Scientific Research Publishing

Open Journal of Statistics, 2025, 15(3), 337-360
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Evaluating Utility of Machine Learning-Based Imputation Methods to Account for Attrition in Multi-Stage Epilepsy Prevalence Surveys

Daniel M. Mwanga[1,2*], Isaac C. Kipchirchir[1], George O. Muhua[1], Charles R. Newton[3,4], Damazo T. Kadengye[2]

[1]Department of Mathematics, University of Nairobi, Nairobi, Kenya
[2]African Population and Health Research Center, Nairobi, Kenya
[3]Department of Psychiatry, University of Oxford, Oxford, United Kingdom
[4]Kenya Medical Research Institute—Wellcome Trust Research Programme, Kilifi, Kenya
Email: *mtaimwanga@gmail.com

How to cite this paper: Mwanga, D.M., Kipchirchir, I.C., Muhua, G.O., Newton, C.R. and Kadengye, D.T. (2025) Evaluating Utility of Machine Learning-Based Imputation Methods to Account for Attrition in Multi-Stage Epilepsy Prevalence Surveys. *Open Journal of Statistics*, **15**, 337-360.
https://doi.org/10.4236/ojs.2025.153018

Received: May 7, 2025
Accepted: June 27, 2025
Published: June 30, 2025

Abstract

Attrition is a common challenge in statistical analysis for longitudinal or multi-stage cross-sectional studies. While strategies to reduce attrition should ideally be implemented during the study design phase, they remain common in real-world research, necessitating statistical methods to address them. Traditional approaches like multiple imputation (MI) and inverse probability weighting (IPW) rely on the assumption that data is missing at random (MAR), which is not always plausible. Recent developments in machine learning (ML) based methods offer promising alternatives because of their ability to capture complex patterns in data and handle non-linear relationships more effectively. This study examines four ML-based imputation methods to account for attrition and compares them with conventional MI and IPW in a two-stage epilepsy population-based prevalence survey involving 56,425 participants. Simulated attrition levels from 5% to 50% were applied following the MAR mechanism to assess the performance of the different methods. This was replicated 100 times using different random seeds. Results showed that bias increased with an increase in attrition levels. Complete case analysis had the largest bias in all scenarios. *k*-nearest neighbor (KNN) and sequential KNN (sKNN) performed similarly to MI under MAR but exhibited less bias than MI and IPW when data were MNAR. While IPW performed similarly to MI under MAR, it had greater bias under MNAR. Both missForest and the MI implemented using random forest were outperformed by sKNN and KNN. We have demonstrated that even a small attrition proportion of 5% can significantly bias estimates if not properly addressed. While MI is still the most

preferred for missing data assuming MAR, ML methods, particularly sKNN and KNN demonstrated potential for addressing attrition when data are MNAR. Choosing the appropriate method to address missing data should be preceded by an evaluation of different available methods that could be suitable for the data being analysed. Future research should explore ML methods in various study designs and consider integrating ML into the very robust MI framework to improve prediction accuracy for missing data due to attrition.

Keywords

Prevalence, Missing Data, Machine Learning, Multiple Imputation, Inverse Probability Weighting, Attrition, Epilepsy, Population-Based Studies

1. Introduction

Missing data, whether due to attrition or other causes, poses a common challenge in any statistical analysis. In practice, for studies with multiple timepoints such as longitudinal and multi-stage cross-sectional studies, attrition is often inevitable. Attrition results in data missing on outcomes or covariates of interest at that specific time point. The process of accounting for attrition should ideally start at the point of study design, where strategies can be put in place to enhance response rates. While attrition can be minimized, in practice, it cannot be entirely eliminated. Thus, there is a need to consider methods which researchers can use to account for attrition.

For diseases such as epilepsy, the application area for this study, the most common method of estimating prevalence consists of at least two stages [1] [2]. The first stage is used for screening all individuals in the target area (through a census) to detect the possible cases of epilepsy, and the subsequent stages are used for confirmation by a trained physician (most robustly, a neurologist) [3] [4]. This design often faces the challenge of attrition, which occurs when participants screened in the first stage fail to participate in the follow-up confirmation stage(s).

A number of methods exist for accounting for attrition. They range from simple methods such as complete case analysis (ignoring missingness), last observation carried forward (LOCF), single imputations methods such as mean imputations, regression imputation and maximum likelihood estimation (also called direct likelihood) [5], to more advanced methods such as inverse probability weights (IPW) and multiple imputation (MI). Below, we discuss three commonly used methods, namely CCA, MI and IPW, which are three of the most commonly applied methods in recent literature [6].

Complete case analysis works by restricting the analysis to the observed data and thus ignores missingness. This method yields unbiased results when the missing data pattern is completely at random (MCAR) [6]. In practice, however, MCAR is uncommon, which means that using CCA has an increased risk of producing biased estimates.

The inverse probability weighting method works by assigning weights to each observation based on the probability of being observed, thereby giving more weight to observations that are less likely to be missing. The common approach is using propensity scores generated as predicted values from a fitted model that includes covariates related to the missingness [6], for example, socio-demographic factors associated with non-response. Inverse probability weights are calculated as the inverse of the predicted probability of being observed. This means that observations with low probability of being observed (that is, are more likely to be missing) would have higher weights, while observations with high probability of being observed would have lower weights. Each individual observation is weighted by its corresponding inverse probability weight. This means that observations with higher weight would have greater influence on the analysis effectively giving more importance to observations that are less likely to be missing. Analysis is conducted using the weighted data.

Multiple imputation (MI) is commonly used when the missing data mechanism is at least missing at random (MAR). It can be applied to impute continuous, binary, and categorical variables [7] [8]. MI replaces missing values with plausible values drawn from the posterior predictive distribution of the missing data, conditional on the observed data.

A major assumption that must be met to apply MI and IPW is that the data must be MAR or MCAR. Not negating its important role in helping analysis navigate the problem of missing data, MI has limitations in some settings [9]-[14]. One of the main limitations is that it is not appropriate when data are MNAR and the MAR assumption can not be tested with empirical data. Further, its efficiency may not be guaranteed if the missing proportion is greater than 40% [15], and especially if the MAR assumption is implausible. As noted by Kristman *et al.* [14], attrition is rarely random and MNAR seriously biases estimates. MI is computationally intensive and involves a lot of approximations.

A recent study has shown that, while MAR and MCAR could be sufficient conditions for consistent estimation with specific methods, they may not always directly determine the best approach for handling the missing data in question [16] with sensitivity analysis needed to test plausibility [17]. Further, the most commonly used model in the MI framework is logistic regression for binary outcomes. However, in comparison with newer approaches such as machine learning, logistic regression has often been outperformed by algorithms like random forest and extreme gradient boosting methods. As the development and application of machine learning (ML) methods continue to evolve, there is attention to their potential in addressing the challenges posed by missing data due to attrition.

Machine learning methods, which are able to learn patterns in a dataset, identify trends and make predictions based on large datasets, offer a promising avenue for handling missing data in a way that goes beyond traditional imputation and weighting techniques. One of the recent developments has been application of machine learning algorithms to handle missing data include use of random forest,

missForest and k-nearest neighbors (KNN) implemented through common statistical software such as *mice* and *caret* packages in the R software [18] [19]. This is an active area of research to determine how ML methods perform or complement the established MI and IPW methods. It remains largely unexplored, how the new ML methods perform in the context of attrition in the analysis of prevalence using multi-stage population-based surveys.

In this paper, using a real dataset on epilepsy, we evaluate the performance of four ML-based imputation methods, namely KNN, sequential KNN, an iterative imputation method called missForest [20], which uses the random forest algorithm, and multiple imputation implemented using random forest as the imputation model. We also compare their performance against the common approaches such as MI and IPW. By leveraging the predictive power of machine learning algorithms, researchers can improve the accuracy, efficiency, and robustness of imputation procedures. We emphasize the importance of understanding the underlying assumptions and considerations when applying different methods for accounting for attrition and highlight avenues for future research in the field.

