Comparing Multiple Imputation and Machine Learning Techniques for Longitudinal Data

[1]Sanjana Rajamani
Department of Statistics and Data Science
Christ (Deemed to be University)Bangalore, India

[2]Seena Thomas
Department of Statistics and Data Science
Christ (Deemed to be University) Bangalore, India

Abstract:- It is an essential part of research to find ways to impute the missing values in a data set. The missingness is unavoidable as it could be due to natural or non-natural reasons. Missing information is inevitable in longitudinal or multilevel studies, and can result in biased estimates, loss of power, variability and inaccuracy in results. For this study a complete data which showed the resistance scores of intellectually disabled children on giving behavioral skill training was considered in order to compare the various imputation techniques. The secondary data collected was longitudinal in nature. The resistance score was noted before the training and at four different time points after the training. A random missingness was created under varying percentages in the complete data (5%, 10%, 15%, 20%, 30%) using the MAR mechanism. The obtained values after imputation were compared with full data using a linear mixed model. Various models built under the multiple imputation and machine learning techniques for imputing different features which are used to predict the resistance score, using the coefficients taken from the real data and the same mechanism was implemented for simulated data as well. The methods based on machine learning techniques were the most suited for the imputation of missing values and led to a significant enhancement of prognosis accuracy when compared to multiple imputation techniques using linear mixed models.

Keywords:- *Multiple Imputation, MAR Mechanisms, Machine Learning Techniques, Linear Mixed Effect Model.*

I. INTRODUCTION

This Dealing with the incidence of missing data remains a concern when using real world data to supplement Clinical Trials [1]. It is not always practical to remove a record from your data in order to make predictions. In this study a lot of variation in the results as a consequence of missing data, and this can also lead to bias, which could further lead to flawed and inaccurate results of the study. By obtaining these values using a variety of methods, the missing values in the data may be replaced. A study design that includes observing the same variables repeatedly throughout a range of time-frames is called a longitudinal data. Since longitudinal studies [2] are frequently employed in clinical trials and social- personality psychology to examine quick changes in actions, thoughts, and emotions from one moment to the next or day to the next, they are frequently referred to as a type of observational studies. These missing [3] values were imputed using machine learning and several types of multiple imputation techniques. Using linear mixed model LMM [4], a comparison of how accurate these values were, was studied.

A. Missing Data

Longitudinal data [5] frequently contain missing data. This is particularly true in long-term biomedical investigations of people since it is hard to guarantee a 100% protocol compliance. A linear mixed model can take into account missing values and build a model to give out results and help predict any variable unlike other models which require a complete data set to fit a model and predict the estimates, although the validity of the parameters we want to estimate has to be of a certain type. Many a times the underlying reason as to why there is an occurrence of missing data might not be known. In order to find these values, the reason or cause has to be known well in advance. In Missing [3] data literature mainly there are three distinct mechanisms how the data could be possibly missing.

- *Missing at Random (MAR)*

This indicates that there is a systematic link between the probability of missing values and the observed data, that is the missing value can be anticipated using the other attributes in the data set.

- *Missing Completely at Random (MCAR)*

In contrast to MAR, this form of missingness [6] shows that there is no connection between the missing value and the other features in the data set.

Since there is no logic involved to understand why there is missingness of a certain value, this form of missing values is the easiest to comprehend

- *Missing not at Random (MNAR)*

The hardest data to discover and work with both in terms of finding and using is MNAR data. The missingness and lack of data are caused by factorsthat we ignored to take into consideration.

All the real-world data sets have missing values and this definitely has to be taken care of. There are several approaches to handling missing data since the outcomes we aim to get might be significantly impacted by these missing values [7]. Data recording errors, damaged raw data columns, and human error are some of the common reasons of missing data. Missing data poses some significant obstacles for the study's findings. if important information is missed, it is difficult to conclude any results [8]. Hence, study of several tactics is necessary to ascertain which approach will result in adequate outcomes and a sound model [9].

B. *Imputation*

Missing data may add a significant amount of bias, make processing and analyzing the data more difficult, and reduce efficiency. That is, omissions (introduced in collection or processing) may result in certain sub-groups of the target population from being excluded in the analysis of the data set, and in turn increasing the risk of biased estimates, reducing the power of inferential statistics and increasing the uncertainty of estimates and inferences derived from the data. By substituting missing data with an estimated value based on other available information, imputation preserves all situations [8]. The data set may be analyzed using methods used for complete data once all missing values have been imputed.

➢ *Multiple Imputation Techniques*

The distribution of each variable with missing values when using the multiple imputation [10] process has to be modeled. It is a generic solution to the issue of missing data that is included in a number of widely used statistical software programs [7]. By constructing many distinct plausible imputed data sets and correctly merging the findings from each one, it seeks to account for the uncertainty around the missing data. Multiple imputation is a method for addressing non-response bias that is based on a Bayesian approach. The steps involved in Multiple Imputation [11]: Missing values are imputed m times (m >1), resulting in m complete data sets. The imputed values are drawn from distributions modelled specifically for each missing entry. The standard suggestion for relative efficiency is 4 - 5 imputations, while a higher number of imputations would give more accurate results. Each of these data sets is analyzed using the statistical model and hence it creates m set of estimates. The m analysis estimates are combined toone Multiple Imputation estimate.

- *Fully Conditional Specification (FCS)*

We refer to this method as FCS Standard since it uses all of the repeated measurements of the time-varying covariate as predictors in each of the univariate imputation models. Due tomodel overfitting / collinearity when thereare a lot of associated repeated measures, this method is vulnerable to convergence issues [10].

- *Fully Conditional Specification, Linear mixed Model (FCS-LMM)*

The cycles over the univariate imputation models, using a multilevel LMM to impute missing values in each incomplete time-dependent variable given all the others instead of considering repeated measurements as separate variables. The linear two-level model with homogeneous within-subject variances, a specific instance of a multivariate LMM [5], is implemented using the Gibbs sampler in this method.

- *Joint Modelling (JM)*

Employing a joint multivariate [12] LMM for imputing multiple incomplete longitudinal variables [11] rather than considering repeated measurements as separate variables is the algorithm [13] behind this technique. In order to account for dependence among people over time, this technique presupposes that all the incomplete variables are continuous with subject specific random effects. Similar to the univariate LMM [14], this approach makes the assumption that measurement errors and random effects have a normal distribution with constant error covariance across all subjects.

➢ *Machine Learning Technique*

A large number of important machine learning methods [15] have emerged since the 1980s and 1990s, such as back propagation neural network and random forest (RF), which had a profound impact on the medical field including clinical decision making in presence of missing data. Before that, the traditional methods used to process the missing data in clinical decision [16] making mainly included complete case analysis, k- nearest neighbors (KNN), expectation maximization and so on. With the in-depth application of machine learning models in this field, researchers found that machine learning models can restore the true distribution of data from missing data sets more accurately than the traditional missing data processing models.

- *K- Nearest Neighbors*

An unclassified sample point is given the classification of the nearest previously classified point via the nearest neighbor judgement process. The sample points classifications and the underlying joint distribution have no bearing on this rule. In contrast, we shall demonstrate in a large sample analysis that for all appropriately [16] smooth underlying distributions, the M category is where these restrictions are the tightest. It may be claimed that the nearest neighbor holds half of the classification information in an infinite sample set.

- *Random Forest*

An effective method of imputation, Random Forest virtually meets all the criteria for becoming the best [15] imputation method. The Random Forests are fairly good at scaling to large data settings, and they can tolerate outliers and non-linearity in the data. Mixed-type data can be held by Random Forests (both numerical and categorical). They additionally offer a built-in feature selection method. Random Forests can easily outperform KNN and other techniques because of its unique advantages.

C. *Data Description*

Behavioral Skills Training (BST) was conducted on intellectually disabled children's knowledge of sexual abuse [17] and capacity for resistance. A pilot research was carried out to look at how children who had experienced sexual abuse in the past and those who had not responded to the personal safety program. Parents or other caregivers trained 60 children between the ages of 3 and 7 who had been subjected to abuse and 60 children who had not. Before and after the intervention, the children were evaluated using the Personal Safety Questionnaire (PSQ). A six-month at home behavioral skills training program on sexual abuse prevention was offered [18]. The findings indicated that both groups had increased knowledge and expertise in sexual abuse prevention. After participating in the program, children who had a history of sexual abuse significantly reduced their improper sexual behavior [21].

II. METHODOLOGY

- A complete data set was taken, with the resistance score as well as several features that significantly influence the prediction of the score was considered, and it was then constructed with various percentages (5%, 10%, 15%, 20%) of missingness.
- Given is the plot depicting where missingness in the data was created. No missingness was created in the resistance column, since it is the predictor variable.

- By considering the technique's relevance to the longitudinal data, a comparison was done with different multiple (FCS, FCS-LMM, and JM [20]) and machine learning imputation techniques (KNN and Random Forest). These methods were all programmed using the RStudio program.

