Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2020, 8, 110-133
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Using Statistical Learning to Treat Missing Data: A Case of HIV/TB Co-Infection in Kenya

Joshua O. Mwaro, Linda Chaba, Collins Odhiambo*

Strathmore Institute of Mathematical Sciences, Strathmore University, Nairobi, Kenya
Email: *codhiambo@strathmore.edu

How to cite this paper: Mwaro, J.O., Chaba, L. and Odhiambo, C. (2020) Using Statistical Learning to Treat Missing Data: A Case of HIV/TB Co-Infection in Kenya. *Journal of Data Analysis and Information Processing*, **8**, 110-133.
https://doi.org/10.4236/jdaip.2020.83007

Received: April 29, 2020
Accepted: July 28, 2020
Published: July 31, 2020

Abstract

In this study, we investigate the effects of missing data when estimating HIV/TB co-infection. We revisit the concept of missing data and examine three available approaches for dealing with missingness. The main objective is to identify the best method for correcting missing data in TB/HIV Co-infection setting. We employ both empirical data analysis and extensive simulation study to examine the effects of missing data, the accuracy, sensitivity, specificity and train and test error for different approaches. The novelty of this work hinges on the use of modern statistical learning algorithm when treating missingness. In the empirical analysis, both HIV data and TB-HIV co-infection data imputations were performed, and the missing values were imputed using different approaches. In the simulation study, sets of 0% (Complete case), 10%, 30%, 50% and 80% of the data were drawn randomly and replaced with missing values. Results show complete cases only had a co-infection rate (95% Confidence Interval band) of 29% (25%, 33%), weighted method 27% (23%, 31%), likelihood-based approach 26% (24%, 28%) and multiple imputation approach 21% (20%, 22%). In conclusion, MI remains the best approach for dealing with missing data and failure to apply it, results to overestimation of HIV/TB co-infection rate by 8%.

Keywords

Missing Data, HIV/TB Co-Infection, Imputation, Missing at Random, Count Data

1. Introduction

Dealing with missing data remains a major challenge in any field. It is therefore important to carefully examine any given data to identify the type and pattern of missingness and apply the correct method in a given setting. Missing data always occur in research, they undermine the validity of research findings yet this is overlooked by most researchers. Most of the researchers understand the existence of missing data but never report how they dealt with it, some address the issue by assuming that data is missing completely at random (MCAR, which rarely happens in practice) and apply ad hoc approaches (Listwise deletion, Pairwise Deletion, Single imputation) during analysis.

A few researchers go further to apply Imputation methods, advanced methods like Multiple Imputation and Maximum Likelihood, but tend to apply them across all the different missing data mechanisms. All these methods of dealing with missing data have not applied to HIV/TB co-infection data setting. Here we seek to identify the best method to correct for missingness under different mechanisms in HIV/TB co-infection setting.

TB is the leading comorbidity disease among PLWHIV as well as the leading cause of morbidity among HIV patients. Globally, HIV and TB are the leading infectious diseases that cause death. This makes it important to have valid and accurate statistics about the co-infection rates of these killer diseases.

An important statistic is the prevalence rate of the co-infection. Globally, there are wide variations from the true to the reported prevalence rates both in-country and between countries. One of the reasons attributed to the variations is under-reporting which results from missing data.

Missing data is mainly caused by irregular collection of information from HIV patients, no show of the patients for the baseline or consecutive checkup and nondisclosure of some information. Most of the researchers in public health address the missing data problem by using the default complete case analysis available in the analysis software and/or single imputation. The two techniques are commonly applied without considering the mechanism under which data are missing thus yielding biasses and loss of power in the study.

We therefore need to identify the best method to correct missing data for each of the missing data mechanism (MCAR, MAR and MNAR). As we do this, we are considering varying proportions of missing data as well as varying sample sizes.

The main objective of this work is to identify the best method of correcting missing data in TB/HIV co-infection setting. To do this, we systematically assess the following methods of dealing with missing data: Complete case, Mean/Single imputation method, Maximum Likelihood Estimation, Multiple Imputation.

Background

Generally, missing data fail to fulfill the MAR assumption if cases with missing data on a particular variable tend to have lower/higher values on that variable than those with observed data, controlling for other observed variables [1].

Data is MCAR if the likelihood of missingness does not depend on the observed data Y_{obs}, or on the missing data Y_{mis}.

The other mechanism is MNAR, which occurs when the probability of missing data on a variable depends on the value of the variable itself. An example is being likely to have missing data on HIV/TB co-infection if HIV TB co-infected clients who do not attend clinics are more likely not to provide information as compared to those who attend clinics.

Let our data set have p variables denoted: $Y_1, Y_2, \cdots, Y_p$. A data set is said to have a **univariate** pattern of missingness if the same records have missing data on one or more of the p variables.

A missing data pattern is **monotone** if the variables can be arranged in such a way that, when Y_j is missing, $Y_{j+1}, Y_{j+2}, \cdots, Y_p$ are missing as well. The monotone pattern frequently occurs in longitudinal studies after dropouts.

The **arbitrary** missing data pattern refers to when missing data occur in any variable for any participant in a random manner, this is computationally harder to handle than the other two patterns [2] [3].

Due to the serious consequences of missing data, ways have been developed to solve the issues of having missing data. Missing data solutions can be classified into both simple and advanced. Unfortunately most researchers go for the simple solutions (ad hoc techniques) which [4] indicate that do not always work.

The most common method which is always the default way in analysis softwares is complete case analysis. Complete case analysis can be classified into Listwise deletion and pairwise deletion.

Listwise deletion.

This refers to the removal of records with missing data on any variable. The method is the easiest. However, it greatly reduces the sample size and assumes that the data is MCAR.

Pairwise deletion.

This is an improvement of listwise deletion. This method works with pairs of variables in that: it calculates the covariance estimates for cases with complete observations on both variables, *i.e.* cases are removed if they have missing data on any of the variables involved in the analysis [5]. Pairwise deletion can be problematic as it results into varying sample sizes for different variable combinations within the same data set.

Single imputation.

An improvement over the two methods is the single imputation. Here, missing data on a continuous variable are replaced with the mean of observed data for the variable while missing data on a categorical variable are replaced with the mode of observed data for the variable. The pitfall of single imputation is that it ignores the variability within cases.