2. Materials and Methods

2.1. Study Setting and the Motivating Study

The data used in this analysis are based on an epilepsy prevalence study conducted in the two informal settlements, under the Epilepsy Pathway Innovation in Africa (EPInA) project, conducted in Nairobi (Protocol reference: NIHR200134) [21]. It was set up to improve epilepsy treatment pathways, including prevention, diagnosis, treatment and awareness. The Nairobi site covered two urban informal settlements, namely Viwandani and Korogocho, which form the Nairobi Urban Health and Demographic Surveillance System (NUHDSS) that is led by the African Population and Health Research Center (APHRC). Like most other urban informal settlements in Nairobi, Viwandani and Korogocho are characterized by lack of basic infrastructure, poor sanitation, overcrowding, high unemployment rate, poverty, and inadequate health infrastructure. Epilepsy studies have been conducted more predominantly in rural settings. This site was selected because it represents urban poor settlements in Nairobi. Viwandani is a more mobile population where most residents are workers of the nearby companies in the industrial area of Nairobi. Korogocho is a more settled population where most residents have stayed there all their life. The two settings provide a suitable environment to study attrition in urban settings, which is the focus of this paper. Detailed information about the NUHDSS is available elsewhere [22] [23].

2.2. Study Design

The data are from a population-based cross-sectional prevalence survey (census) conducted in the NUHDSS in Nairobi, under the EPInA project. The survey had two stages of screening patients for epilepsy. In the first stage, trained field interviewers administered a standardized validated screening questionnaire with

14 items [23] to the head of household or an adult representative in the household to identify persons with history of epilepsy. Socio-demographic characteristics of all members of the household were collected at this stage, including age and sex. Participants identified as possible cases of epilepsy in the first stage would then be invited for assessment by the neurologist at a nearby facility (second stage). The participants were invited through scheduled appointments, and those who missed appointments were physically traced using confidential contact and residential information they provided in the first stage. The first stage of screening was conducted between 21st September 2021 and 21st December 2021, and the second stage between 14th April 2022 and 6th August 2022.

2.3. The dataset and simulation

The entire EPInA dataset in the Nairobi site consisted of 56,425 participants, of whom 1126 were screened as possible cases of epilepsy in the first stage of screening (at household level) and 873 of the possible cases completed the second stage (assessment by a neurologist at a health clinic). Data with 0% attrition is not feasible in practice. Therefore, for this analysis, we construct a hypothetical '*gold standard*' dataset based on the complete observations from the EPInA dataset. We exclude possible cases that were not screened by the neurologist. Thus, we consider the dataset with 56,172 records as the dataset with *no attrition*, for the purpose of comparison and determining the methods that better account for attrition in a population-based epilepsy prevalence survey. This excludes the 253 individuals lost to follow-up from stage 2.

We simulated attrition at different levels, denoted by λ. For each attrition level, a new variable was generated to reflect the induced missingness. Attrition rates of λ = 5%, 10%, 20%, 30%, 40%, 50% were randomly imposed on the data, with the process repeated 100 times using different random seed values. Attrition was imposed only among the 1126 possible cases, reflecting real-world follow-up loss. The reported estimates were obtained by computing the mean across all 100 replications. These proportions were selected to represent small, moderate, and high levels of attrition. As a result, the analytical dataset includes a variable with complete information (*no attrition*) and new variables with incomplete information at varying levels of attrition (λ). We simulated two missingness mechanisms: MAR (Missing at Random), by introducing differential attrition between the Viwandani and Korogocho sites; and MNAR (Missing Not at Random), by manipulating the missingness for sex and age variables to ensure that missingness is related to the missing data itself. This dual approach allows for an evaluation of how each method performs under more realistic and challenging missing data scenarios. MNAR was only evaluated when examining the relationship between the outcome and covariates.

2.4. Outcome

The primary outcome of this study is the prevalence of epilepsy. It is measured as

the proportion of individuals who were confirmed as having epilepsy by a neurologist in the second stage out of the population size captured in the first stage.

2.5. Covariates

In addition to prevalence estimation, for the purpose of determining utility of the imputed datasets, we also examine the association between epilepsy and key socio-demographic characteristics namely site, sex and age. These covariates were chosen just for the purposes of comparing the methods to account for attrition and because they are the most commonly analyzed demographic variables in epidemiological studies.

2.6. Statistical Models

2.6.1. Notations

The notation used in the statistical models is as follows:

Y is a binary outcome indicating epilepsy diagnosis ($Y=1$ if confirmed, $Y=0$ if not confirmed).

$\boldsymbol{x}$ is a vector of covariates used in the regression models.

$\hat{P}\left(Y=1|\boldsymbol{x}\right)$ is predicted probability of epilepsy given the covariates.

$\hat{\theta}$ is the estimated prevalence of epilepsy.

R_i is the response indicator ($R_i=1$ if observed, $R_i=0$ if missing).

ω_i is the inverse probability weight for observation i.

M is the number of imputations used in multiple imputation.

$\hat{\beta}_m$ is the estimate from the m^{th} imputed dataset.

$\bar{\beta}$ is the mean estimate across the M imputations.

MCE is the Monte Carlo Error of the estimates from the imputation.

In all tables in the results section, σ denotes the standard error from the dataset with *no attrition*, σ_m denotes the standard error from the dataset with some level of attrition, τ denotes the p-value, and ζ denotes the attrition bias.

2.6.2. The Logistic Regression Model

Let Y be a binary random variable such that $P\left(Y=1|\boldsymbol{x}\right)$ denotes the probability of being diagnosed with epilepsy, and $P\left(Y=0|\boldsymbol{x}\right)$ denotes the probability of not being diagnosed.

$$Y=\begin{cases}1, & \text{if participant was confirmed as having epilepsy;}\\ 0, & \text{if the participant was screened negative at stage 1 or assessed}\\ & \text{by a neurologist and not diagnosed with epilepsy at stage 2.}\end{cases}$$

Our objectives are to estimate the prevalence of epilepsy (θ), and identify associated factors using a logistic regression model, specified as

$$\log\left(\frac{P\left(Y=1|\boldsymbol{x}\right)}{1-P\left(Y=1|\boldsymbol{x}\right)}\right)=\boldsymbol{x}^{\top}\boldsymbol{\beta} \tag{1}$$

where $\boldsymbol{x}$ is the vector of covariates (including an intercept) and $\boldsymbol{\beta}$ is the vector

of regression coefficients.

2.6.3. The Multiple Imputation Model

The multiple imputation model included both the covariates from the substantive model (Equation 1) and screening variables from validated epilepsy screening tools [2] [24]. Sociodemographic variables used in the substantive model were also included in the imputation model, following best practice recommendations [8] [25]. All methods were evaluated across attrition levels $\lambda = \{5\%, 10\%, 20\%, 30\%, 40\%, 50\%\}$.

Let Y_j denote the binary epilepsy diagnosis variable with missing values. We model the missing values using logistic regression:

$$P\left(Y_j = 1 \mid \boldsymbol{x}_{\text{obs}}\right) = \frac{1}{1 + \mathrm{e}^{-\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\right)}} \tag{2}$$

1) For each missing value Y_j in $\boldsymbol{Y}_{\text{mis}}$, predicted probabilities are computed based on the logistic regression model

$$\hat{P}\left(Y_j = 1 \mid \boldsymbol{x}_{\text{obs}}\right) = \frac{1}{1 + \mathrm{e}^{-\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k\right)}}$$

where $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ are the estimated regression coefficients from the logistic regression model.

2) For each missing Y_j, we generate a random value from a Bernoulli distribution with success probability $\hat{P}\left(Y_j = 1 \mid \boldsymbol{x}_{\text{obs}}\right)$. This step ensures that the imputed values reflect the uncertainty in the predicted probabilities.

3) We perform this imputation multiple times (such as, M times) to create M complete datasets, where each dataset has a different set of imputed values for $\boldsymbol{Y}_{\text{mis}}$. For this study, we set $M = 50$. Although a minimum of $M = 5$ is commonly used, larger values are preferred to reduce Monte Carlo Error (MCE) of the estimate, which is computed as the standard deviation of the estimates across all M. More specifically,

$$\text{MCE} = \sqrt{\frac{\sum_{m=1}^{M}\left(\hat{\beta}_m - \bar{\beta}\right)^2}{M\left(M-1\right)}} \tag{3}$$

where $\hat{\beta}_m$ is the estimate based on the m^{th} imputation, and $\bar{\beta}$ is the arithmetic mean of the estimates from all the M imputed datasets. Computation of the MCE is the same also for the prevalence estimare $\hat{\theta}$.