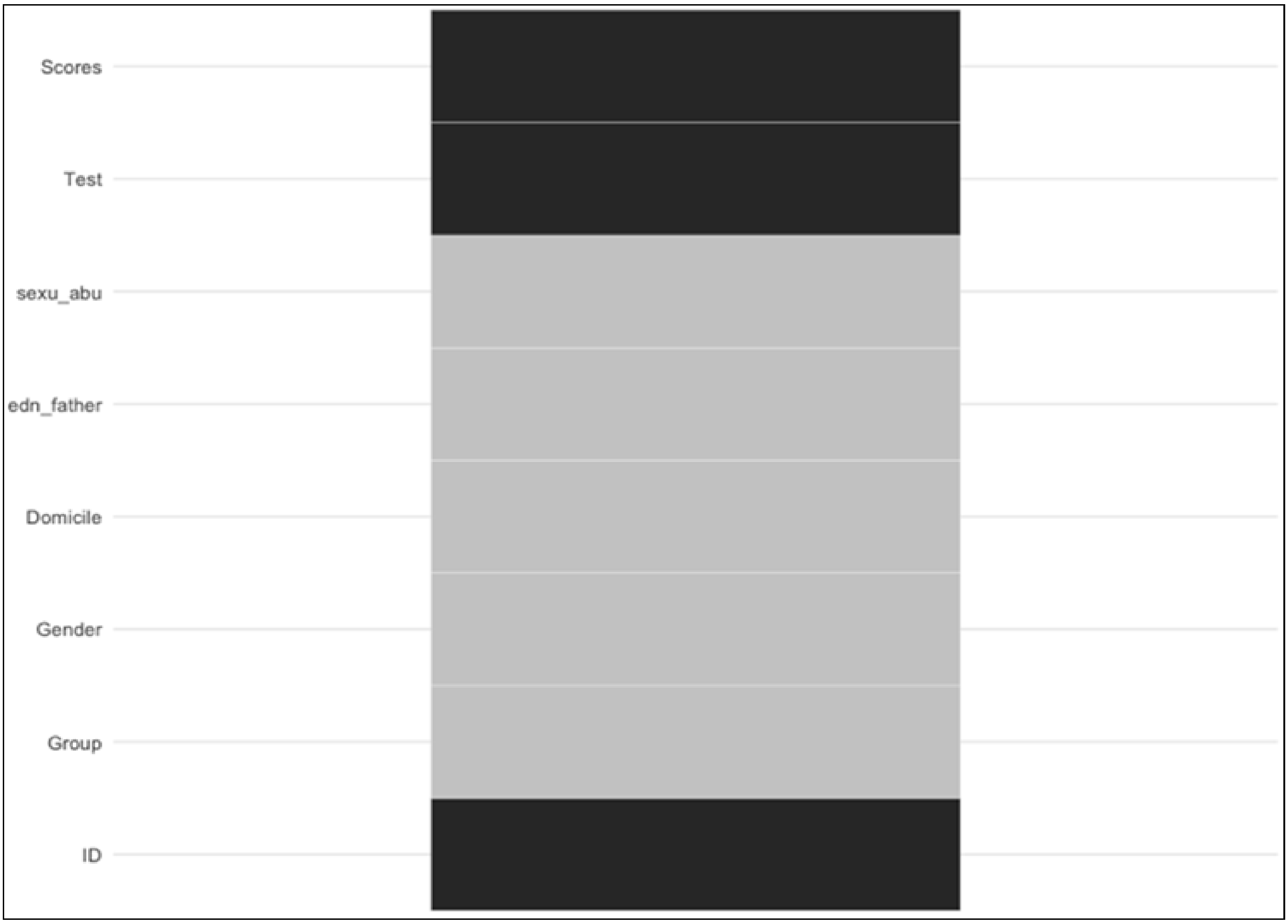
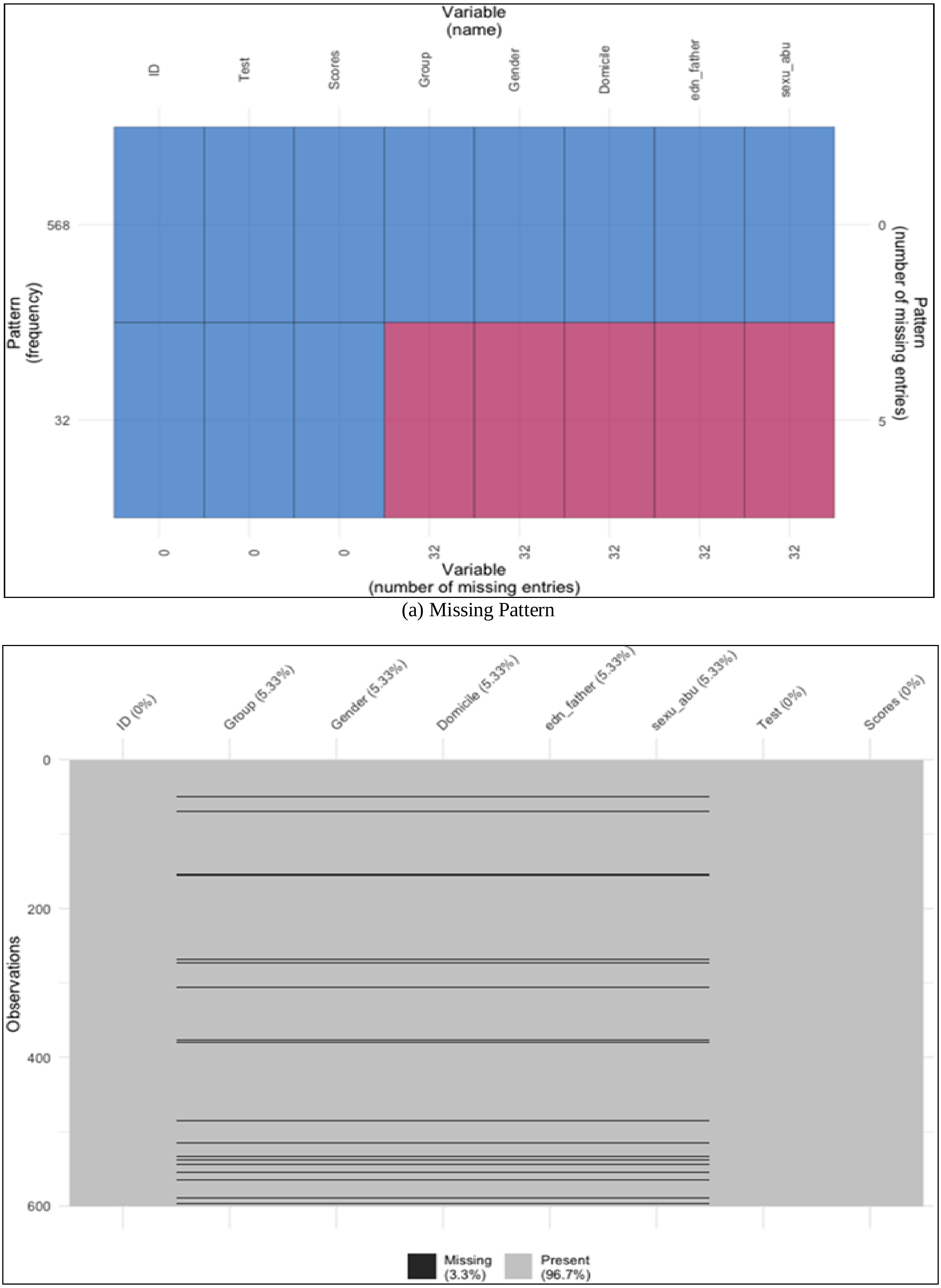


Fig 1 Variables with Missing Values

- Based on the estimates that the various models using the imputed methodologies provided, the various techniques were compared using linear mixed modelling. These findings were tabulated in order to determine which of these techniques generated [9] the most accurate outcomes compatible with the original model we developedfor the entire data set.
- Combining together all these predicted values for all missingness percentages to check whether the techniques can produce accurate estimates for all levels of missingness.
- A conclusion on which method provides a better estimate for the missing values by comparing it to the original values was computed.