The weighting of the available data to compensate for the missing data, or estimating means, variances and correlations, may lead to biased results because their main assumption is that the data is MCAR [6]. To correct for the bias, multiple imputation technique is recommended Multiple Imputation (MI). MI is described as a simulation based technique that replaces missing values with a number m of plausible values [7]. Each of the m complete data set is analyzed by use of the standard complete-data procedures and parameter estimates and their respective standard errors are obtained. The results are later combined to produce estimates (multiple imputation estimates), the confidence intervals incorporate missing-data uncertainty [8] outlined the three steps of Multiple imputation :

Step 1. Creation of plausible values.

MI creates $m > 1$ plausible value for the missing values. The procedure draws the plausible values from a distribution specifically modeled for a missing value. For a given incomplete variable *v*, an imputation model is constructed that regresses v on variables with complete data, e.g. $v_1, v_2, \cdots, v_k$, among individuals with the observed *v*. This results into m imputed complete datasets, the datasets are identical in observed values but differ in imputed values.

Step 2. Parameter estimation.

The standard analytical procedures are applied on to the imputed datasets as would have been on a complete data set. Variations are expected in the results due to the different imputed values.

Step 3. Pooling results.

Finally, the m parameter estimates are combined into an overall estimate and variance-covariance matrix using Rubin's rules [6]. The combined variance-covariance matrix incorporates the within-imputation (sampling variance) variance and the between-imputation (caused by missing data) variance. Mathematically:

Suppose $\hat{\theta}_j$ estimates the univariate or multivariate quantity of interest like regression coefficient obtained from the *j*th imputed data set and W_j the estimated variance of $\hat{\theta}_j$. The combined estimate $\hat{\theta}$ is given by:

$$\hat{\theta} = \frac{1}{m} \sum \sum_{j=1}^{m} \hat{\theta}_j . \tag{1}$$

within-imputation variance of $\hat{\theta}$ is:

$$W = \frac{1}{m} \sum_{j=1}^{m} W_j , \tag{2}$$

the between-imputation variance :

$$B = \frac{1}{m-1} \sum_{j=1}^{m} \left(\hat{\theta}_j - \hat{\theta} \right)^2 \tag{3}$$

and the total variance of $\hat{\theta}$ is given by combining both the within and between imputation variances:

$$Var\left(\hat{\theta}\right) = W + \left(1 + \frac{1}{m}\right) B. \tag{4}$$

This is proper imputation as it incorporates all sources of variability and uncertainty in the imputed values, it also includes prediction errors of the individual values and errors of estimation in the fitted coefficients of the imputation model.

Figure 1 illustrates multiple imputation steps. It shows imputation involving three (m) imputed data sets, the imputed data sets are stored in class

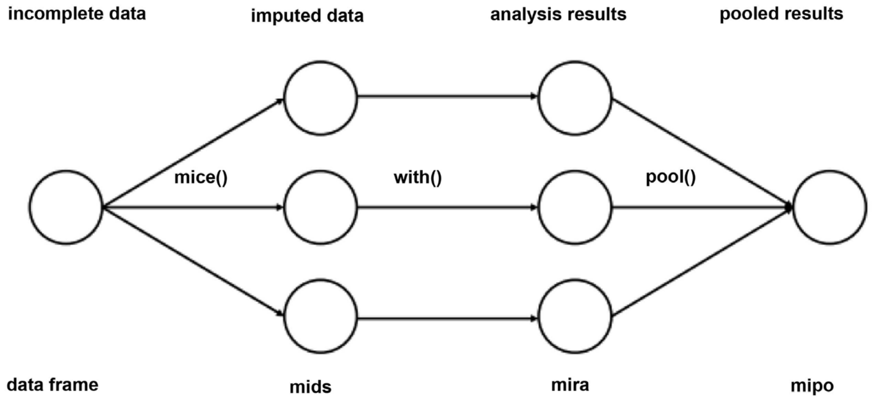


Figure 1. Multiple imputation steps. The imputation involving three (m) imputed data sets, and the imputed data sets are stored in class midst (multiply imputed dataset).

midst (multiply imputed dataset). The next step is an analysis of each of the imputed datasets using the function with(), analysis results are stored in class mira (multiply imputed repeated analysis). Finally the pool() function is used to pool results using Rubin's rules and store the pooled results in an object of class mipo (multiple imputed pooled outcomes) [9].

Here we defined statistical learning as a vast set of tools for understating data. The set of tools can be classified into supervised and unsupervised. Supervised statistical learning refers to a statistical learning method in which a statistical model is built to predict or estimate an output based on one or more inputs. Unsupervised statistical learning refers to a statistical learning method in which we have input(s) but no supervising output the main aim here is to learn relationships and structure from the data.

Let Y be a quantitative response with p predictors: $X_1, X_2, \cdots, X_p$. We assume that there is a relationship between Y and $X = X_1, X_2, \cdots, X_p$. The relationship between X and Y can be written as: $Y = f(X) + \in$. Where f is fixed but unknown function of $X_1, X_2, \cdots, X_p$ which represents the systematic information that X provides about Y and $\in$ is a random error term independent of X with mean zero. **Statistical learning** refers to a set of approaches for estimating f.

Reasons for estimating f.

1) Prediction.

This is motivated by the fact that inputs (X) are readily available but the output Y cannot be easily obtained. Now that $\in$ averages to zero, Y is predicted using: $\hat{Y} = \hat{f}(X)$.

Here, $\hat{f}$ is the estimate of f and $\hat{Y}$ is the resulting prediction for Y. We are not interested in the exact form of $\hat{f}$, as long as it results into accurate predictions for Y hence $\hat{f}$ is often treated as a black box.

Two quantities: reducible error and irreducible error determine the accuracy of $\hat{Y}$ as a prediction of Y.

Generally, $\hat{f}$ cannot be a perfect estimate of f, this introduces a reducible error. The error is reducible since we are able to improve the accuracy of $\hat{f}$

using the best method of statistical learning to estimate $f(\hat{Y} = \hat{f}(X))$

Our prediction would still have an irreducible error because Y is also a function of $\in$ which from the definition cannot be predicted by *X*. No matter how well we estimate *f*, we cannot reduce the error introduced by $\in$.

2) Inference.