The final step in the multiple imputation process involves combining the results from the M imputed datasets using Rubin's combination rule [6] [26]. According to Rubin's rules, the combined estimate $\hat{\beta}$ of a parameter β is given by

$$\hat{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}^{(m)} \tag{4}$$

where $\hat{\beta}^{(m)}$ is the estimate of β from the m^{th} imputed dataset. The variance $\text{Var}\left(\hat{\beta}\right)$ is obtained as

$$\mathrm{Var}\left(\hat{\beta}\right)=\frac{1}{M}\sum_{m=1}^{M}\mathrm{Var}^{(m)}\left(\hat{\beta}\right)+\frac{1+\frac{1}{M}}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}^{(m)}-\hat{\beta}\right)^{2} \tag{5}$$

This accounts for both within-imputation variability (the first term) and between-imputation variability (the second term). The same combination rules were applied to both prevalence estimates $\hat{\theta}$ and regression coefficients $\hat{\beta}$ from the multiply imputed datasets.

2.6.4. Inverse Probability Weighting Model

To adjust for attrition bias, we modeled the probability of response ($R_i=1$) using logistic regression:

$$\log\left(\frac{P\left(R_i=1\mid \boldsymbol{x}_i\right)}{1-P\left(R_i=1\mid \boldsymbol{x}_i\right)}\right)=\boldsymbol{x}_i^{\top}\boldsymbol{\gamma} \tag{6}$$

The vector $\boldsymbol{\gamma}$ represents the regression coefficients corresponding to the covariates $\boldsymbol{x}$. These coefficients quantify the association between each covariate and the log odds of being observed, that is, $R_i = 1$.

Weights were computed as:

$$\omega_i=\frac{1}{\hat{p}_i} \tag{7}$$

A weighted logistic regression was then fitted to the observed outcome:

$$\log\left(\frac{P\left(Y_i=1\mid \boldsymbol{x}_i\right)}{1-P\left(Y_i=1\mid \boldsymbol{x}_i\right)}\right)=\boldsymbol{x}_i^{\top}\boldsymbol{\beta} \tag{8}$$

The weights ω_i adjust for selection bias due to attrition, giving more influence to underrepresented individuals and improving the robustness of the parameter estimates.

2.7. Machine Learning-Based Methods

2.7.1. missForest

missForest is a random forest-based approach used to handle missing data. It is particularly effective for mixed-type data and captures nonlinear relationships well. It works by building a series of random forest models, one for each variable with missing values, using the other variables as predictors. The idea is to use the patterns in the observed data to estimate the missing parts.

In practice, the algorithm starts by filling in missing values using a simple method like the mean or mode. Then, for each variable with missing data, a random forest model is trained using only the complete cases. This model is used to predict the missing values in that variable. Once all variables have been processed, the algorithm checks how much the new imputations differ from the previous round. This cycle is repeated until the changes between iterations are small enough to stop.

2.7.2. *k*-Nearest Neighbour (KNN)

KNN is a distance-based algorithm commonly used in classification and regression tasks. In the context of imputation, it estimates missing values by identifying the *k* nearest observations in the dataset based on available data. Distance is typically calculated using metrics like Euclidean or Gower distance, depending on variable types.

For continuous variables, the imputed value is usually the mean of the *k* nearest neighbors. For binary or categorical variables, the mode of the neighbors is used instead. When imputing binary outcomes—such as epilepsy diagnosis—the algorithm determines which of the two classes (such as, 0 or 1) appears most frequently among the neighbors and assigns that as the imputed value. This approach preserves the binary nature of the data while still leveraging the similarity structure in the observed dataset.

2.7.3. Sequential *k*-Nearest Neighbour

Sequential KNN extends the basic KNN method by iteratively imputing one variable at a time. At each step, KNN is applied to fill in missing values for a single variable, using the currently available and previously imputed data as inputs. After each round, the dataset is updated, and the process continues with the next variable.

As with standard KNN, binary variables are imputed by identifying the *k* nearest neighbors and selecting the most frequent class among them. This majority-vote mechanism ensures that the imputed values remain binary. The sequential structure allows for improved accuracy, particularly when multiple variables have missing data, by incorporating more information as the algorithm progresses.

Choice of *k* in KNN and sKNN

The performance of *k*-nearest neighbors (KNN) and sequential KNN (sKNN) imputation methods depend on the choice of the parameter k, which determines the number of nearest neighbors considered when imputing missing values. In this study, we used $k=5$, a commonly used default in the literature for binary and categorical data [27] [28]. For binary outcome variables, this means that each missing value is imputed using a majority among the five nearest neighbors with observed values. For example, if at least 3 out of the 5 nearest neighbors have the value 1, the imputed value is set to 1; otherwise, it is set to 0. This approach balances sensitivity to local data structure with stability across the dataset.

We selected $k=5$ based on preliminary testing and practical considerations. Larger values of k tend to smooth over local variation while smaller values (for example, $k=1$ or $k=3$) can introduce noise due to overfitting in some instances. To assess the robustness of this choice, we conducted a sensitivity analysis using $k=3$ and $k=7$. The results were consistent across the different values of k.

2.7.4. Multiple Imputation Using Random Forest

Multiple imputation (MI) implemented together with random forest (MI with RF) models as the underlying imputation model leverages the strengths of machine learning to capture complex, nonlinear relationships between variables during the imputation process. Similar to other machine learning models, random forest is able to learn trends and patterns in the training dataset and use it to predict a new set of data.

In this approach, for each incomplete variable Y_j, a random forest model is fit using the other observed variables $\boldsymbol{X}_{-j}$ as predictors. In this context, $\boldsymbol{X}_{-j}$ represents all variables except the j-th variable Y_j, which is the target of the current imputation. The model predicts the missing values of Y_j by sampling from the conditional distribution estimated by the random forest rather than simply using point predictions. This stochastic element allows for proper variability between imputations. Stochastic sampling can be implemented using methods such as predictive mean matching or drawing from the distribution of trees in the forest to reflect imputation uncertainty.

For each variable Y_j containing missing data, a random forest model is trained using the observed values Y_j^{obs} and the other variables $\boldsymbol{X}_{-j}$ as predictors. Imputed values Y_j^{imp} for the missing entries are then generated by sampling from the predictive distribution estimated by the random forest. This procedure is repeated sequentially for all variables with missing data, with imputations updated based on the most recent values of other variables. Finally, the entire iterative imputation process is performed m times to produce m completed datasets, each reflecting the uncertainty inherent in the imputation.

For missing data in variable Y_j, the imputed values at iteration t can be written as:

$$Y_j^{imp,(t)} \sim f_{\mathrm{RF}}\left(Y_j \mid \boldsymbol{X}_{-j}^{imp,(t-1)}\right),$$

where f_{RF} is the conditional distribution modeled by the random forest, and $\boldsymbol{X}_{-j}^{imp,(t-1)}$ denotes the latest imputed predictors from the previous iteration.

After obtaining the m completed datasets, analyses are performed separately on each, and results are combined using Rubin's rules as shown in equations 4 and 5.

2.8. Statistical Analysis

Descriptive statistics were used to summarize the data including means and standard deviations for approximately normally distributed continuous variables, medians and interquartile ranges for skewed continuous variables, and frequencies or proportions for categorical variables. We present results from the dataset with *no attrition*, alongside those from a dataset in which missing data were imputed. To compare how different methods accounted for attrition, we report the point estimate their standard errors (σ), 95% confidence intervals, and the attrition bias (ζ), defined as the absolute difference between the estimate from the

attrition-affected dataset and that from the dataset with *no attrition*. For this analysis, the best model is defined as the one that minimizes attrition bias.