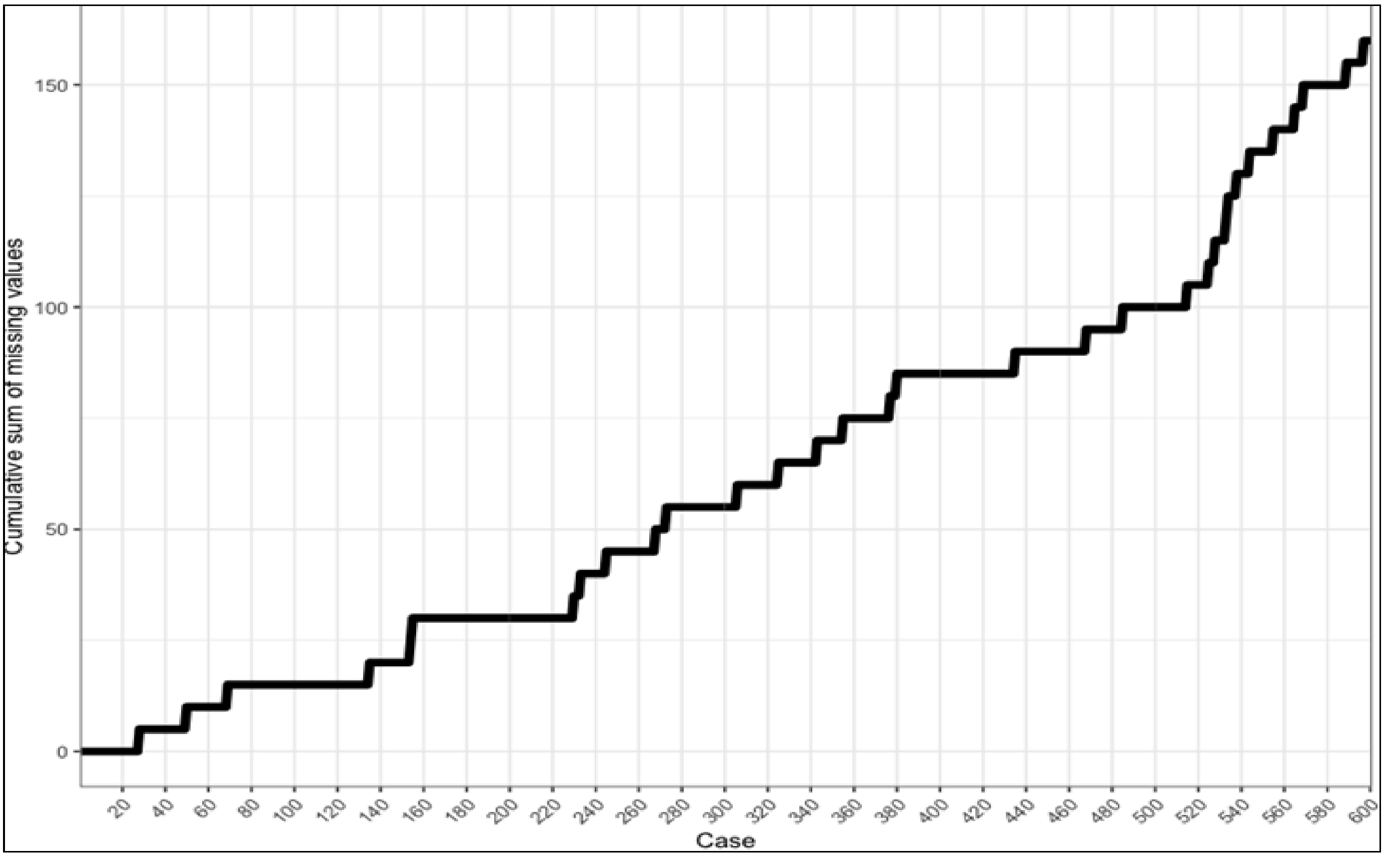
III. RESULTS



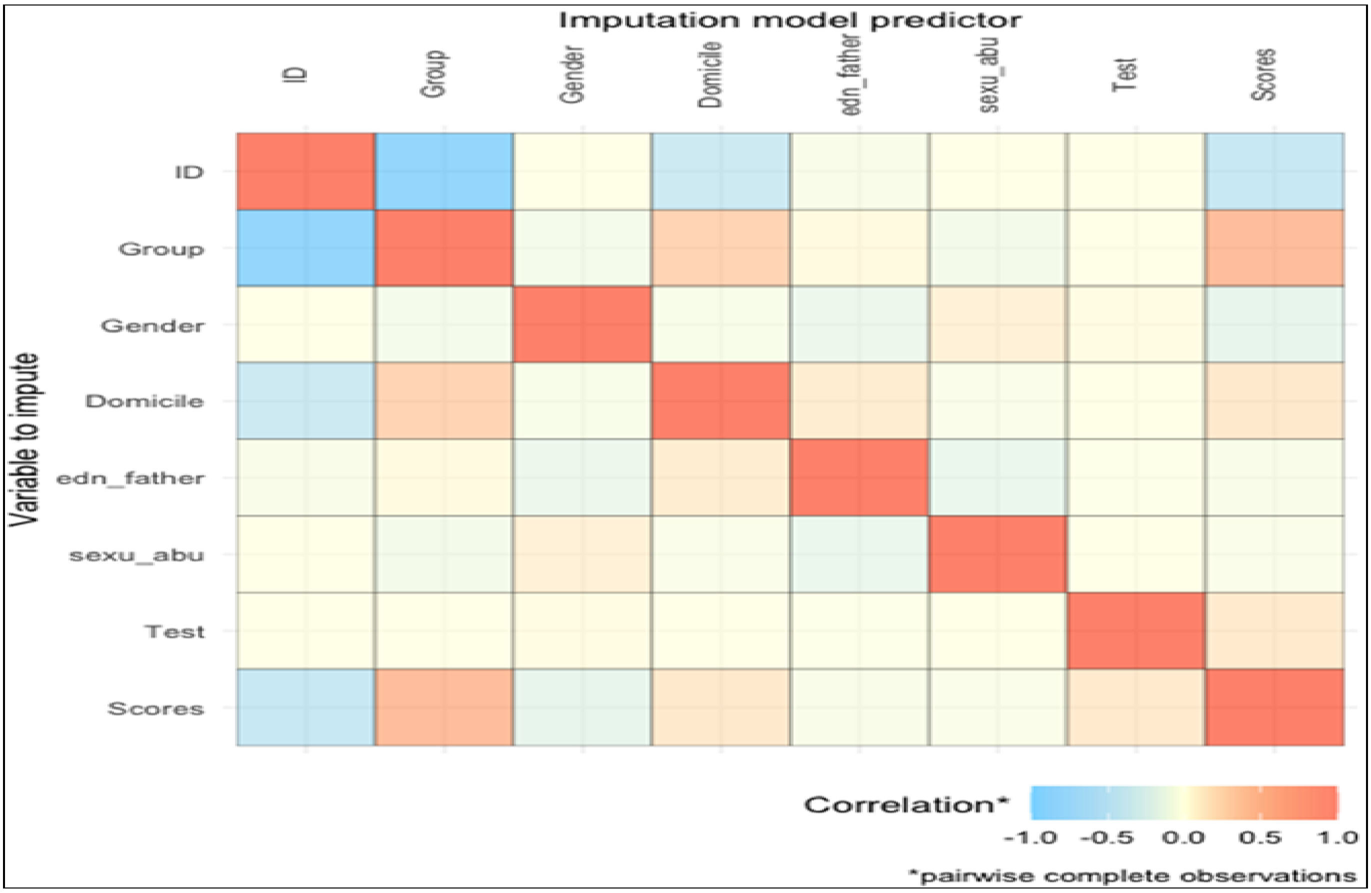
(a) Missing Pattern

(b) Percentage of Missingness in Each Variable

Fig 2 The pattern of missingness in the data where 5%missingness was generated and visualization of the amount of missing data. Showing in black is the location of missing values, and also providing information on the overall percentage of missing values in each variable.