Here, we are interested in understanding the way Y is affected by how X changes. We estimate f but our main goal is not making predictions of Y but understanding the relationship between *X* and *Y*, how *Y* changes as a function of *X*. Therefore, $\hat{f}$ cannot be treated as a black box as its exact form needs to be known.

Supervised and Unsupervised learning.

Under supervised learning, for each observation of the predictor measurement(s) $x_i, i = 1, 2, \cdots, n$, there is an associated response measurement y_i. Our model relates the response to the predictors with the aim of accurately predicting the response for future observations (prediction) or understanding the relationship between predictors an the response (inference).

For unsupervised learning, for every observation $i = 1, 2, \cdots, n$ we have a vector of measurements x_i but no associated response y_i. We are working blind because of lack of a response variable to supervise our analysis. A statistical tool that we can use in this setting is cluster analysis in which the goal is to ascertain, on the basis of $x_1, x_2, \cdots, x_n$, if the observations fall into relatively distinct groups.

Regression and Classification problems

Models with a quantitative response are referred to as regression models, while the models with a qualitative response are called classification models.

Least squares linear regression is used for regression models while logistic regression is used for classification models. In statistical learning methods predictor variable type is less important as long as any qualitative predictors are properly coded before analysis.

TB is the leading opportunistic infection and the leading cause of morbidity and mortality among HIV-infected persons. TB is a global health problem with yearly new infections of 9 million and around 2 million annual deaths. The most affected countries are in Africa which account for 85% of the global rates of infection. There are 22 high-burden countries (Kenya among the 22) that account for 80% of the worlds TB cases [10]. Approximately 30% of HIV-infected persons are estimated TB. Globally, 13 million individuals are co-infected with HIV and TB, 70% of the worldwide HIV TB co-infections live in sub-Saharan Africa [11]. HIV and Mycobacterium tuberculosis highly interact each increasing progression of the other. Treatment of TB is made difficult by frequent drug-interactions with highly active antiretroviral therapy (HAART) and adverse drug reactions are more common among HIV-infected patients. Active TB is over 8 times higher in HIV-positive than HIV-negative individuals in Africa. Active TB is the first sign of AIDS in HIV-infected persons. Both active TB and HIV accelerate the progression of the other with the former decreasing the number of CD4+

lymphocytes thereby increasing HIV viral replication and eventually shortening the lives of HIV-positive persons. The fatality rate of HIV-related TB is over 50% [12]. Further information on TB and HIV can be found at [13].

In Kenya, the [14] report indicates that 98% of TB patients were tested for HIV, 27% of the TB patients tested were found to be co-infected with HIV. This was close to the 28% 2017 co-infection rate [15]. The prevalence rates for the two years are questionable due to what was reported on strategies for finding missing people with TB. The key strategy was the Active Case Finding (ACF) which entails a systematic screening for TB among all patients presenting to health facilities regardless of whether they present with TB symptoms or not. The key motivation to our study comes from one of the challenges encountered during the ACF strategy implementation the challenge was called "System challenges" and quoted as: "incomplete documentation in the presumptive TB registers leading to leakage e.g. missing lab results was also noted during the implementation".

2. Materials and Method

2.1. Statistical Learning Algorithm

Let ψ be an observed quantitative response and η different predictors, $X_1, X_2, \cdots, X_\eta$. Assuming that there is some relationship between ψ and $X = (X_1, X_2, \cdots, X_\eta)$, which can be written in the form

$$\psi = f(X) + \epsilon. \tag{5}$$

where f is some fixed but unknown function of $X_1, X_2, \cdots, X_\eta$, and ϵ is a random error term, which is independent of X and has mean zero. In this formulation, f represents the systematic information that X provides about ψ.

Assuming that there is a true underlying parameter vector $\Omega \in \mathbb{R}^d$ which governs the outputs. For each $i = 1, 2, \cdots, n$:

$$\psi_i = X_i \Omega + \epsilon_i$$

The problem of statistical learning involves a set of approaches for estimating f as well as tools for evaluating the estimates obtained. The two main reasons for estimating f are for prediction and inference. For, prediction, a set of inputs X are readily available, but the output ψ cannot be easily obtained. In this setting, since the error term averages to zero, we can predict ψ using

$$\hat{\psi} = \hat{f}(X), \tag{6}$$

where $\hat{f}$ represents our estimate for f, and $\hat{\psi}$ represents the resulting prediction for ψ. In this setting, $\hat{f}$ is treated as a black box, in the sense that one is not typically concerned with the exact form of $\hat{f}$, provided that it yields accurate predictions for ψ.

At training segment, we observe one realization of $\psi_1, \psi_2, \cdots, \psi_n$. The data matrices are given as:

$$X = [X_1, X_1, \cdots, X_n]^{\mathrm{T}} \in \mathbb{R}^{n\times\eta}.$$

$$\epsilon = [\epsilon_1, \epsilon_2, \cdots, \epsilon_n]^{\mathrm{T}} \in \mathbb{R}^{\eta}.$$

$$\psi = \left[\psi_1, \psi_2, \cdots, \psi_n\right]^{\mathrm{T}} \in \mathbb{R}^{\eta}.$$

$$\Sigma = \frac{1}{n} X^{\mathrm{T}} X \in \mathbb{R}^{\eta \times \eta}.$$

The goal is to naturally minimize the expected risk.

2.2. Handling Missing Data

Here we employ the complete case scenario and compare with three methods *i.e.* Weighted Method, Maximum likelihood approach and MI.

2.2.1. Complete Case

By complete cases, we refer to available data analysis, or pairwise deletion, uses all available data to generate estimate of the parameters of interest. This approach is illustrated in Equation (2) where required statistics is generated on different sets of cases.

In pairwise deletion, all cases would be used to estimate the mean of ϕ, but only the complete cases would contribute to an estimate of Φ, and the correlation between ϕ and Φ. Different sets of cases are used to estimate parameters of interest in the data. When variables are highly correlated, available case analysis provides estimates that are inferior to complete case results [16].