All statistical tests considered in this study were conducted at a 5% significance level ($\alpha = 0.05$). We report both the lower class boundary (LCB) and upper class boundary (UCB) of the 95% confidence intervals for all estimates. The focus of fitting the logistic regression model is to determine the association between epilepsy and key socio-demographic characteristics. The dependent variable was binary (epilepsy diagnosis), and the independent variables included site (1 = 'Korogocho', 0 = 'Viwandani'), age (1 = 'five years or younger', 2 = '6 to 12 years', 3 = '13 to 18 years', 4 = '19 to 28 years', 5 = '29 to 49 years', and 6 = '50 years or older'), and sex (0 = 'female', 1 = 'male'). These covariates and their categorization were selected for demonstration purposes and to simplify comparisons across different methodologies. We included all the three covariates in all the models.

2.9. Training Machine Learning Models

We evaluated four machine learning-based imputation models: *missForest* [20], k-nearest neighbour (KNN), sequential KNN (sKNN) and multiple imputation implemented with random forest model (MI with RF). These models were selected because they are widely used and have demonstrated strong performance in similar studies [29]-[31]. Training was performed on the dataset with *no attrition*, while testing was conducted using datasets with varying levels of missingness due to attrition (λ). The performance of the ML-based imputation models was evaluated using the following metrics.

2.9.1. Accuracy

$$\text{Accuracy} = \frac{\text{Number of correctly classified instances}}{\text{Total number of instances}} \tag{12}$$

Accuracy ranges from 0 to 1, with higher values indicating greater classification performance.

2.9.2. F1 Score

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

Precision is the proportion of true positive predictions among all positive predictions, while Recall (also known as sensitivity or true positive rate) is the proportion of true positives among all actual positive instances. The F1 score ranges from 0 to 1, with higher values indicating better performance. An F1 score above 0.7 is generally recommended [32].

2.9.3. Area under the Receiver Operating Characteristic Curve (AUC)

$$\text{AUC} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{14}$$

Here, sensitivity is the proportion of true positives correctly identified, and specificity is the proportion of true negatives correctly identified. AUC values

range from 0 to 1, with higher values indicating better discriminatory ability. AUC values above 0.7 are considered acceptable [33].

3. Results

3.1. The Substantive Logistic Regression Model

In this paper, the substantive model is estimated from the data with *no attrition*, the results against which subsequent findings based on the various methods used to account for attrition are compared. Here, we estimate prevalence of epilepsy and fit the logistic regression model to determine the association between site, sex and age of the participant. Table 1 presents the prevalence of epilepsy expressed per 1000 people and the 95% confidence interval.

Table 1. Prevalence based on the dataset with *no attrition*.

revalence/1000	Lower CI (L)	Upper CI (U)	U-L
9.40	8.60	10.20	1.60

Overall, the prevalence estimate against which the missing data methods are compared is 9.4 cases per 1000 people, and a 95% confidence interval of 8.6 to 10.2. Table 2 shows the estimates from the subtantive logistic regression model, against which the estimates from the logistic regression models based on datasets with missing data accounted for by different approaches are compared.

Table 2. Logistic regression model based dataset with *no attrition*.

	β	σ	τ	Lower 95% boundary	Upper 95% boundary
Site (Ref = Viwandani)					
Korogocho	0.304	0.089	0.001	0.130	0.478
Sex (Ref = Male)					
Female	0.101	0.088	0.251	−0.071	0.273
Age in years (Ref = under 5 years)					
6 - 12 years	0.691	0.199	0.001	0.301	1.081
13 - 18 years	0.791	0.206	<0.001	0.388	1.194
19 - 28 years	0.717	0.184	<0.001	0.356	1.078
29 - 49 years	0.746	0.179	<0.001	0.395	1.097
50 years or older	0.368	0.243	0.130	−0.108	0.845
Constant	−5.468	0.173	<0.001	−5.808	−5.129

Notes: Ref = Reference category, τ is p-value, σ is standard error and β in this table are the log odds of being diagnosed with epilepsy given the covariates.

3.2. Complete Case Analysis, Multiple Imputation and Inverse Probability Weighting

Below, we compare prevalence obtained by CCA, MI and IPW for different levels of missingness (assuming MAR) against the prevalence from the dataset with *no attrition*. We compare the attrition bias and precision. Precision is assessed by how the missing data methods estimate the confidence intervals and the standard errors. Table 3 presents prevalence estimates, confidence intervals when missing data is handled using CCA, MI and IPW. It also presents attrition bias, which is the difference between the estimate by CCA, IPW and MI and the estimate and the estimate from the dataset with *no attrition.*

Table 3. Prevalence based on data analyzed using CCA and when accounted for MI and IPW under MAR.

Attrition/Methods	Prevalence/1000	σ_m	LCB (L)	UCB (U)	U-L	ζ
0% (no attrition)	9.40	0.41	8.60	10.20	1.60	-
CCA						
5%	8.91	0.40	8.13	9.69	1.56	0.49
10%	8.38	0.39	7.63	9.14	1.51	1.02
20%	7.36	0.36	6.65	8.07	1.42	2.04
30%	6.27	0.33	5.61	6.92	1.31	3.13
40%	5.63	0.32	5.01	6.25	1.24	3.77
50%	4.45	0.28	3.90	5.01	1.11	4.95
MI						
5%	9.36	0.41	8.56	10.17	1.61	0.04
10%	9.32	0.41	8.51	10.13	1.62	0.08
20%	9.29	0.43	8.45	10.13	1.68	0.11
30%	9.16	0.44	8.30	10.03	1.73	0.24
40%	9.31	0.46	8.40	10.22	1.82	0.09
50%	8.94	0.44	8.07	9.80	1.73	0.46
IPW						
5%	10.06	0.46	9.15	10.97	1.82	0.66
10%	9.44	0.44	8.57	10.30	1.73	0.04
20%	9.44	0.48	8.49	10.38	1.89	0.04
30%	9.27	0.53	8.22	10.31	2.09	0.13
40%	9.31	0.54	8.25	10.36	2.11	0.09
50%	8.97	0.60	7.79	10.14	2.35	0.43

As shown in Table 3, both MI and IPW resulted in prevalence estimates that are closer ($\zeta < 1$) to the value based on data with *no attrition* compared to CCA, across all levels of attrition. Generally, bias increased with increase in the proportion of attrition, particularly for CCA.

Further, we fit the logistic regression model to determine association between socio-demographic characteristics and prevalence under complete case analysis and when attrition is accounted for using MI and IPW. Table 4 presents estimates comparing the bias in odds ratio estimates when complete case is used, and when attrition is accounted for by MI and IPW.

Table 4. Attrition bias on the log odds of the covariates in a logistic regression model under CCA and after accounting for attrition using MI and IPW.

	CCA						MI						IPW					
λ	5%	10%	20%	30%	40%	50%	5%	10%	20%	30%	40%	50%	5%	10%	20%	30%	40%	50%
MAR																		
Site (Ref = Viwandani)																		
Korogocho	0.037	0.099	0.369	0.672	0.482	0.675	0.002	0.005	0.001	0.021	0.050	0.096	0.049	0.033	0.031	0.061	0.040	0.064
Sex (Ref = Male)																		
Female	0.032	0.016	0.044	0.079	0.089	0.054	0.004	0.013	0.015	0.025	0.002	0.011	0.004	0.011	0.024	0.050	0.081	0.064
Age in years (Ref = 0 - 5 years)																		
6 - 12 y	0.029	0.007	0.071	0.147	0.126	0.083	0.018	0.035	0.075	0.033	0.184	0.021	0.018	0.032	0.008	0.032	0.053	0.077
13 - 18 y	0.041	0.016	0.009	0.134	0.073	0.037	0.034	0.050	0.123	0.163	0.075	0.107	0.086	0.033	0.061	0.075	0.029	0.105
19 - 28 y	0.032	0.019	0.055	0.164	0.206	0.215	0.006	0.028	0.056	0.008	0.152	0.041	0.051	0.022	0.009	0.147	0.209	0.222
29 - 49 y	0.023	0.013	0.049	0.152	0.176	0.153	0.012	0.018	0.076	0.018	0.098	0.012	0.100	0.043	0.038	0.073	0.130	0.108
50 y or older	0.081	0.033	0.014	0.102	0.020	0.132	0.046	0.088	0.072	0.029	0.037	0.076	0.046	0.052	0.081	0.107	0.089	0.126
MNAR																		
Site (Ref = Viwandani)																		
Korogocho	0.030	0.059	0.034	0.149	0.148	0.131	0.012	0.015	0.032	0.044	0.026	0.016	0.003	0.122	0.163	0.289	0.209	0.156
Sex (Ref = Male)																		
Female	0.100	0.415	0.727	1.383	2.919	3.820	0.014	0.020	0.050	0.112	0.240	0.343	0.037	0.390	0.693	1.352	2.944	3.924
Age in years (Ref = 0 - 5 years)																		
6 - 12 y	0.993	0.055	0.051	0.055	0.052	0.050	0.089	0.142	0.034	0.073	0.031	0.092	1.228	0.162	0.073	0.056	0.015	0.013
13 - 18 y	0.993	0.725	0.229	0.232	0.222	0.218	0.090	0.145	0.191	0.188	0.187	0.167	1.062	0.803	0.276	0.276	0.195	0.205
19 - 28 y	0.989	0.986	0.763	0.428	0.423	0.423	0.087	0.080	0.029	0.051	0.028	0.029	1.324	1.166	0.841	0.492	0.362	0.378
29 - 49 y	0.996	1.014	1.034	0.857	0.426	0.435	0.088	0.078	0.018	0.021	0.011	0.005	1.255	1.149	1.105	0.906	0.369	0.385
50 y or older	1.008	1.049	1.082	1.164	1.248	0.920	0.090	0.078	0.017	0.009	0.111	0.091	1.243	1.083	1.143	1.206	1.239	0.897