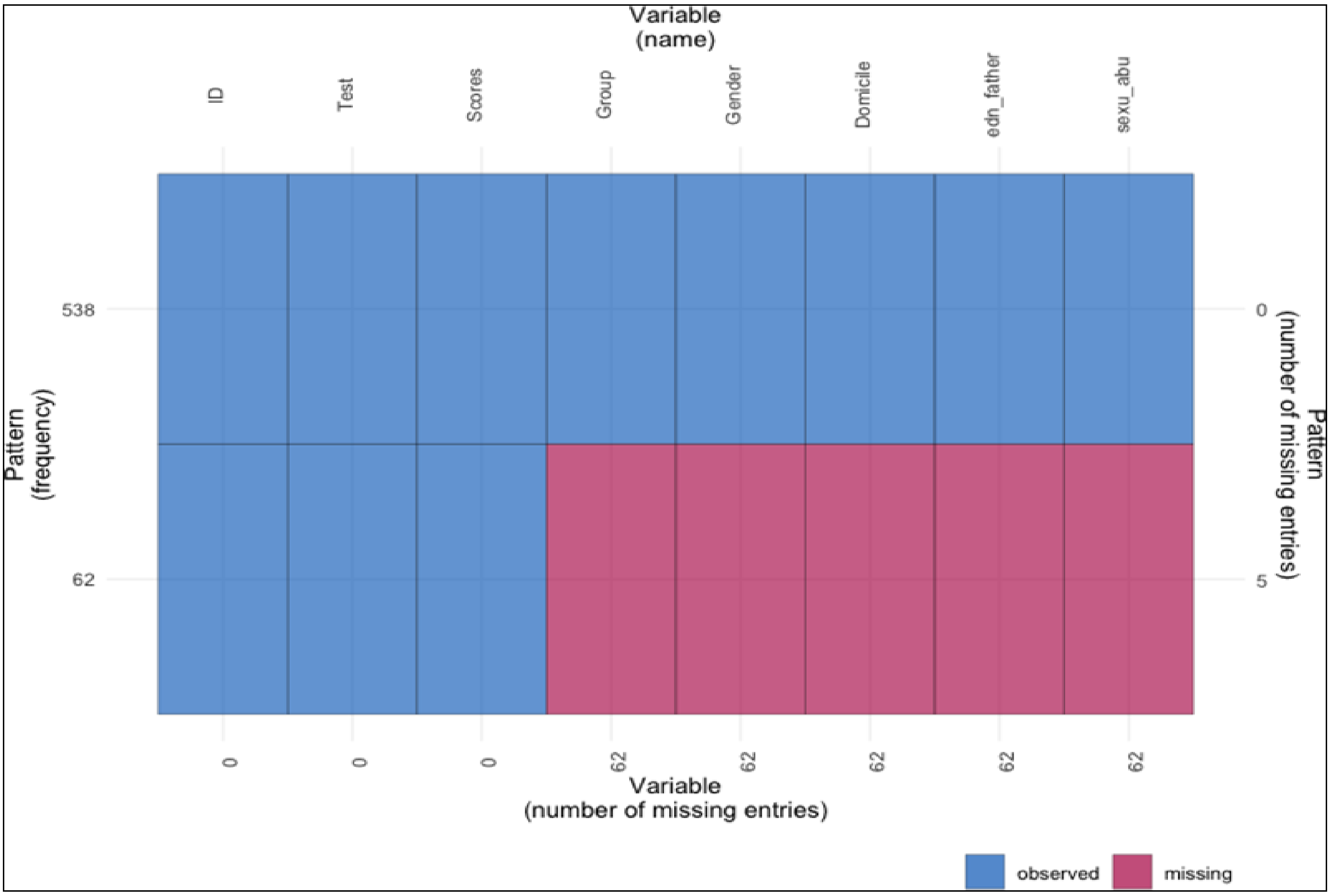


(a) Cumulative Sum of Missing Values

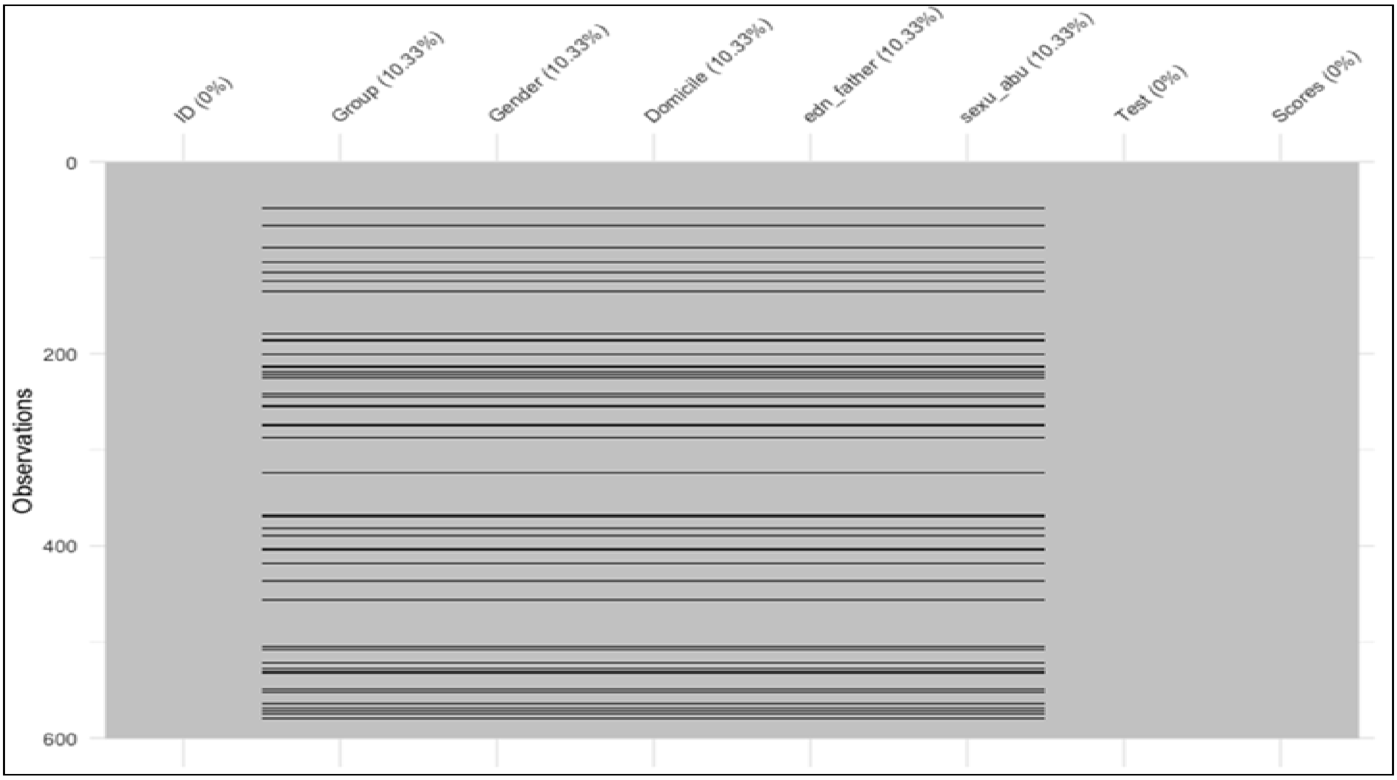


(b) Correlation Plot

Fig 3 The plot shows the cumulative sum of missing values, reading the rows of the data set from the top to bottom and the correlation between the variables after generating 5% missingness in the data set.