Let ψ_i be a given i^{th} observation. Then

$$\Psi_{m,n} = \begin{pmatrix} \psi_{1,1} & \psi_{1,2} & \cdots & \psi_{1,n} \\ \psi_{2,1} & \psi_{2,2} & \cdots & \psi_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{m,1} & \psi_{m,2} & \cdots & \psi_{m,n} \end{pmatrix}$$

$$\Psi_{p,q} = \begin{pmatrix} \psi_{m1,1} & \psi_{m1,2} & \cdots & \psi_{1,1n} \\ \psi_{m2,1} & \psi_{m2,2} & \cdots & \psi_{2,2n} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{mp,1} & \psi_{mp,2} & \cdots & \psi_{mp,nq} \end{pmatrix}$$

where $\Psi_{m,n}$ are complete cases and $\Psi_{p,q}$ refers to incomplete cases. The estimates for the mean, standard error can be obtained as

$$\bar{\psi}_1 = \frac{\sum_i^n \psi_{1i}}{n}$$

$$\overline{s}.e(\psi) = \frac{\sum_i^n \left(\psi_{1i} - \bar{\psi}_1\right)^2}{n}$$

2.2.2. Weighted Method

In this approach, the missing value is replaced with a plausible figure, such as the mean for observed cases. This approach is repeated as if the data are completely observed. While this allows the inclusion of all cases in a standard analysis procedure, replacing missing values with a single value changes the distribution of that variable by decreasing the variance that is likely present.

Impute all missing values of ψ by (weighted) respondent mean of $\bar{\psi}$, if ψ

continuous. Impute by respondent mode if categorical variable. Define homogeneous classes and impute mean or mode within class Want ψ mean (or mode) same for respondents and missing values within classes. We may use segmentation algorithm to develop homogenous classes.

2.2.3. Maximum Likelihood Method

The principle here is when missing data occur; we base estimation on the likelihood of the observed data. The difficulty lies in specifying the likelihood of the observed data. Intrinsically, we use iterative solution, termed the EM algorithm, to find the estimate of a parameter (such as the means and covariance matrix) when closed form solutions to the maximization of a likelihood are not possible.

The basic idea is to maximize the conditional likelihood using the set of respondents, those with $\varphi_{iT} = 1$, for whom the response probability is reversed in the sense that, instead of the original probability, the conditional probability of $\varphi_{it-1} = 1$ given that $\varphi_{iT} = 1$ is considered. The conditional likelihood for the special case $T = 1$, *i.e.*, no followup.

The approach based on conditional likelihood consists of two steps. In the first step, the reverse conditional probability $q_{it} = pr\left(\varphi_{i,t} = 1 \mid \varphi_{i,t+1} = 1, \psi_i, \upsilon_i\right)$ is derived using Baye's formula from the assumed response model.

We can obtain $q_{it} = \dfrac{\Xi_{it}}{1+\Xi_{it}}$ where

$$\begin{aligned}\Xi_{it} &= \frac{pr\left(\varphi_{it} = 1 \mid \psi_i, \upsilon_i, \varphi_{i,t+1} = 1\right)}{pr\left(\varphi_{it} = 0 \mid \psi_i, \upsilon_i, \varphi_{i,t+1} = 1\right)} \\ &= \frac{pr\left(\varphi_{it} = 1, \varphi_{i,t+1} = 1 \mid \psi_i, \upsilon_i\right)}{pr\left(\varphi_{it} = 0, \varphi_{i,t+1} = 1 \mid \psi_i, \upsilon_i\right)} \\ &= \frac{pr\left(\varphi_{it} = 1 \mid \psi_i, \upsilon_i, \varphi_{i,t+1} = 1\right)}{pr\left(\varphi_{it} = 1 \mid \psi_i, \upsilon_i, \varphi_{i,t+1} = 0\right)} \frac{pr\left(\varphi_{it} = 1 \mid \psi_i, \upsilon_i\right)}{pr\left(\varphi_{it} = 1 \mid \psi_i, \upsilon_i\right)} \\ &= \frac{1}{p_{i,t+1}} \frac{\overline{\Upsilon}_{it}}{1-\overline{\Upsilon}_{it}}\end{aligned}$$

where $\overline{\Upsilon}_{it} = \sum_t^{j=1}\left[p_{ij}\prod_{j-}^{k=1}\left(1-p_{ij}\right)\right] = \sum_t^{j=1}\upsilon_j$. More of the parameter estimation can be found in [17].

multiple Imputation

Generally, there are three steps for executing MI *i.e.*

1) Imputation: Generate a set of $m > 1$ plausible values for $\theta^{mis} = \left(\tau^{mis}; \upsilon^{mis}\right)$

2) Analysis: Analyze the datasets using complete-case methods.

3) Combination: Combine the results from the m analyses.

Imputation step relies upon assumptions regarding the missing mechanism that generated the observed sample. The goal of the imputation is to account for the relationships between unobserved and observed variables, while considering the uncertainty of the imputation. The MAR assumption (which is generally assumed for many missing data methods, and as previously noted is intestable without additional information) is key to the validity of multiple imputation. Use of this assumption allows the researcher to generate imputations $\left(\theta^1,\theta^2,\cdots,\theta^m\right)$.

When missingness is monotone, simple methods are employed, including (for continuous variables) propensity methods, predictive mean matching, and (for discrete variables) discriminant analysis or logistic regression. However, for more complicated missingness, Markov Chain Monte Carlo (MCMC) approaches are applied. Both the predictive mean matching and MCMC approaches require assumptions of multivariate normality.

2.3. Simulation

We adopted the missing data methodology proposed by [18]. The methodology comprises of the following steps:

Step 1. Simulation of a multivariate, complete data set to be considered the population of interest.

Step 2. Making the dataset incomplete.

Step 3. Estimating the incomplete data by methods of correcting for missing data.

Step 4. Comparing the Statistical inferences obtained for the original, complete data set and after dealing with the missing values to get an indication of the performance of the missing data method. We then apply real HIV-TB coinfection data.

2.4. Step 1: Simulated Multivariate Complete Data Set

We reproduced a simulated data set "Default" in R. The "Default" dataset is found in ISLR package of R. The dataset contains 10,000 observations with four variables. The variables are: "default"; A factor with levels "No" and "Yes" indicating whether the customer defaulted on their debt, "student"; A factor with levels "No" and "Yes" indicating whether the customer is a student, "balance"; The average balance that the customer has remaining on their credit card after making their monthly payment and "income"; Income of customer [19].

The statistical inferences of interest are the Accuracy, Sensitivity, Specificity and AUC of the original complete dataset that we would compare with the statistical inferences of the original dataset after creating missingness (under MAR, MNAR and MCAR) and correcting for the missingness by: complete case analysis, single imputation, multiple imputation and maximum likelihood methods.