Notes: Ref = Reference category.

For the MAR scenario, MI generally shows lower biases in log odds, followed by IPW. The bias was largest for complete case analysis. For instance, the bias for site variable ranged from 0.037 to 0.675 for 5% to 50% attrition when CCA was used, but ranged from 0.001 to 0.096 when MI was used and 0.031 to 0.064 for

IPW. Similarly, for the sex variable, bias in the log odds ranged from 0.016 to 0.089 when CCA was but from 0.002 to 0.025 when MI was used and from 0.004 to 0.081 for IPW.

In the MNAR scenario, the bias increased, especially for the sex and age variable. While bias generally increased, it was greater for CCA and IPW than MI. For instance, for sex, bias ranged from 0.100 to 3.820 for CCA, from 0.014 to 0.343 for MI and from 0.037 to 3.924 for IPW. Similarly, for age, for instance those aged 19 to 28 years old, bias ranged from 0.423 to 0.989 for CCA, from 0.028 to 0.087 for MI and from 0.362 to 1.324 for IPW.

3.3. Evaluation of ML-Based Imputation Methods

Figure 1 presents key metrics used to evaluate the four ML-based methods for different levels of attrition.

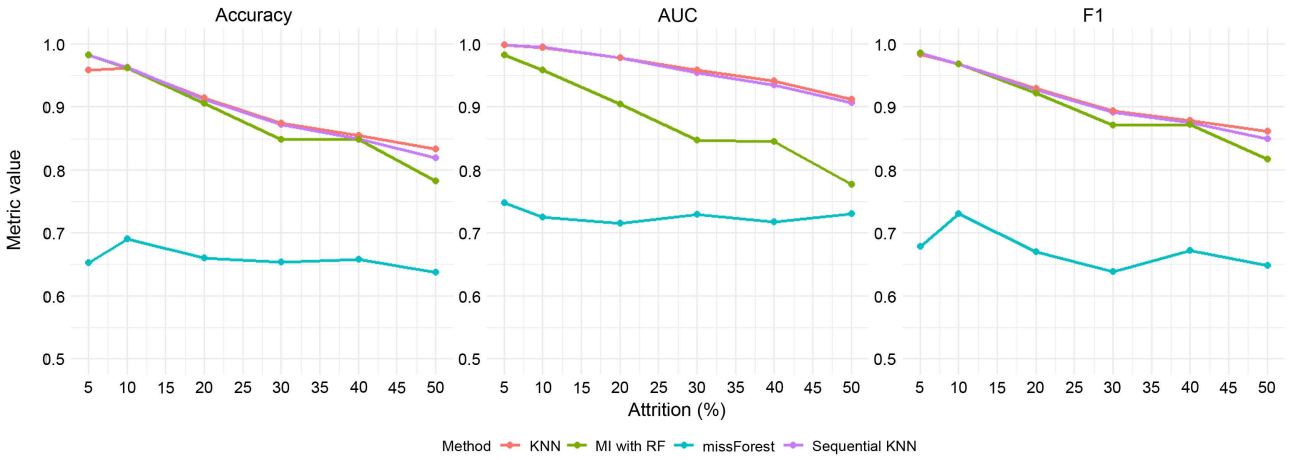

Figure 1. Evaluation of ML-based imputation methods under varying levels of attrition.

Based on all performance metrics used KNN and sKNN performed better than both missForest and random forest implemented within the MI framework. Overall, sequential KNN and KNN had similary the best performance across all metrics (accuracy, F1 score, and AUC). The performance however reduced with increase in attrition levels. The AUC for sKNN ranged from 0.907 for 50% attrition to 0.999 for 5% attrition. It ranged from 0.912 for 50% to 0.999 for 5% attrition under MAR assumption and 0.945 to 0.996 under MNAR. This suggests that sKNN and KNN are better performing models in predicting the missing data due to attrition in our study. MI with RF also performed well, closely following KNN methods up to 30% attrition, but with a steeper performance drop at higher attrition. The missForest model showed consistently lower performance across all metrics and attrition levels, with modest variation and no clear advantage as missingness increased. These findings suggest that sequential KNN and MI with RF are more robust to missing data and have good potential for addressing attrition for binary outcome variables, especially at lower to moderate levels of missingness.

3.4. Prevalence based on Data Imputed by ML-Based Methods

Table 5 shows the prevalence estimates based on data imputed using missForest, sKNN and KNN under different levels of attrition assuming MAR. It also shows the amount of attrition bias when compared to the actual prevalence (that is, the estimate assuming attrition did not happen).

Table 5. Prevalence based on data imputed by ML-based methods under MAR mechanism.

Attrition/Methods	Prevalence/1000	σ	LCB (L)	UCB (U)	U-L	ζ
0% (no attrition)	9.40	0.41	8.60	10.20	1.60	-
missForest						
5%	9.17	0.40	8.38	9.96	1.58	0.23
10%	9.11	0.40	8.33	9.90	1.57	0.29
20%	9.36	0.41	8.57	10.16	1.59	0.04
30%	9.26	0.40	8.47	10.05	1.58	0.14
40%	9.04	0.40	8.26	9.83	1.57	0.36
50%	8.53	0.39	7.77	9.29	1.52	0.87
sKNN						
5%	9.35	0.41	8.55	10.14	1.59	0.05
10%	9.42	0.41	8.62	10.22	1.60	0.02
20%	9.56	0.41	8.76	10.36	1.60	0.16
30%	9.38	0.41	8.58	10.18	1.60	0.02
40%	9.60	0.41	8.79	10.40	1.61	0.20
50%	9.44	0.41	8.64	10.23	1.59	0.04
KNN						
5%	9.36	0.41	8.57	10.16	1.59	0.04
10%	9.42	0.41	8.62	10.22	1.60	0.02
20%	9.47	0.41	8.67	10.27	1.60	0.07
30%	9.24	0.40	8.45	10.03	1.58	0.16
40%	9.45	0.41	8.65	10.25	1.60	0.05
50%	9.29	0.40	8.50	10.09	1.59	0.11
MI with RF						
5%	9.31	0.41	8.52	10.10	1.58	0.09
10%	9.49	0.41	8.69	10.29	1.60	0.09
20%	9.19	0.40	8.40	9.98	1.58	0.21

Continued

30%	8.95	0.40	8.18	9.73	1.55	0.45
40%	9.06	0.40	8.28	9.85	1.57	0.34
50%	9.04	0.40	8.26	9.83	1.57	0.36

Attrition bias for prevalence estimate ranged from 0.04 to 0.87 across different levels of attrition when data were imputed using missForest, from 0.02 to 0.20 when imputed using sKNN and from 0.02 to 0.11 when using KNN. Standard error remained unchanged from the original value of 0.41 across all the four methods implying strong precision of the estimates. From these results, KNN showed the lowest bias overall, followed closely by sKNN, suggesting they have better performance in preserving the true prevalence estimate. MI with RF performed better than missForest at higher levels of attrition but worse than KNN and sKNN, particularly at 30% attrition and above, where its bias exceeded 0.3.