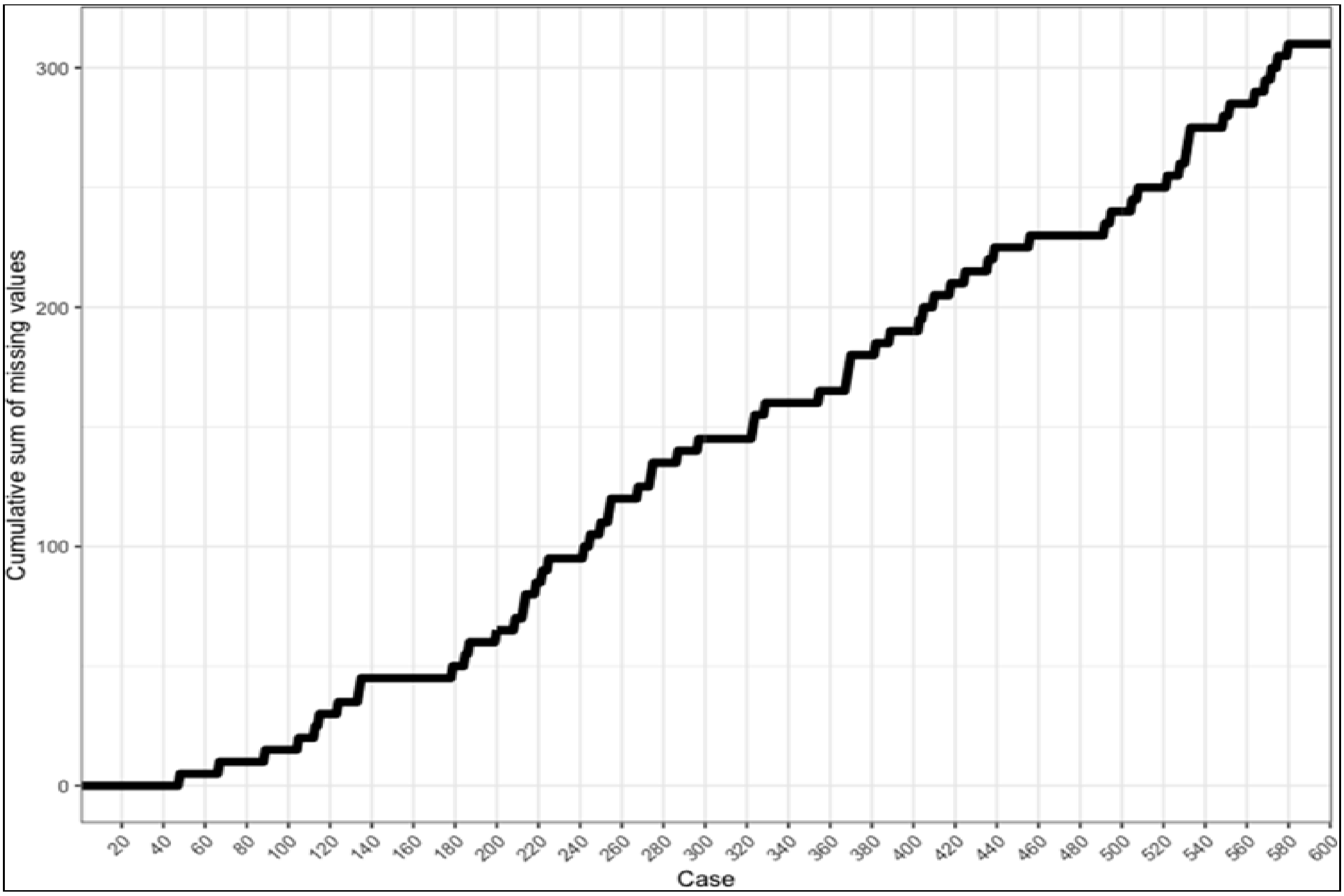


(a) Missing Pattern

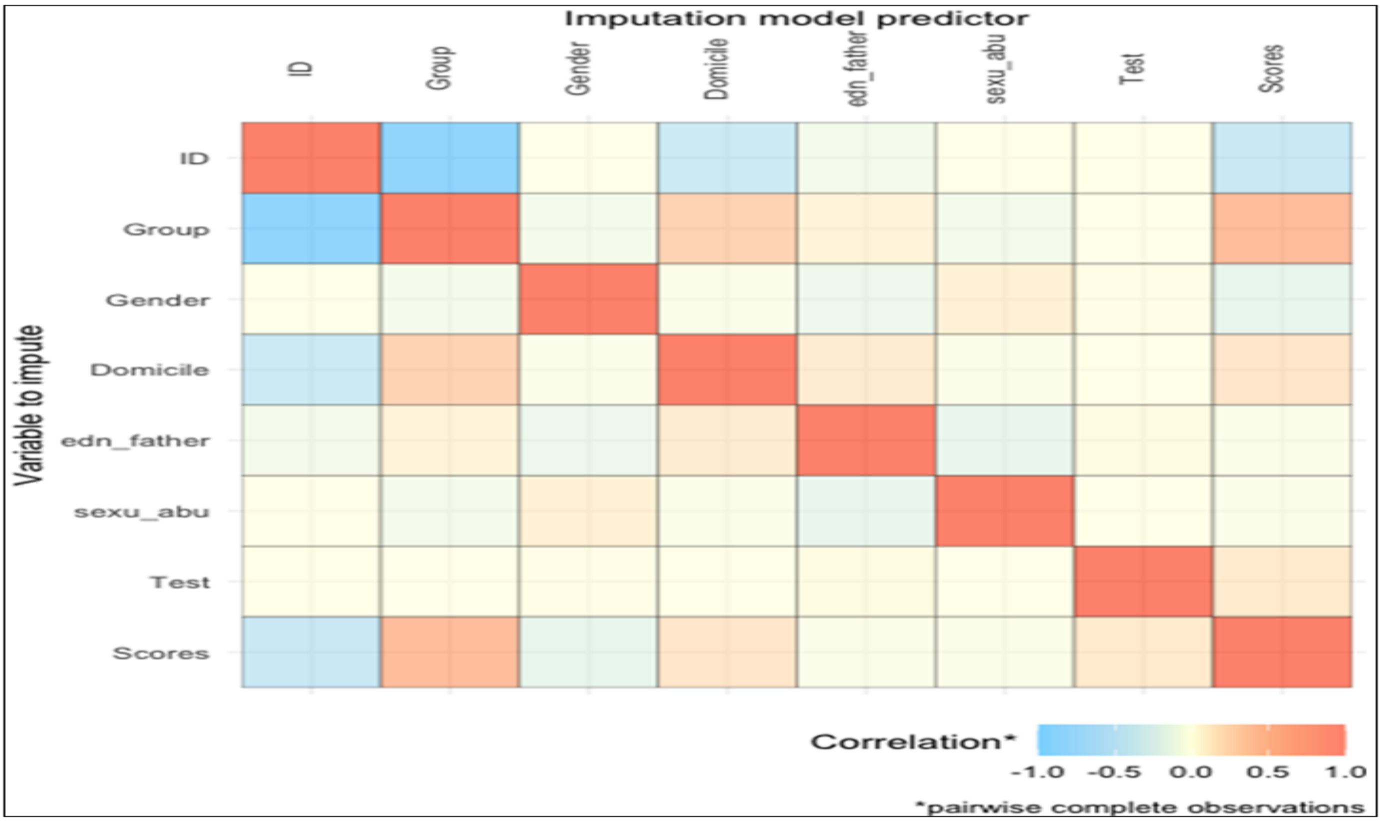


(b) Percentage of Missingness in Each Variable

Fig 4 The pattern of missingness in the data where 10% missingness was generated and visualization of theamount of missing data, showing in black the location ofmissing values, and also providing information on the overall percentage of missing values and in each variable, are shown above.

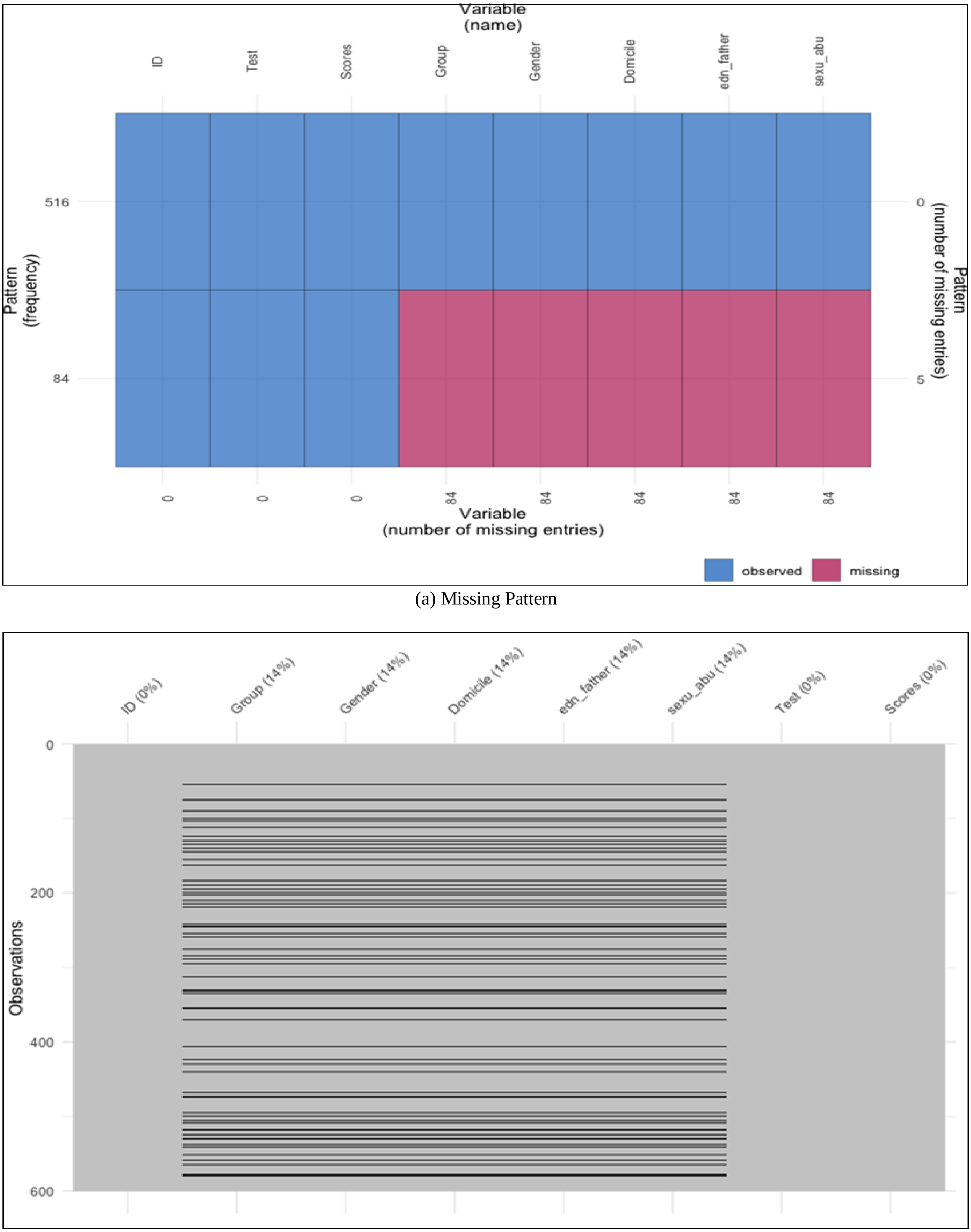


(a) Cumulative Sum of Missing Values



(b) Correlation Plot

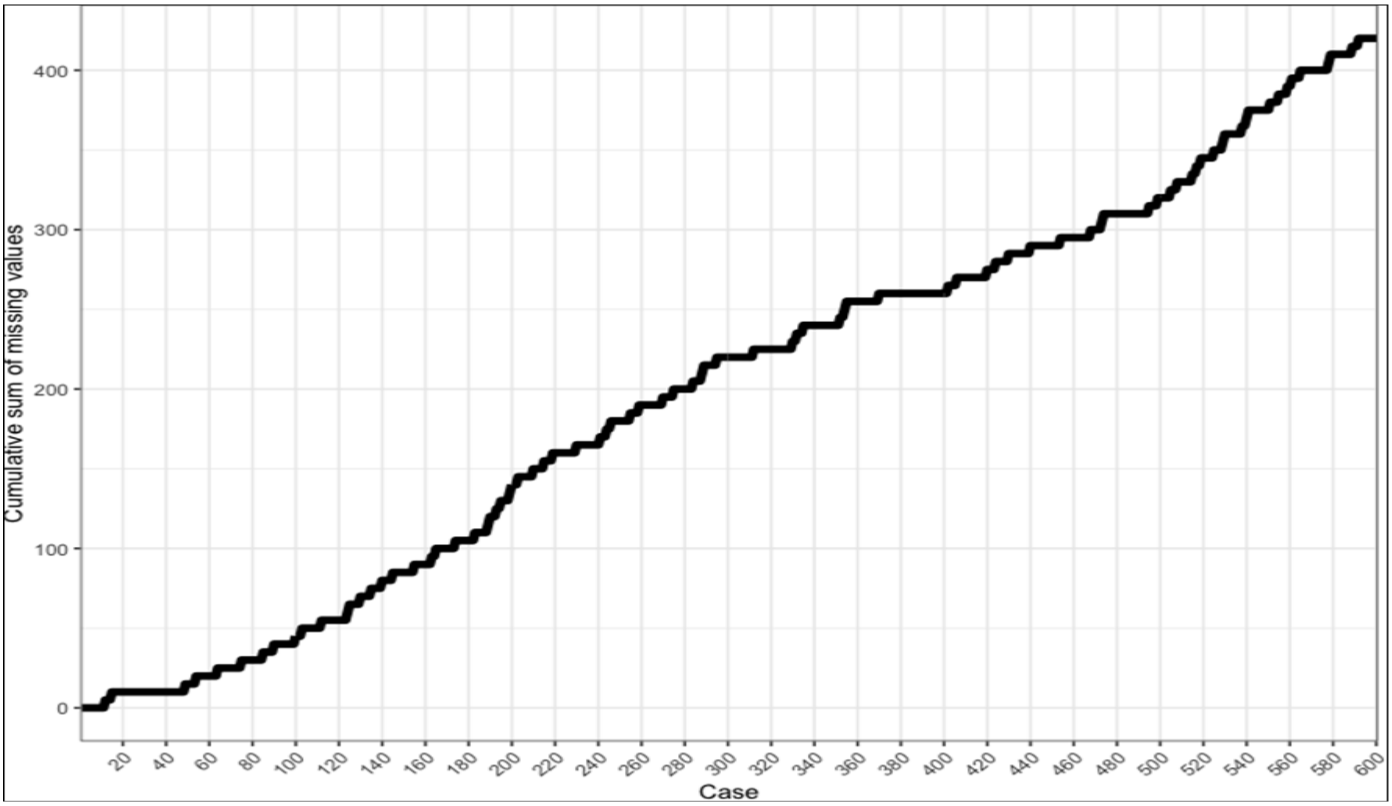
Fig 5 The plots show the cumulative sum of missing values, reading the rows of the data set from thetop to bottom and the correlation between the variables after generating 10% missingness in the data set.