2.5. Step 2: Missing Data Amputation

The amputation (generation of missing values) was done using the "ampute" function which is available in mice package of R software, a multivariate amputation procedure as explained by [18] was used. Missing data were generated in three mechanisms: MCAR, MAR and MNAR with varying proportion of missingness (7%, 10%, 30%, 50% and 80%). The default patterns (combinations of variables with; coded 0 and without; coded 1 missing values) in ampute function of R was adopted, here each pattern had missingness on one variable only. Our frequency (a vector of length number of patterns containing the relative frequency with which the patterns should occur.) too was default (equal probability for each pattern) *i.e.* (0.25, 0.25, 0.25, 0.25). Specification of weights (a matrix/data frame of size number of patterns by number of variables, weights are used to calculate weighted sum of scores) was as follows: For MAR mechanism, we assigned weights of value zero to the variables that would be made incomplete. Whereas for MNAR mechanism, we assigned weights of value zero to the variables that would remain completely observed and weights of value one to the variables that would be made incomplete. The MCAR mechanism does not use weights.

2.6. Step 3: Correcting for the Missing Data

We started the procedure of correcting for missing values by first reconverting the categorical which had been converted (during amputation) into numerical into their original categorical form. Checked to ensure the data was coded correctly, identified missing values and patterns within each variable and graphically represented the missingness. Finally, we corrected for the missingness in each of the amputed (under the three mechanisms) datasets using each of these methods: Complete case analysis, Single imputation, Multiple imputation and the likelihood method.

2.7. Step 4: Comparing the Statistical Inferences Obtained for the Original, Complete Data Set and after Correcting for the Missing Values

We repeated the analysis in step 1 on each of the datasets that we had corrected for missingness. The goal was to obtain similar statistical inferences (Accuracy, Sensitivity, Specificity and AUC) for comparison with those obtained in the first step.

3. Results

Simulation Results

Prediction of "default" using "balance", "income" and "student" of the "Default" data set in R's ISLR package, resulted into 0.9726, 0.3058824, 0.9960663 and 0.9510306 as the Accuracy, Sensitivity, Specificity and AUC of our prediction model respectively. These are the standard inferences that we used to compare to get an indication of the performance of the missing data method. We generated missingness in the simulated "Default" data set of the ISLR package in R. The proportion of missingness ranged from 7% to 80%. Then used four approaches of dealing with missing data: Complete case analysis (List wise deletion), Mean/Single imputation, Multiple imputation and Maximum Likelihood Estimation Imputation method. We re-analyzed the data sets to obtain new: Accuracy, Sensitivity, Specificity and the area under the receiver-operator curve (AUC) for comparison with the complete case values.

Table 1 shows analysis where missing data are corrected for by List wise deletion; complete case analysis. Under MAR, We see lower (compared to Complete data) values of sensitivity which systematically decrease as the proportion of missingness increase. However, we observe inconsistencies in the AUC values. We are unable to obtain inferences for the highest proportion of missingness, this is attributed to the few observed cases which result into extremely few training and testing data sets. When the mechanism is MCAR, we observe irregular inference values, this is caused by the complete randomness in the amputation of the missing values.

Table 1 shows analysis when missing data are corrected for by Mean/Mode single imputation. We replaced missing values with mean and mode of observed cases for continuous and categorical variables respectively. There is a systematic decrease in the sensitivity and AUC values which are greatly lower than the

Table 1. Complete case analysis.

Listwise deletion: Complete Case Analysis (CCA)				
Missingness under MAR				
Percentage Missing	Accuracy	Sensitivity	Specificity	AUC
Original Complete Data	**0.9726**	**0.3058824**	**0.9960663**	**0.9510306**
7%	0.9862543	0.1666667	0.9991274	0.937443
10%	0.98733896	0.08333333	0.99954975	0.9465462
30%	0.99242424	0.01960784	1.000000	0.9642857
50%	0.9946667	0	0.999665	0.9107317
80%	-	-	-	0.9230769
Missingness under MNAR				
Percentage Missing	Accuracy	Sensitivity	Specificity	AUC
Original Complete Data	**0.9726**	**0.3058824**	**0.9960663**	**0.9510306**
7%	0.9783076	0.2586207	0.9966960	0.9570831
10%	0.9824522	0.2282609	0.9981859	0.95467
30%	0.9851515	0.2203390	0.9990744	0.9755696
50%	0.9840000	0.3064516	0.9982982	0.9683496
80%	0.9839519	0.1176471	0.9989796	0.9497348
Missingness under MCAR				
Percentage Missing	Accuracy	Sensitivity	Specificity	AUC
Original Complete Data	**0.9726**	**0.3058824**	**0.9960663**	**0.9510306**
7%	0.974348	0.316129	0.9969047	0.955583
10%	0.9724244	0.2792208	0.9968029	0.9501156
30%	0.9743954	0.3478261	0.9955882	0.9374005
50%	0.9742063	0.3956044	0.9958831	0.9388399
80%	0.9714286	0.2413793	0.9936909	0.9545671

complete data values. For MCAR, we observe a similar trend of systematic decrease in sensitivity and AUC, but values (sensitivity and AUC) for lower proportions (7% and 10%) of missingness are closer to the complete data values.

Table 2 shows analysis when missing data are corrected for by Single imputation. We used the MICE package in R to carry out multiple imputations. We imputed three data sets using all the variables ("default", "student", "balance" and "income"). Results indicate similar characteristics among the three mechanisms: the values for Accuracy, Specificity and AUC are very close to the complete data set values. The Sensitivity values are similar across the mechanism but slightly higher than the complete data set Sensitivity values.

Table 3 shows results when missing data are corrected for by Multiple Imputation method. Under MAR and MNAR: we observe fluctuations in the sensitivity values which are close to the sensitivity values of the complete data set. Fluctuations

Table 2. Mean/mode single imputation method.