When compared with the conventional methods, sKNN and KNN had similar performance with MI, but better than IPW as far as analysis of the prevalence estimate is concerned. KNN and sKNN outperformed both missForest and MI with RF in terms of minimizing attrition bias in the prevalence estimate. The standard error remained consistent across all methods, indicating stable precision. Based on bias alone, KNN and sKNN outperform missForest and MI with RF, particularly at higher attrition levels. Therefore, KNN and sKNN are the most reliable in recovering prevalence estimates with minimal bias.

Figure 2 visualizes the performance of the both machine-learning based models and conventional models.

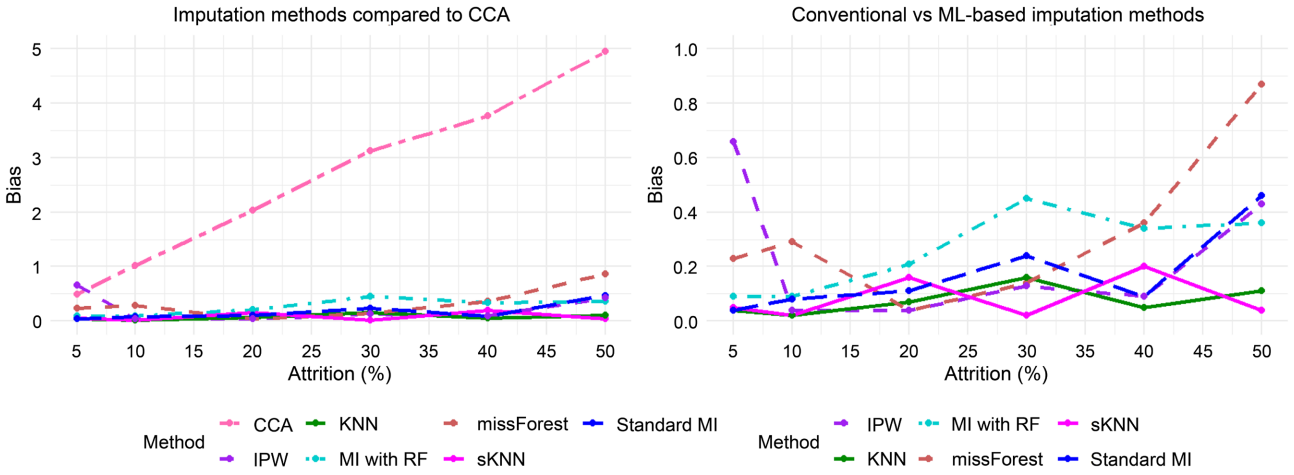

Figure 2. Attrition bias on prevalence estimation based on various missing data methods.

3.5. Logistic Regression Model Based on Data Imputed by ML-Based Methods

Logistic regression model is used to determine association between socio-demographic characteristics and prevalence under complete case analysis and

when attrition is accounted for using ML-based imputation methods; missForest, sKNN and KNN. Table 6 presents estimates comparing the bias in odds ratio estimates when complete case is used, and when attrition is accounted for by the ML-based methods, assuming both MAR and MNAR.

Table 6. Attrition bias on the log odds of the covariates in a logistic regression model for data imputed using ML-based imputation methods max.

	missForest						sKNN						KNN						MI with RF					
λ	5%	10%	20%	30%	40%	50%	5%	10%	20%	30%	40%	50%	5%	10%	20%	30%	40%	50%	5%	10%	20%	30%	40%	50%
MAR																								
Site (Ref = Viwandani)																								
Korogocho	0.032	0.045	0.001	0.022	0.151	0.276	0.005	0.009	0.042	0.010	0.028	0.106	0.010	0.001	0.020	0.017	0.057	0.139	0.000	0.022	0.008	0.051	0.129	0.083
Sex (Ref = Male)																								
Female	0.023	0.003	0.023	0.008	0.034	0.034	0.025	0.025	0.030	0.004	0.010	0.032	0.021	0.010	0.034	0.003	0.019	0.011	0.077	0.039	0.038	0.020	0.081	0.105
Age in years (Ref = 0 - 5 years)																								
6 – 12 y	0.062	0.102	0.035	0.219	0.216	0.280	0.010	0.003	0.001	0.015	0.018	0.114	0.009	0.028	0.076	0.160	0.156	0.313	0.027	0.014	0.027	0.045	0.022	0.073
13 - 18 y	0.047	0.042	0.057	0.242	0.136	0.069	0.005	0.045	0.036	0.023	0.076	0.223	0.005	0.071	0.041	0.129	0.177	0.294	0.033	0.034	0.076	0.110	0.015	0.042
19 - 28 y	0.023	0.052	0.048	0.215	0.155	0.229	0.002	0.007	0.033	0.017	0.016	0.130	0.002	0.047	0.090	0.178	0.205	0.269	0.008	0.016	0.103	0.028	0.086	0.073
29 - 49 y	0.016	0.038	0.105	0.037	0.029	0.018	0.023	0.056	0.080	0.177	0.217	0.333	0.017	0.076	0.159	0.319	0.346	0.397	0.011	0.018	0.056	0.009	0.031	0.010
40 y or older	0.019	0.044	0.064	0.347	0.306	0.296	0.001	0.008	0.043	0.041	0.074	0.050	0.000	0.020	0.033	0.005	0.013	0.114	0.041	0.039	0.156	0.051	0.059	0.035
MNAR																								
Site (Ref = Viwandani)																								
Korogocho	0.002	0.002	0.071	0.099	0.157	0.106	0.018	0.039	0.022	0.035	0.050	0.007	0.014	0.004	0.039	0.094	0.131	0.081	0.014	0.033	0.009	0.030	0.105	0.001
Sex (Ref = Male)																								
Female	0.036	0.061	0.300	0.105	0.115	0.167	0.067	0.038	0.035	0.046	0.014	0.005	0.070	0.065	0.033	0.032	0.063	0.016	0.083	0.033	0.013	0.099	0.151	0.035
Age in years (Ref = 0 - 5 years)																								
6 - 12 y	0.268	0.012	0.212	0.070	0.052	0.055	0.375	0.048	0.047	0.043	0.042	0.045	0.392	0.347	0.376	0.380	0.382	0.379	0.236	0.252	0.333	0.344	0.219	0.268
13 - 18 y	0.267	0.042	0.026	0.079	0.078	0.108	0.375	0.050	0.036	0.044	0.045	0.040	0.393	0.405	0.410	0.418	0.422	0.416	0.237	0.128	0.328	0.279	0.189	0.349
19 - 28 y	0.268	0.108	0.152	0.037	0.045	0.098	0.372	0.056	0.070	0.162	0.120	0.123	0.390	0.373	0.424	0.495	0.476	0.479	0.234	0.122	0.191	0.199	0.146	0.198
29 - 49 y	0.270	0.113	0.291	0.020	0.062	0.081	0.376	0.059	0.058	0.105	0.089	0.065	0.395	0.378	0.373	0.416	0.477	0.424	0.228	0.120	0.190	0.173	0.090	0.245
50 y or older	0.274	0.117	0.296	0.054	0.022	0.033	0.382	0.053	0.054	0.051	0.056	0.022	0.400	0.379	0.380	0.378	0.379	0.257	0.227	0.124	0.191	0.139	0.054	0.117

Overall, sKNN and KNN showed largely similar performance with MI when estimating the log odds under MAR. For example, the bias in the log odds for the sex variable ranged from 0.003 to 0.034 for missForest, 0.004 to 0.032 for sKNN, and 0.003 to 0.034 for KNN, which, though slightly higher, is comparable with MI (0.002 to 0.025). Bias for the sex variable ranged from 0.020 to 0.105 for MI with random forest and 0.004 to 0.081 for IPW. For age groups and site variables, biases were generally low across all methods under MAR.