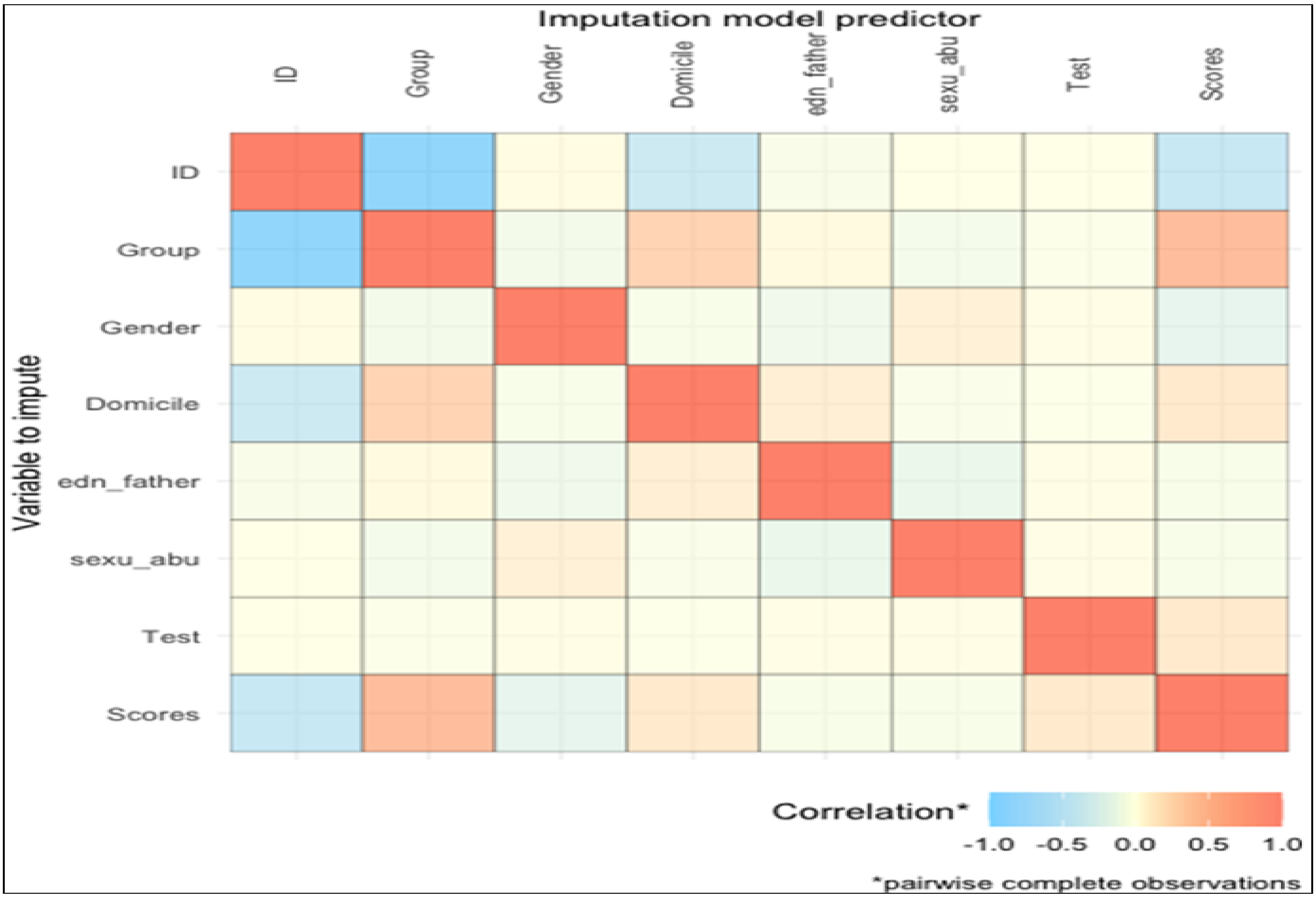
(a) Missing Pattern

(b) Percentage of Missingness in Each Variable

Fig 6 The pattern of missingness in the data where15% missingness was generated and visualization of theamount of missing data, showing in black the location ofmissing values, and also providing information on the overall percentage of missing values overall and in eachvariable, are shown above.