Mean/Mode Single Imputation Method				
Missingness under MAR				
Percentage Missing	Accuracy	Sensitivity	Specificity	AUC
Original Complete Data	**0.9726**	**0.3058824**	**0.9960663**	**0.9510306**
7%	0.9694000	0.06962025	0.99876084	0.8638239
10%	0.9692000	0.04545455	0.99855551	0.8589932
30%	0.9726000	0.02919708	0.99917746	0.8333297
50%	0.9732000	0.02255639	0.99917814	0.8213744
80%	0.9742000	0.0234375	0.999179	0.8067571
Missingness under MNAR				
Percentage Missing	Accuracy	Sensitivity	Specificity	AUC
Original Complete Data	**0.9726**	**0.3058824**	**0.9960663**	**0.9510306**
7%	0.9748000	0.1461538	0.9969199	0.9123312
10%	0.9742000	0.1015625	0.9971264	0.9043917
30%	0.9748000	0.0320000	0.9989744	0.8552557
50%	0.97480000	0.02419355	0.99897457	0.8448201
80%	0.97500000	0.02419355	0.99917966	0.8192981
Missingness under MCAR				
Percentage Missing	Accuracy	Sensitivity	Specificity	AUC
Original Complete Data	**0.9726**	**0.3058824**	**0.9960663**	**0.9510306**
7%	0.9728000	0.297619	0.9962748	0.9414881
10%	0.9730000	0.297619	0.9964818	0.9415958
30%	0.9718000	0.21875	0.9966942	0.9363456
50%	0.9732000	0.1458333	0.9977348	0.9110347
80%	0.9738000	0.04580153	0.99876771	0.8360101

Table 3. Multiple imputation method.

Multiple Imputation Method				
Missingness under MAR				
Percentage Missing	Accuracy	Sensitivity	Specificity	AUC
Original Complete Data	**0.9726**	**0.3058824**	**0.9960663**	**0.9510306**
7%	0.9757333	0.3388773	0.9968317	0.9501267
10%	0.9742667	0.3190184	0.9963476	0.9509557
30%	0.9756000	0.3702970	0.9966885	0.9493033
50%	0.9751333	0.3447581	0.9966906	0.9542059
80%	0.9752667	0.3265306	0.9971744	0.9479768
Missingness under MNAR				
Percentage Missing	Accuracy	Sensitivity	Specificity	AUC
Original Complete Data	**0.9726**	**0.3058824**	**0.9960663**	**0.9510306**
7%	0.9757333	0.3388773	0.9968317	0.9501267
10%	0.9742667	0.3190184	0.9963476	0.9509557
30%	0.9756000	0.3702970	0.9966885	0.9493033
50%	0.9751333	0.3447581	0.9966906	0.9542059
80%	0.9752667	0.3265306	0.9971744	0.9479768
Missingness under MCAR				
Percentage Missing	Accuracy	Sensitivity	Specificity	AUC
Original Complete Data	**0.9726**	**0.3058824**	**0.9960663**	**0.9510306**
7%	0.9750000	0.3326572	0.9968291	0.9505351
10%	0.9748667	0.3360324	0.9966221	0.9499557
30%	0.9746667	0.3490196	0.9966874	0.9507702
50%	0.9761333	0.3382353	0.9970394	0.9552228
80%	0.9742000	0.3340000	0.9962759	0.9501424

are also witnessed in the AUC values but all the values are lower than the complete data AUC values.

Table 4 shows results when missing data are corrected for by MLE Imputation method. Under MAR and MNAR: we observe fluctuations in the sensitivity values which are close to the sensitivity values of the complete data set. The statistical analysis could not however, run missingness under MCAR.

CRUDE Co-infection rate:

We generated a crude co-infection rate as displayed in Table 5.

Plotting Missingness

Plot the missing values for HIV Patients Screened for TB

Plot the missing values for HIV/TB Coinfection

Results show in Table 6; complete cases only had a co-infection rate (95% Confidence Interval band) of 29% (25%, 33%), weighted method 27% (23%,

Table 4. MLE imputation method.

MLE Imputation Method				
Missingness under MAR				
Percentage Missing	Accuracy	Sensitivity	Specificity	AUC
Original Complete Data	**0.9726**	**0.3058824**	**0.9960663**	**0.9510306**
7%	0.9718000	0.2500000	0.9962779	0.9378383
10%	0.9736000	0.3173653	0.9962756	0.9460736
30%	0.9730000	0.3048780	0.9956576	0.9344419
50%	0.9736000	0.2926829	0.9966915	0.9283667
80%	0.9712000	0.2931034	0.9956486	0.9464979
Missingness under MNAR				
Percentage Missing	Accuracy	Sensitivity	Specificity	AUC
Original Complete Data	**0.9726**	**0.3058824**	**0.9960663**	**0.9510306**
7%	0.9718000	0.2500000	0.9962779	0.9378383
10%	0.9736000	0.3154762	0.9964818	0.9423608
30%	0.9738000	0.3372781	0.9960671	0.9418557
50%	0.9736000	0.2926829	0.9966915	0.9283667
80%	0.9712000	0.2937853	0.9960605	0.9338375
Missingness under MCAR				
Percentage Missing	Accuracy	Sensitivity	Specificity	AUC
Original Complete Data	**0.9726**	**0.3058824**	**0.9960663**	**0.9510306**
7%				
10%				
30%				
50%				
80%				

Table 5. CRUDE coinfection rate.

No. Tested	TB/HIV	HIV/TB Coinfection Rate
7,379,664	2,112,688	29% (LCI 25%, UCI 33%)

Table 6. Comparison of four approaches.

Approach	Estimated HIV/TB Co-infection Rate	95% LCI	95% UCI
Complete Cases Only	29%	25%	33%
Weighted Method	27%	23%	31%
Likelihood Based approach	26%	24%	28%
Multiple Imputation Approach	21%	20%	22%

31%), likelihood-based approach 26% (24%, 28%) and multiple imputation approach 21% (20%, 22%). In conclusion, MI remains the best approach for dealing with missing data and failure to apply it results to overestimation of HIV/TB co-infection rate by 8%. **Comparison of four approaches:**

Complete cases only

Results show a cyclic trend

Weighted Method

Shows a deep downward trend in the recent months

Likelihood Method

Shows an upward trend in the recent months

Multiple Imputation

Cyclic trend but generally lower

4. Discussion

The aim of this work was to revisit and review the topic of dealing with missing data in the context of estimating national HIV/TB co-infection. In addition, we wanted to assess the accuracy of MI using an example where models from imputed data could be compared with models derived from actual data using a modern approach *i.e.* statistical learning.