When data were MNAR, sKNN (bias range: 0.005 - 0.070) and KNN (bias

range: 0.013 - 0.099) consistently provided slightly better estimates for sex across different attrition levels compared to MI (0.014 - 0.343), MI with random forest (0.035 - 0.151), and IPW (0.037 - 0.930), where bias tended to increase markedly with higher attrition. For age categories under MNAR, all ML-based methods, including missForest, MI with RF, sKNN, and KNN, showed substantial bias increases as attrition rose. The site variable estimates were less affected overall but showed some variation across methods. This suggests that ML-based nearest neighbor methods might offer an alternative approach to MI for addressing missing data where the assumption of MAR can not be guaranteed.

4. Discussion and Conclusion

In summary, we found that sKNN and KNN performed similarly to MI in terms of estimating prevalence under both MAR and MNAR. Based on the logistic regression model under MAR, sKNN and KNN performed comparably to MI. Both sKNN and KNN showed promising results when covariates were affected by MNAR compared to other models that we evaluated. This indicates the potential of ML-based methods to address the persistent challenge of MNAR when imputing missing data, though more research is still needed on this topic. Conversely, complete case analysis produced the most biased estimates under both MAR and MNAR. Complete case analysis only produces unbiased estimates when data are missing completely at random (MCAR) [6], but MCAR is rare in practice [14]. IPW performed similarly to MI under MAR but exhibited larger bias than MI for MNAR data.

The ideal practice to account for attrition is to design studies that minimize attrition. In clinical settings for example, this can be achieved through strategies such as targeted mobilization to improve response rates and scheduling favorable appointment dates for patients. However, for longitudinal or multi-stage cross-sectional studies, attrition is often inevitable but can be minimized and accounted for during analysis, as demonstrated in our study. As shown, on average, bias increased with an increase in the proportion of attrition.

Recent studies have compared common methods used for missing data including CCA, IPW and MI, and found that while IPW and MI are better than CCA, MI is more favourable [6] [34], especially when missing data is 5% or more. We have extended this work by comparing the three conventional methods for handling attrition with machine learning based imputation methods, which are gaining popularity in research as the development of data science methodologies continues to advance. The key advantage of using machine learning models is that they provide flexibility, address the complex non-linear interactions [20] [31] and provide internally cross-validated error estimates [35].

In addition, our findings regarding the limitations of CCA and IPW are corroborated by recent studies. Zhou *et al.* [36] conducted a comprehensive review of missing data techniques and found that CCA consistently produced the most biased estimates unless the data were MCAR. Furthermore, their analysis

showed that while IPW can perform well under MAR, it tends to introduce significant bias under MNAR scenarios. These findings resonate with our results, where CCA and IPW showed substantial limitations compared to ML-based methods. This growing body of literature emphasizes the need for careful method selection based on the specific missing data mechanism and highlights the potential of machine learning approaches in improving the accuracy of epidemiological and clinical research outcomes.

Several recent studies have highlighted the advantages of ML-based methods over traditional methods [37]. The ability of ML methods to adapt to different data structures and missingness mechanisms makes them particularly attractive alternatives for handling the attrition often encountered in longitudinal and population-based studies. Our study demonstrates that ML-based methods such as sKNN and KNN can provide reliable and less biased estimates compared to traditional methods.

In conclusion, our study demonstrates that even a small attrition proportion of 5% can significantly bias estimates if not properly addressed. Our findings indicate that sKNN and KNN perform similarly to MI under MAR and outperform IPW and MI in some scenarios under MNAR. This suggests that ML-based methods are viable alternatives to MI in various situations. While our findings may not be generalized to real-world MNAR data where the mechanism is unknown, our findings show that ML-based methods may have potential in addressing this persistent challenge. Multiple imputations with random forest did not perform differently from those with missForest. This could mean that random forest may not be well suited for our dataset, indicating the need for researchers to first evaluate which models work best for their data under study before selecting the appropriate method to use. It is advisable to avoid using CCA in the presence of any level of attrition. As noted by [14], attrition is rarely random, and one should assume that attrition is MNAR and make efforts at the study design stage to maximize response rates as much as possible.

Our study underscores the importance of using appropriate methods for accounting for attrition in population-based studies. While MI and IPW have been widely used, ML-based methods offer promising alternatives, particularly in dealing with situations where MAR is not plausible. An examination of different methods for accounting for attrition is necessary before settling on one because the underlying assumptions may be data-specific. Future research should continue to explore the potential of these advanced methods in various study designs and contexts. For instance, their application to rare outcomes which tend to have imbalanced outcome classes. For instance, in diseases like epilepsy, <1% of the population could screen positive for a disease, which results in having imbalanced outcome classes. The research considerations could include the development and integration of ML-based imputation algorithms within the robust MI frameworks to improve the accuracy of prediction and incorporating them in common statistical software to allow for their wider application, especially as computational capabilities

continue to improve.

Acknowledgements

The authors acknowledge the data collection team at both stages of the study and the Nairobi City County Health Department leadership for allowing the team to use the public health facilities in Nairobi to conduct the assessments. The team also acknowledges the Epilepsy Pathway Innovation in Africa (EPInA) scientific committee for the support of this study and the whole EPInA team.

Ethical Consideration

The study was approved by Scientific Ethics Review Unit (SERU) at the Kenya Medical Research Institute (KEMRI) (Reference Number: KEMRI/RES/7/3/1).

Informed consent

Written informed consent was obtained from all study participants.

Author Contribution

Daniel M. Mwanga: Conceptualization, Methodology, Data curation, Formal analysis, Writing: review and editing, Writing: original draft, Project administration. Isaac C. Kipchirchir: Supervision, Conceptualization, Methodology, Writing: review and editing. George O. Muhua: Supervision, Conceptualization, Methodology, Writing: review and editing. Charles R. Newton: Funding acquisition, Supervision, Conceptualization, Methodology, Writing: review and editing. Damazo T. Kadengye: Funding acquisition, Supervision, Conceptualization, Methodology, Writing: review and editing.

Funding Statement

This research was commissioned by the National Institute for Health Research (grant number NIHR200134) using Official Development Assistance (ODA) funding. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care.

Conflicts of Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] Ngugi, A.K., Bottomley, C., Chengo, E., Kombe, M.Z., Kazungu, M., Bauni, E., *et al*. (2012) The Validation of a Three-Stage Screening Methodology for Detecting Active Convulsive Epilepsy in Population-Based Studies in Health and Demographic Surveillance Systems. *Emerging Themes in Epidemiology*, **9**, Article No. 8. https://doi.org/10.1186/1742-7622-9-8

[2] Mwanga, D.M., Kadengye, D.T., Otieno, P.O., Wekesah, F.M., Kipchirchir, I.C., Muhua, G.O., *et al*. (2024) Prevalence of All Epilepsies in Urban Informal Settlements in Nairobi, Kenya: A Two-Stage Population-Based Study. *The Lancet Global Health*, **12**, E1323-E1330. https://doi.org/10.1016/s2214-109x(24)00217-1

[3] Ngugi, A.K., Bottomley, C., Kleinschmidt, I., Wagner, R.G., Kakooza-Mwesige, A., Ae-Ngibise, K., *et al*. (2013) Prevalence of Active Convulsive Epilepsy in Sub-Saharan Africa and Associated Risk Factors: Cross-Sectional and Case-Control Studies. *The Lancet Neurology*, **12**, 253-263. https://doi.org/10.1016/s1474-4422(13)70003-6

[4] Kariuki, S.M., Ngugi, A.K., Kombe, M.Z., Kazungu, M., Chengo, E., Odhiambo, R., *et al*. (2021) Prevalence and Mortality of Epilepsies with Convulsive and Non-Convulsive Seizures in Kilifi, Kenya. *Seizure*, **89**, 51-55. https://doi.org/10.1016/j.seizure.2021.04.028

[5] Kadengye, D.T., Ceulemans, E. and Van den Noortgate, W. (2013) Direct Likelihood Analysis and Multiple Imputation for Missing Item Scores in Multilevel Cross-Classification Educational Data. *Applied Psychological Measurement*, **38**, 61-80. https://doi.org/10.1177/0146621613491138

