 200$ observations. We also applied for our study for $n = 200$ units and set the missing data rate to be 5%, 15%, 25% and 40%. Jadhav et al. (2019) reported that the performance of data imputation methods independent of the dataset and percentage of missing values in the dataset. Liu et al. (2020) conducted a study on the effect of missing data rate on classification success. He has found that the higher the missing data rate, the lower the correct classification rate. Mattei (2009) compared the PMM method with two different methods commonly used. They suggested that the PMM approach appeared perform the best among the three alternatives. For our study, as the missing rate increased, the performance of the imputation methods decreased and as the correlation between covariates increased, the performance evaluation ratings of the methods changed. Besides PMM method performed pretty well.

In our study showed which methods can be used to handle missing values in covariates of a PS model and it is suggested to choose the most suitable imputation method according to missing rates and correlation structures of the dataset. It was focused on this study missing values in covariates, because while in routinely collecting data on patient characteristics, missingness of recorded data more generally occurs.

5. Conclusion

This study has special importance because it contains multiple features in comparison to the health field. Ignoring missing values on covariates for PS analyses causes information loss significantly and this

information loss becomes greater as the rate of missing data increases. PS analyses might be biased if missing data on covariates are ignored also. To prevent this information loss and bias, PS analyses should be performed after solving the problem of missing data with MAR mechanism on covariates by regression and PMM methods, which showed statistical superiority compared to other methods in this study.

Additional

Presented as an oral presentation at the "2nd International Applied Statistics Conference (UYIK 2021) Tokat / Turkey," on June 29 - July 02, 2021.

References

[1] Akmam, E.F., et al. Multiple Imputation with Predictive Mean Matching Method for Numerical Missing Data. in 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS). 2019. IEEE.

[2] Austin, P.C., An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research, 2011. 46(3): p. 399-424.

[3] Caliendo, M. and S. Kopeinig, Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 2008. 22(1): p. 31-72.

[4] Cham, H. and S.G. West, Propensity score analysis with missing data. Psychological methods, 2016. 21(3): p. 427-445.

[5] Choi, J., O.M. Dekkers, and S. le Cessie, A comparison of different methods to handle missing data in the context of propensity score analysis. *European Journal of Epidemiology*, 2019. 34(1): p. 23-36.

[6] Coffman, D.L., J. Zhou, and X. Cai, Comparison of methods for handling covariate missingness in propensity score estimation with a binary exposure. *BMC medical research methodology*, 2020. 20(1): p. 1-14.

[7] D'Agostino Jr, R.B., Propensity score methods for bias reduction in the comparison of a treatment to a non randomized control group. *Statistics in medicine*, 1998. 17(19): p. 2265-2281.

[8] D'Agostino Jr, R.B. and D.B. Rubin, Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, 2000. 95(451): p. 749-759.

[9] Engels, J.M. and P. Diehr, Imputation of missing longitudinal data: a comparison of methods. *Journal of Clinical Epidemiology*, 2003. 56(10): p. 968-976.

[10] Hosmer Jr, D.W., S. Lemeshow, and R.X. Sturdivant, *Applied logistic regression. 3rd ed*ition, Vol. 398. 2013: John Wiley & Sons , Hoboken, NJ.

[11] Jacovidis, J.N., Evaluating the performance of propensity score matching methods: A simulation study. 2017, Doctoral dissertation, James Madison University.

[12] Jadhav, A., D. Pramod, and K. Ramanathan, Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 2019. 33(10): p. 913-933.

[13] Jochen, H., et al., Multiple imputation of missing data: a simulation study on a binary response. *Open Journal of Statistics*, 2013. 2013.

[14] Kalton, G. and D. Kasprzyk. Imputing for missing survey responses. in Proceedings of the section on survey research methods, *American Statistical Association*. 1982. American Statistical Association Cincinnati.

[15] Klamroth-Marganska, V., et al., Three-dimensional, task-specific robot therapy of the arm after stroke: a multicentre, parallel-group randomised trial. *The Lancet Neurology*, 2014. 13(2): p. 159-166.

[16] Little, R.J. and D.B. Rubin, *Statistical analysis with missing data*. *3rd ed*ition,Vol. 793. 2019: John Wiley & Sons, Hoboken, NJ.

[17] Liu, C.-H., et al., The feature selection effect on missing value imputation of medical datasets. *Applied Sciences*, 2020. 10(7): p. 2344.

[18] Mattei, A., Estimating and using propensity score in presence of missing background data: an application to assess the impact of childbearing on wellbeing. *Statistical Methods and Applications*, 2009. 18(2): p. 257-273.

[19] Mayer, B. and B. Puschner, Propensity score adjustment of a treatment effect with missing data in psychiatric health services research. *Epidemiol Biostat Public Health*, 2015. 12(1): p. 10.2427.

[20] Meena, N., Singh, B., Firefly optimization based hierarchical clustering algorithm in wireless sensor network. *Journal of Discrete Mathematical Sciences and Cryptography*, 2021. 24(6), p. 1717-1725.

[21] Misztal, M., Imputation of missing data using R package. *Acta Universitatis Lodziensis. Folia Oecologica*, 2012, 269: p. 131-144.

[22] Mitra, R. and J.P. Reiter, Estimating propensity scores with missing covariate data using general location mixture models. *Statistics in medicine*, 2011. 30(6): p. 627-641.

[23] Mitra, R. and J.P. Reiter, A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical methods in medical research*, 2016. 25(1): p. 188-204.

[24] Molnar, F.J., B. Hutton, and D. Fergusson, Does analysis using "last observation carried forward" introduce bias in dementia research? Cmaj, 2008. 179(8): p. 751-753.

[25] Morris, T.P., I.R. White, and P. Royston, Tuning multiple imputation by predictive mean matching and local residual draws. *BMC medical research methodology*, 2014. 14(1): p. 1-13.

[26] Qu, Y. and I. Lipkovich, Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statistics in medicine*, 2009. 28(9): p. 1402-1414.

[27] Rosenbaum, P.R. and D.B. Rubin, Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 1985. 39(1): p. 33-38.

[28] Rubin, D.B., Inference and missing data. Biometrika, 1976. 63(3): p. 581-592.

[29] Vink, G., et al., Predictive mean matching imputation of semicontinuous variables. Statistica Neerlandica, 2014. 68(1): p. 61-90.

[30] Zhang, J., A Comparison of Propensity Score Matching Methods in R with the MatchIt Package: A Simulation Study. 2013, University of Cincinnati, Master's thesis. OhioLINK Electronic Theses and Dissertations Center.

Received October, 2021

Revised May, 2022