It is estimated that HIV/TB co-infection is 14 million globally [20], and TB remains the leading cause of death among PLHIV. HIV infection is estimated to increase the risk of TB 20-fold compared to HIV-seronegative individuals in high HIV-prevalence countries [13]. Of the estimated 8.7 million people who developed TB globally in 2012, 1.1 million (13%) were estimated to be HIV-coinfected. Of the 2.8 million people with TB who were screened for HIV in 2012, 20% tested HIV-positive, including 42% of people with TB in sub-Saharan Africa. More than 75% of the estimated HIV-positive incident TB cases live in just 10 countries (Ethiopia, India, Kenya, Mozambique, Nigeria, South Africa, United Republic of Tanzania, Uganda, Zambia, and Zimbabwe) [13]. The increased incidence of active TB in HIV-infected individuals can be attributed to at least two mechanisms: the increased reactivation of latent TB or increased susceptibility to miliary TB infection. The increased risk of active TB among HIV-infected persons was initially mainly attributed to an increased risk of reactivation of a latent infection.

With all these HIV/TB co-infection estimates at national and international levels ignoring missing data can have serious consequences in HIV programming contexts. Missing data are effectively ignored when using "complete case" analyses and the automated use of step-wise selection procedures often exacerbates this as cases can easily be excluded on the basis of data missing from variables not even part of a final estimation. Ignoring missing data causes problems when the data are not missing completely at random, as is likely for most missing data in HIV/TB setting. We demonstrated using extensive simulation procedure for how to deal with missing data with different settings and adjustments. This procedure can be used in any research area facing similar missing data problems involves identifying missing data (with descriptive statistics); investigate missing data patterns; define variables in the data set which may be related to missing values to be used for the imputation model; impute missing data to give "m" complete data sets; run the models of interest using the "m" imputed data sets; combine the "m" models' parameters; report the final model (as you would have done for any regression model).

The assumptions made by the Statistical Learning Theory framework include the future (*i.e.* test) observations are related to the past (*i.e.* training) ones, so that the feature is stationary. At the core of the theory is a probabilistic model of the phenomenon (or data generation process). Within this model (see Figures 2-9), the relationship between past and future observations is that they both are sampled independently from the same distribution (i.i.d.). The independence assumption means that each new observation yields maximum information. The identical distribution means that the observations give information about the underlying phenomenon (here a probability distribution). An immediate consequence of this very general setting is that one can construct algorithms (e.g.

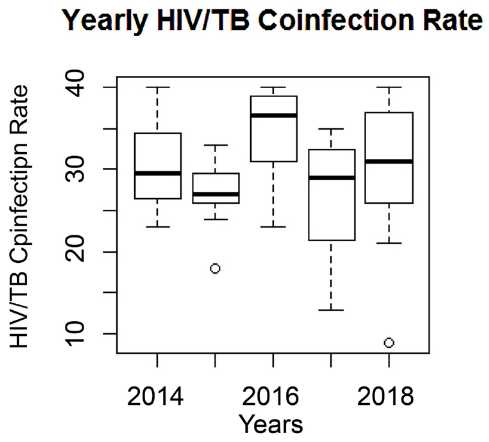


Figure 2. Box plots of yearly HIV/TB coinfection

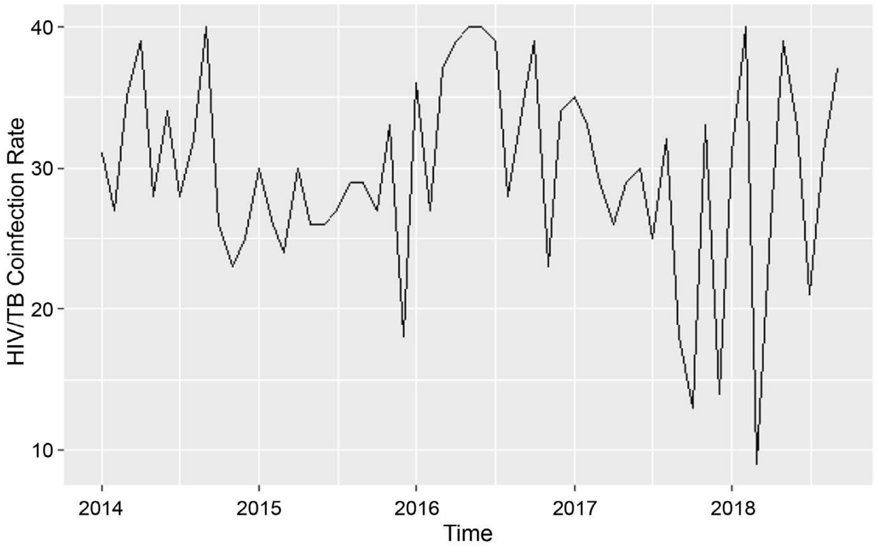


Figure 3. Line graph of yearly HIV/TB coinfection.

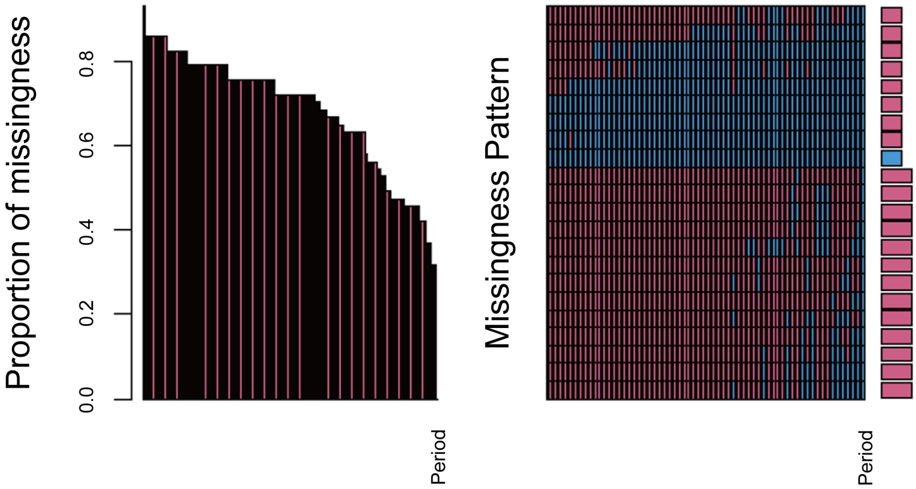


Figure 4. Missingness patterns for HIV patients screened for TB.