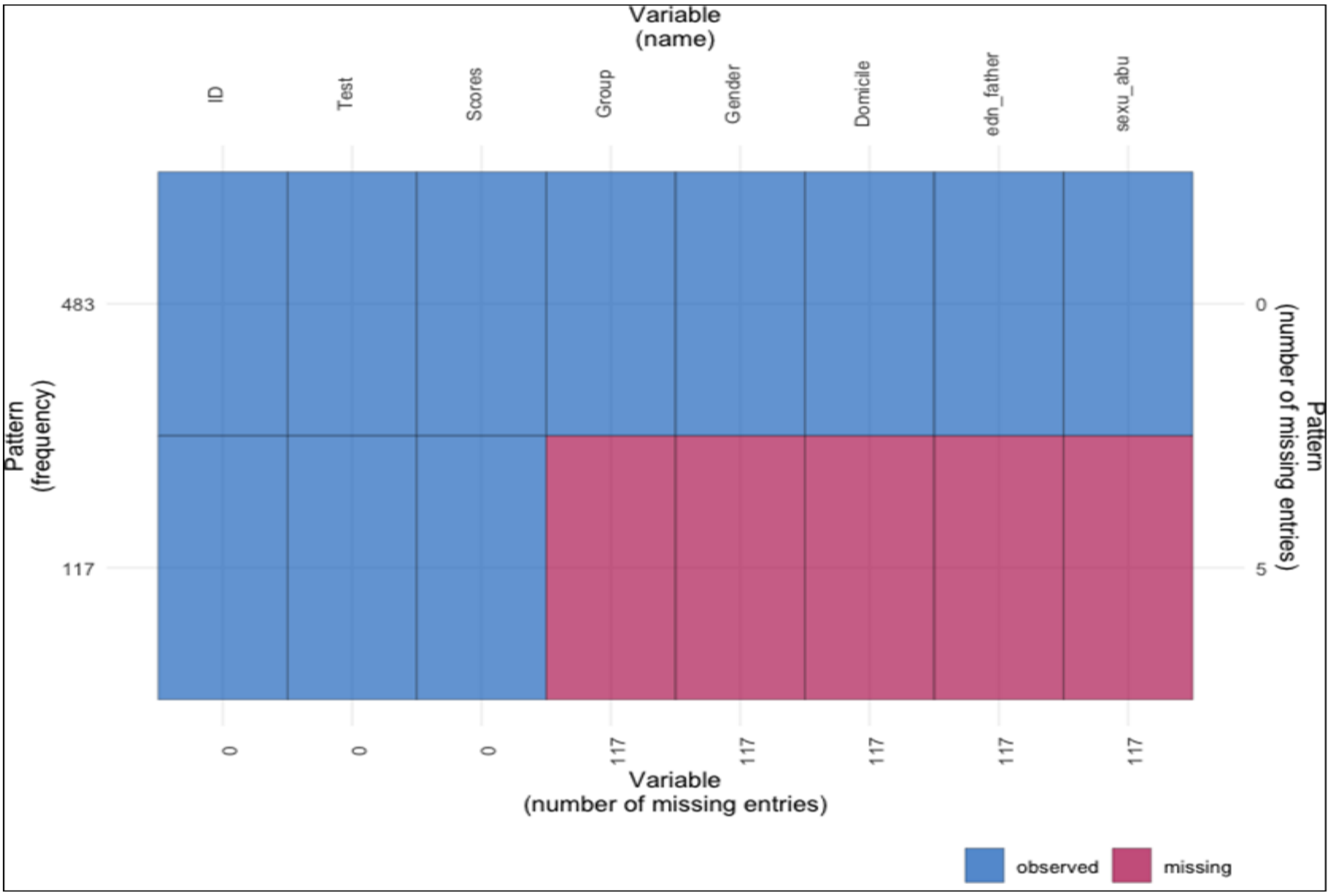


(a) Cumulative sum of missing values

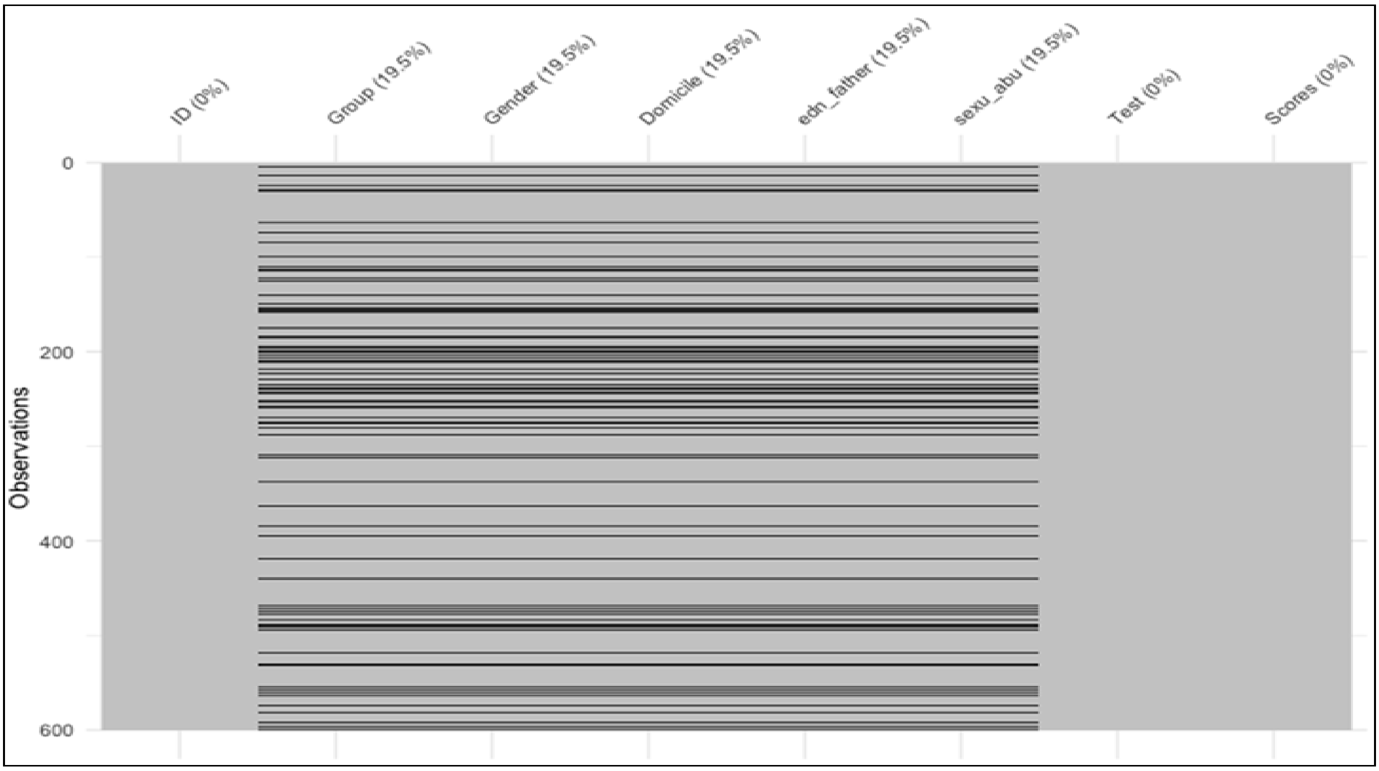


(b) Correlation Plot

Fig 7 The above plots show the cumulative sum of missing values, reading the rows of the data set from the top to bottom and the correlation between the variables after generating 15% missingness in the data set.

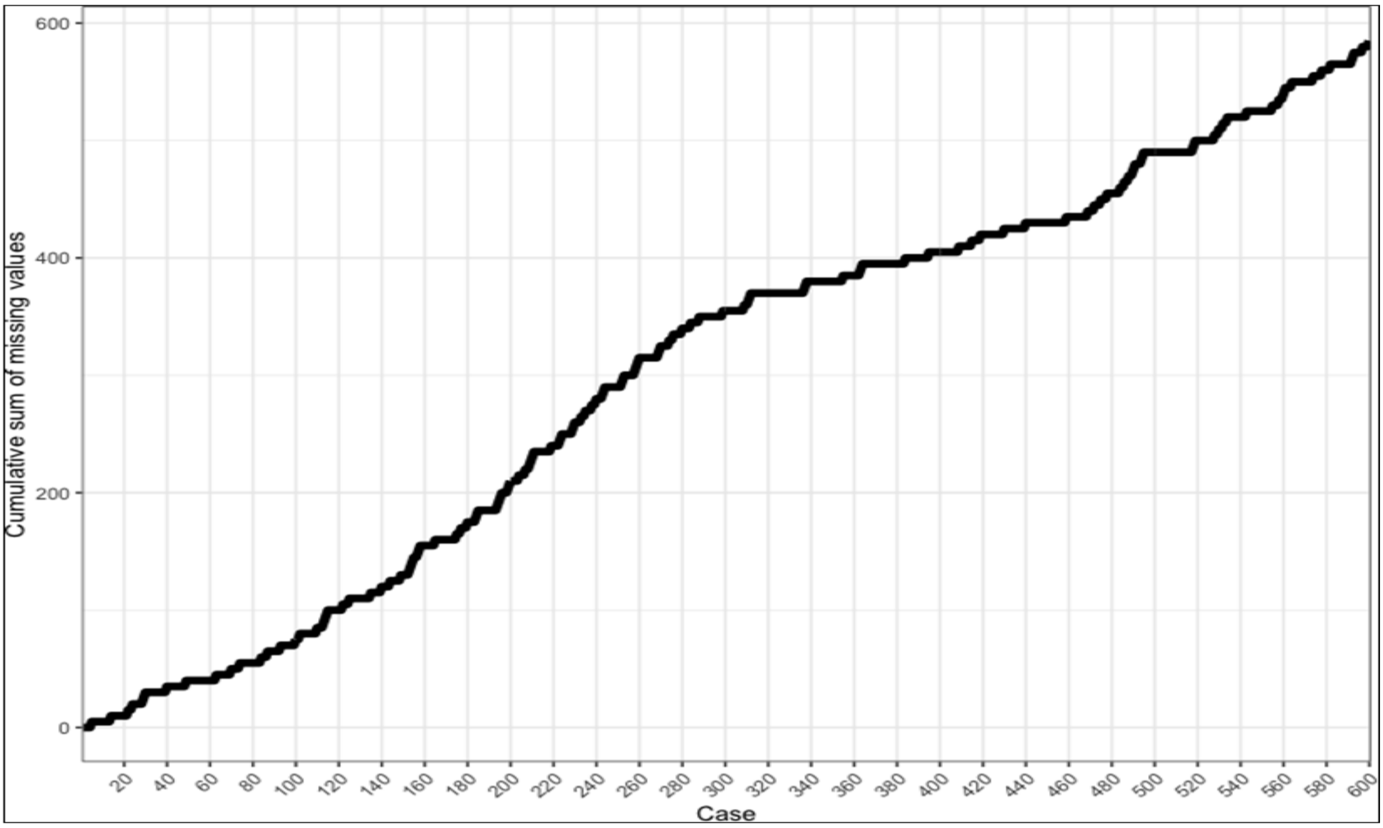


(a) Missing Pattern

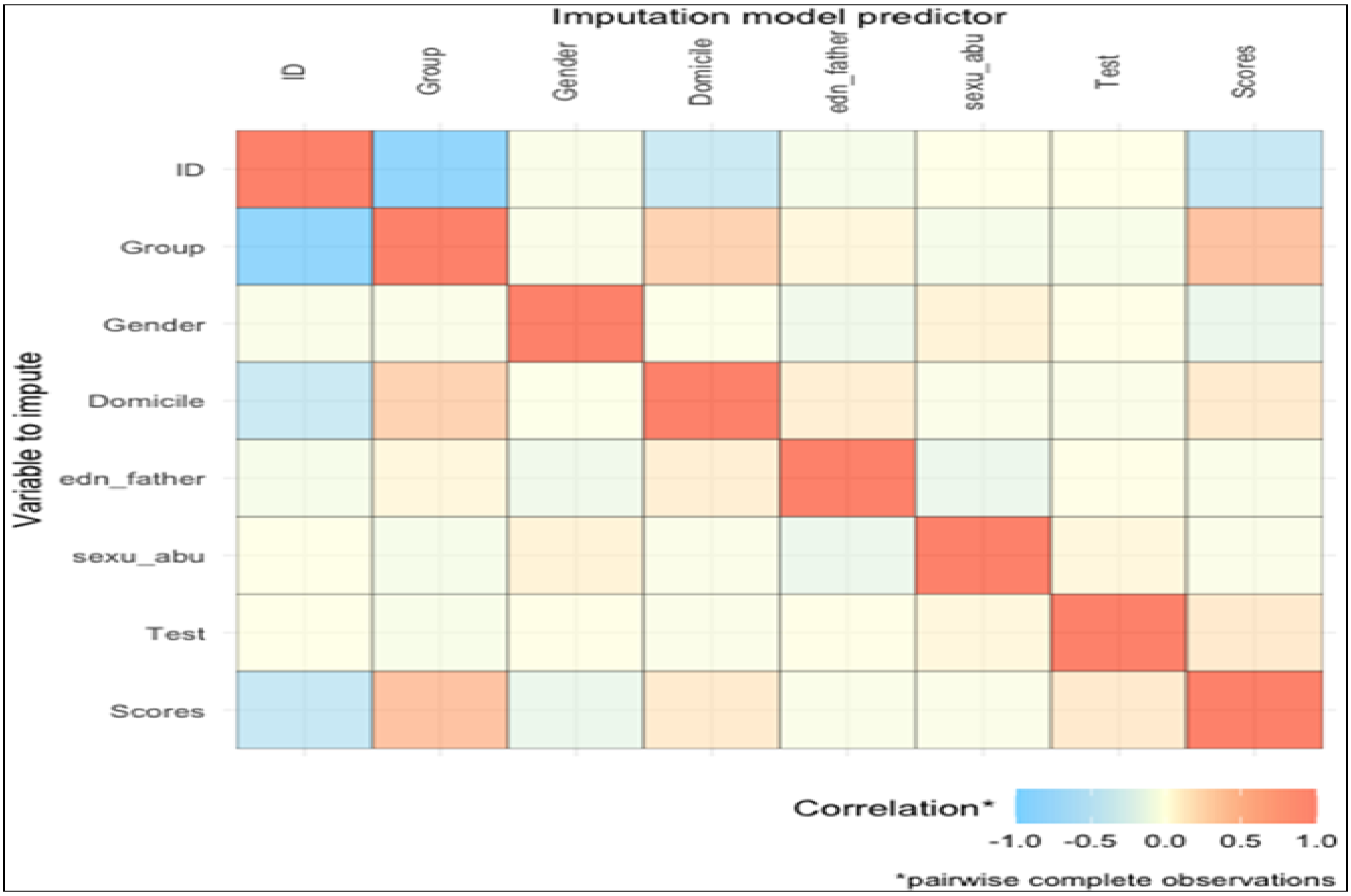


(b) Percentage of Missingness in Each Variable

Fig 8 The pattern of missingness in the data where 20% missingness was generated and visualization of theamount of missing data, showing in black the location ofmissing values, and also providing information on the overall percentage of missing values overall and in eachvariable are shown above.