[6] Little, R.J., Carpenter, J.R. and Lee, K.J. (2022) A Comparison of Three Popular Methods for Handling Missing Data: Complete-Case Analysis, Inverse Probability Weighting, and Multiple Imputation. *Sociological Methods & Research*, **53**, 1105-1135. https://doi.org/10.1177/00491241221113873

[7] Rubin, D.B. (1976) Inference and Missing Data. *Biometrika*, **63**, 581-592. https://doi.org/10.1093/biomet/63.3.581

[8] Carpenter, J.R. and Smuk, M. (2021) Missing Data: A Statistical Framework for Practice. *Biometrical Journal*, **63**, 915-947. https://doi.org/10.1002/bimj.202000196

[9] Sterne, J.A.C., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., *et al*. (2009) Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls. *BMJ*, **338**, b2393. https://doi.org/10.1136/bmj.b2393

[10] Jakobsen, J.C., Gluud, C., Wetterslev, J. and Winkel, P. (2017) When and How Should Multiple Imputation Be Used for Handling Missing Data in Randomised Clinical Trials—A Practical Guide with Flowcharts. *BMC Medical Research Methodology*, **17**, Article No. 162. https://doi.org/10.1186/s12874-017-0442-1

[11] Little, R.J., D'Agostino, R., Cohen, M.L., Dickersin, K., Emerson, S.S., Farrar, J.T., *et al*. (2012) The Prevention and Treatment of Missing Data in Clinical Trials. *New England Journal of Medicine*, **367**, 1355-1360. https://doi.org/10.1056/nejmsr1203730

[12] Morris, T.P., Kahan, B.C. and White, I.R. (2014) Choosing Sensitivity Analyses for Randomised Trials: Principles. *BMC Medical Research Methodology*, **14**, Article No. 11. https://doi.org/10.1186/1471-2288-14-11

[13] Clark, T.G. and Altman, D.G. (2003) Developing a Prognostic Model in the Presence of Missing Data. *Journal of Clinical Epidemiology*, **56**, 28-37. https://doi.org/10.1016/s0895-4356(02)00539-5

[14] Kristman, V., Manno, M. and Côté, P. (2004) Loss to Follow-up in Cohort Studies: How Much Is Too Much? *European Journal of Epidemiology*, **19**, 751-760. https://doi.org/10.1023/b:ejep.0000036568.02655.f8

[15] Seaman, S.R., White, I.R., Copas, A.J. and Li, L. (2012) Combining Multiple Imputation and Inverse-Probability Weighting. *Biometrics*, **68**, 129-137. https://doi.org/10.1111/j.1541-0420.2011.01666.x

[16] Lee, K.J., Carlin, J.B., Simpson, J.A. and Moreno-Betancur, M. (2023) Assumptions and Analysis Planning in Studies with Missing Data in Multiple Variables: Moving

Beyond the MCAR/MAR/MNAR Classification. *International Journal of Epidemiology*, **52**, 1268-1275. https://doi.org/10.1093/ije/dyad008

[17] Gachau, S., Quartagno, M., Njagi, E.N., Owuor, N., English, M. and Ayieko, P. (2020) Handling Missing Data in Modelling Quality of Clinician-Prescribed Routine Care: Sensitivity Analysis of Departure from Missing at Random Assumption. *Statistical Methods in Medical Research*, **29**, 3076-3092. https://doi.org/10.1177/0962280220918279

[18] Kuhn, M. (2008) Building Predictive Models in R Using the Caret Package. *Journal of Statistical Software*, **28**, 1-26. https://doi.org/10.18637/jss.v028.i05

[19] RCoreTeam, R. (2013) A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/

[20] Stekhoven, D.J. and Bühlmann, P. (2012) Missforest—Non-Parametric Missing Value Imputation for Mixed-Type Data. *Bioinformatics*, **28**, 112-118. https://doi.org/10.1093/bioinformatics/btr597

[21] National Institute of Health Research (NIHR) (2020) Research and Innovation for Global Health Epilepsy Pathway Innovation in Africa (EPInA). https://epina.web.ox.ac.uk/

[22] Beguy, D., Elung'ata, P., Mberu, B., Oduor, C., Wamukoya, M., Nganyi, B., *et al*. (2015) Health & Demographic Surveillance System Profile: The Nairobi Urban Health and Demographic Surveillance System (NUHDSS). *International Journal of Epidemiology*, **44**, 462-471. https://doi.org/10.1093/ije/dyu251

[23] Emina, J., Beguy, D., Zulu, E.M., Ezeh, A.C., Muindi, K., Elung'ata, P., *et al*. (2011) Monitoring of Health and Demographic Outcomes in Poor Urban Settlements: Evidence from the Nairobi Urban Health and Demographic Surveillance System. *Journal of Urban Health*, **88**, 200-218. https://doi.org/10.1007/s11524-011-9594-1

[24] Placencia, M., Sander, J.W.A.S., Shorvon, S.D., Ellison, R.H. and Cascante, S.M. (1992) Validation of a Screening Questionnaire for the Detection of Epileptic Seizures in Epidemiological Studies. *Brain*, **115**, 783-794. https://doi.org/10.1093/brain/115.3.783

[25] Rubin, D.B. and Schenker, N. (1991) Multiple Imputation in Health-Are Databases: An Overview and Some Applications. *Statistics in Medicine*, **10**, 585-598. https://doi.org/10.1002/sim.4780100410

[26] Little, R.J.A. and Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Inc.

[27] Batista, G.E.A.P.A., Prati, R.C. and Monard, M.C. (2004) A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD Explorations Newsletter*, **6**, 20-29. https://doi.org/10.1145/1007730.1007735

[28] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., *et al*. (2001) Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics*, **17**, 520-525. https://doi.org/10.1093/bioinformatics/17.6.520

[29] Kigo, S.N., Omondi, E.O. and Omolo, B.O. (2023) Assessing Predictive Performance of Supervised Machine Learning Algorithms for a Diamond Pricing Model. *Scientific Reports*, **13**, Article No. 17315. https://doi.org/10.1038/s41598-023-44326-w

[30] Jones, G.D., Kariuki, S.M., Ngugi, A.K., Mwesige, A.K., Masanja, H., Owusu-Agyei, S., Wagner, R., Cross, J.H., Sander, J.-S., Newton, C.R., *et al*. (2023) Development and Validation of a Diagnostic Aid for Convulsive Epilepsy in Sub-Saharan Africa: A Retrospective Case-Control Study. *The Lancet Digital Health*, **5**, e185-e193.

[31] Mensah, J.A., Nortey, E.N.N., Ocran, E., Iddi, S. and Asiedu, L. (2024) De-Occlusion

and Recognition of Frontal Face Images: A Comparative Study of Multiple Imputation Methods. *Journal of Big Data*, **11**, Article No. 60. https://doi.org/10.1186/s40537-024-00925-6

[32] Manning, C.D., Raghavan, P. and Schütze, H. (2008) Introduction to Information Retrieval. Cambridge University Press. https://doi.org/10.1017/cbo9780511809071

[33] Mandrekar, J.N. (2010) Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, **5**, 1315-1316. https://doi.org/10.1097/jto.0b013e3181ec173d

[34] Cai, J., Zeng, D., Li, H., Butera, N.M., Baldoni, P.L., Maitra, P., *et al*. (2023) Comparisons of Statistical Methods for Handling Attrition in a Follow-up Visit with Complex Survey Sampling. *Statistics in Medicine*, **42**, 1641-1668. https://doi.org/10.1002/sim.9692

[35] Waljee, A.K., Mukherjee, A., Singal, A.G., Zhang, Y., Warren, J., Balis, U., *et al*. (2013) Comparison of Imputation Methods for Missing Laboratory Data in Medicine. *BMJ Open*, **3**, e002847. https://doi.org/10.1136/bmjopen-2013-002847

[36] Zhou, Y., Aryal, S. and Bouadjenek, M.R. (2024) Review for Handling Missing Data with Special Missing Mechanism. arXiv:2404.04905.

[37] Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B. and Tabona, O. (2021) A Survey on Missing Data in Machine Learning. *Journal of Big Data*, **8**, Article No. 140. https://doi.org/10.1186/s40537-021-00516-9