(a) Cumulative Sum of Missing Values



(b) Correlation Plot

Fig 9 The above plots show the cumulative sum of missing values, reading the rows of the data set from thetop to bottom and the correlation between the variables after generating 20% missingness in the data set.

Table 1 The estimated coefficients obtained using LMM for different Multiple and Machine Learning Imputation techniques at different percentage (5%, 10%, 15%, 20%) of missingness are displayed in the table (1a, 2b, 3c, 4d). Only p values which are significant have been added to linear mixed model.

Percentage of Missingness created	Number of observations missing (out of 600)	Percentage of missingness reflected in the data
5	32	5.33
10	62	10.33
15	84	15
20	117	19.33

Table 1a 5 %

Variables	Group	Gender	Domicile	Education	Assault
Original	8.61419	-0.72853	0.01083	-0.18508	-0.0273
LWD	8.0000	-0.72853	0.01083	-0.18508	-0.0273
MI	8.521	-0.71	0.012	-0.0182	-0.026
KNN	8.612	-0.723	0.01	-0.183	-0.0269
RF	8.615	-0.726	0.01083	-0.184	-0.0265
HD	8.2	-0.61	0.011	-0.122	-0.026
JM	8.59	-0.727	0.013	-0.186	-0.028
FCS	8.59	-0.727	0.013	-0.186	-0.028
FCS-LMM	-8.54	0.71	0.0122	-0.182	-0.0271

Table 1b 10 %

Variables	Group	Gender	Domicile	Education	Assault
Original	8.61419	-0.72853	0.01083	-0.18508	-0.0273
LWD	8.61419	-0.72853	0.01083	-0.18508	-0.0273
MI	8.62	-0.73	0.011	-0.187	-0.026
KNN	8.613	-0.728	0.0111	-0.185	-0.0269
RF	8.62	-0.657	0.01	-0.184	-0.023
HD	8.2	-0.61	0.011	-0.122	-0.026
JM	8.2	-0.61	0.011	-0.122	-0.026
FCS	8.59	-0.727	0.013	-0.186	-0.028
FCS-LMM	-8.54	0.71	0.0122	-0.182	-0.0271

Table 1c 15 %

Variables	Group	Gender	Domicile	Education	Assault
Original	8.61419	-0.72853	0.01083	-0.18508	-0.0273
LWD	8.1002	-0.72853	0.01083	-0.18508	-0.0273
MI	8.521	-0.71	0.012	-0.0182	-0.026
KNN	8.612	-0.728	0.01	-0.185	-0.0261
RF	8.59	-0.727	0.013	-0.186	-0.028
HD	-8.54	0.71	0.012	-0.182	-0.027
JM	8.612	-0.728	0.01	-0.185	-0.027
FCS	8.614	-0.729	0.011	-0.185	-0.027
FCS-LMM	8.2	-0.61	0.011	-0.122	-0.026

Table 1d 20%

Variables	Group	Gender	Domicile	Education	Assault
Original	8.61419	-0.72853	0.01083	-0.18508	-0.0273
LWD	8.0090	-0.72853	0.01083	-0.18508	-0.0273
MI	8.62	-0.73	0.011	-0.187	-0.026
KNN	8.619	-0.728	0.0111	-0.185	-0.0269
RF	8.59	-0.727	0.013	-0.186	-0.028
HD	-8.54	0.71	0.012	-0.182	-0.027
JM	8.612	-0.728	0.01	-0.185	-0.027
FCS	8.614	-0.729	0.011	-0.185	-0.027
FCS-LMM	8.2	-0.61	0.011	-0.122	-0.026

IV. CONCLUSION

In the current paper, the performance of different MI methods (FCS-Standard , FCS-LMM , JM) and Machine Learning Imputation In the current study, we compared the performance of several different MI methods (FCS-Standard, FCS-LMM,, JM) and Machine learning imputation techniques (K-nearest neighbors and Random Forest)to handle missing values in longitudinal data in the context of fitting Linear mixed effect model with both random intercepts and slopes. Our comparison also revealed that Joint modeling (JM) approach holds great promise for the imputation of longitudinal data. The results from our theoretical exploration revealed, although several MI methods are available for imputing missing values in longitudinal, its is quite evident that Machine learning Imputation techniques can provide much better estimates. Due to its simplicity, easy-understanding and relatively high accuracy we conclude from our study that K nearest neighbor and random forest Machine learning imputation techniques proved to show better and efficient performance on comparison with other methods than Multiple Imputation like FCS and JM techniques on the basis of the coefficients obtaine on fitting the linear mixed model.

FUTURE SCOPE

In many public health contexts where data are collected from individuals repeatedly over time and from groups of people that are clustered within natural units, longitudinal and cluster-correlated data both emerge.By substituting missing data with an estimated value based on other available information, imputation preserves all cases. The data set can be analysed using methods used for complete data once all missing values have been imputed. With the upcoming data that is present in the health sector, there may be many situations where the missing data may arise and there must be ways to uncover that value to analyse the dataThere are numerous imputation approaches that may be utilised to compare longitudinal studies. Although it is becoming easier to access machine learning algorithms and a variety of approaches are being developed, it still takes a lot of processing time for these techniques to produce estimates for data that are effective and almost exact. Numerous methods can be developed to impute categorical and continuous types of data by modifying multiple imputation algorithms.

ACKNOWLEDGEMENT

I would also like to extend my sincere thanks to the Dr. Natasha, Faculty of Health at University of Canberra, Bruce, ACT, Australia, whose willingness to share their knowledge have significantly enhanced the quality of this paper.

REFERENCES

[1]. Van Buuren, S. Flexible Imputation of Missing Data (CRC Press Taylor, Francis Group , Chapman, and Hall Book, 1991).

[2]. Huque, M. H., Carlin, J., Simpson, J. & Lee, K. A comparison of multiple imputation methods for missing data in longitudinal studies. BMC Medical Research Methodology 18 (Dec. 2018).

[3]. in. Linear Mixed Models for Longitudinal Data 221–229 (Springer New York, New York, NY, 2000). isbn: 978-0-387-22775-7. https://doi.org/10.1007/978-0-387- 22775-7_16

[4]. Ben, ˆA. et al. The handling of missing data in trial-based economic evaluations: should data be multiply imputed prior to longitudinal linear mixed-model analyses? The European Journal of Health Economics, 1–15 (Sept. 2022).

[5]. in. Linear Mixed Models for Longitudinal Data 19–29 (Springer New York, New York, NY, 2000). isbn: 978-0-387-22775 https://doi.org/10.1007/978-0-387-22775-7_3.

[6]. Gurka, M. J. & Edwards, L. J. in Essential Statistical Methods for Medical Statistics (eds Rao, C., Miller, J. & Rao, D.) 146–173 (North- Holland, Boston, 2011). isbn: 978-0-444-53737-9. https://www.sciencedirect.com/science/article/pii/B97804445373795 0086.

[7]. Huque, M. H. et al. Multiple imputation methods for handling incomplete longitudinal and clustered data where the target analysis is a linear mixed effects model. Biometrical Journal 62 (2020).

[8]. Berglund, P. A. Multiple Imputation Using the Fully Conditional Specification Method : A Comparison of SAS ® , Stata , IVEware , and in (2015)

[9]. Molenberghs, G. & Verbeke, G. A review on linear mixed models forlongitudinal data, possibly subject to dropout. Statistical Modelling - STAT MODEL 1, 235–269(Dec. 2001).

[10]. 10. Sterne, J. A. C. et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 338 (2009).

[11]. Diggle, P., Liang, K.-Y. & Zeger, S. Analysis of Longitudinal Data. Biometrics 53, 782 (June1997).

[12]. in. Linear Mixed Models for Longitudinal Data 209–219 (Springer New York, New York, NY, 2000). isbn: 978-0-387-22775-7. https://doi.org/10.1007/978-0-387-22775-7_15.

[13]. Quartagno, M., Grund, S. & Carpenter, J. jomo: A Flexible Package for Two-level Joint Modelling Multiple Imputation. The R Journal 9 (Jan. 2019).

[14]. Enders, C., Hayes, T. & Du, H. A Comparison of Multilevel Imputation Schemes for Random Coefficient Models: Fully Conditional Specification and Joint Model Imputation with Random Covariance Matrices. Multivariate Behavioral Research 53, 1–19 (Jan. 2019).

[15]. Cover, T. & Hart, P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13, 21–27 (1967).

[16]. Quartagno, M. & Carpenter, J. Substantive model compatible multilevel multiple imputation: A joint modeling approach. Statistics in Medicine 41 (Aug. 2022).

[17]. Kumar, P. Imputation and characterization of uncoded self-harm in major mental illness using machine learning. Journal of the American Medical Informatics Association: JAMIA 27,1,136–146 (2020).

[18]. Riiser, K., Richardsen, K., Haugen, A., Lund, S. & Løndal, K. Active Play in ASP –A matched-pair cluster-randomized trial investigating the effectiveness of an intervention in After-school programs for supporting children's physical activity (Apr. 2020).

[19]. Laukkanen, A. Physical activity and motor competence in 4–8-year old children: results of a family-based cluster-randomized controlled physical activity trial isbn: ISBN 978-951-39-6582-2 (PDF) (Apr. 2016).

[20]. in. Linear Mixed Models for Longitudinal Data 201–207 (Springer New York, New York, NY, 2000). isbn: 978-0-387-22775-7. https://doi.org/10.1007/978-0-387-22775-7_14.

[21]. Knowledge of sexual abuse and resistance ability among children with intellectual disability. Child Abuse Neglect 136, 105985. issn: 0145-2134.https://www.sciencedirecet.com/science/article/pii/S0145213422005191(2023).