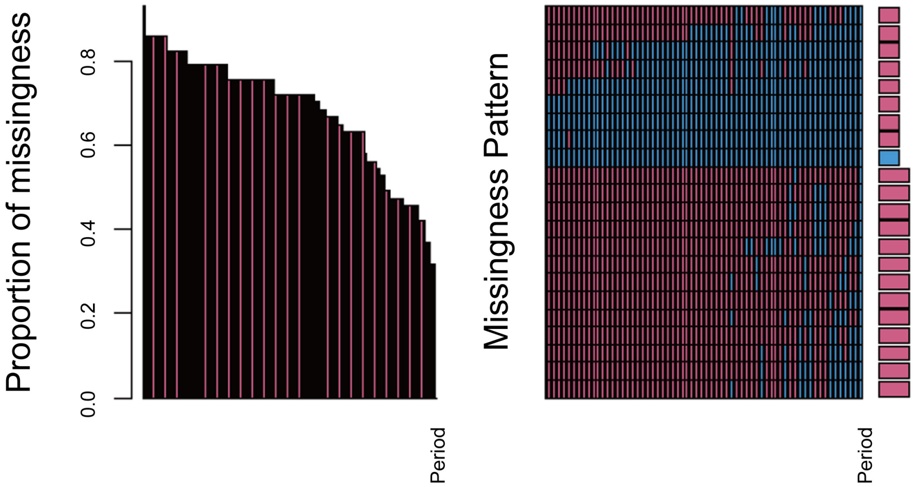


Figure 5. Missingness patterns for HIV/TB coinfection.

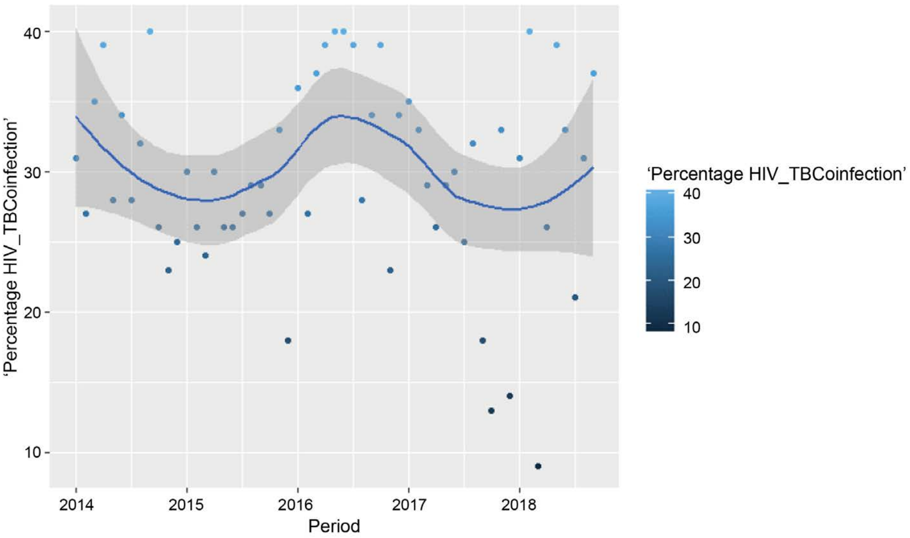


Figure 6. Plot of the complete cases only.

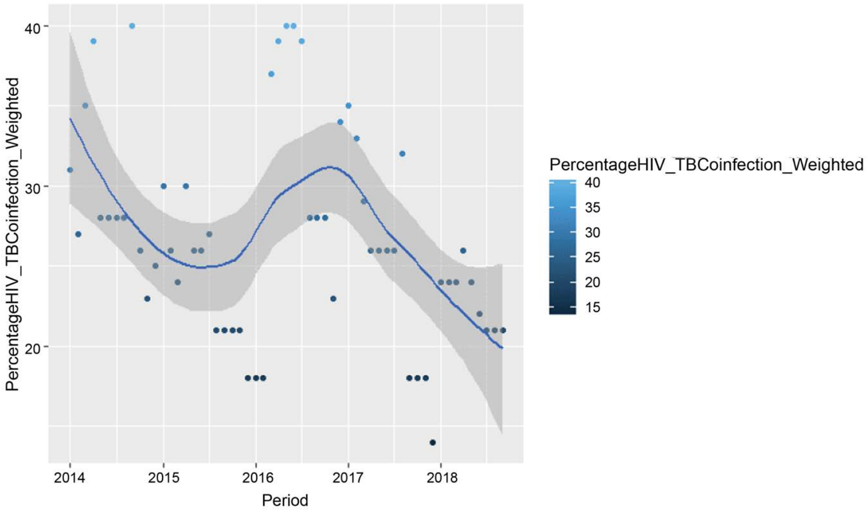


Figure 7. Plot of the weighted data.

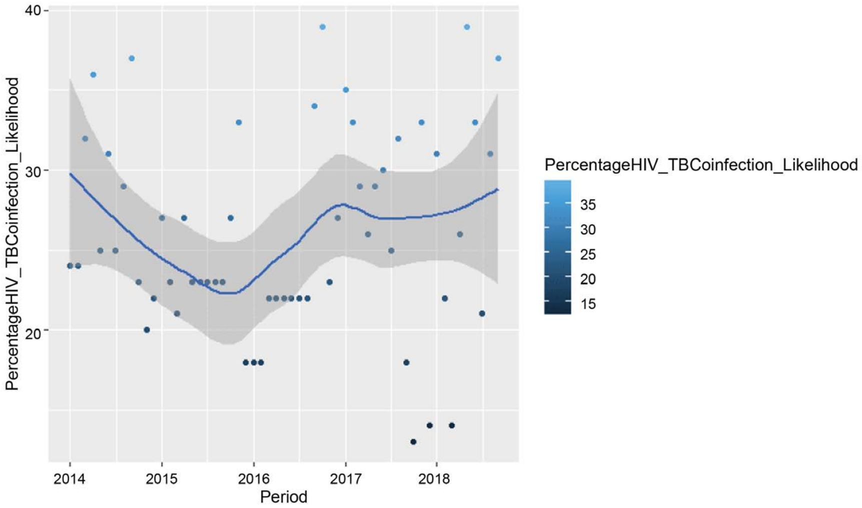


Figure 8. Plot of the likelihood estimated data.

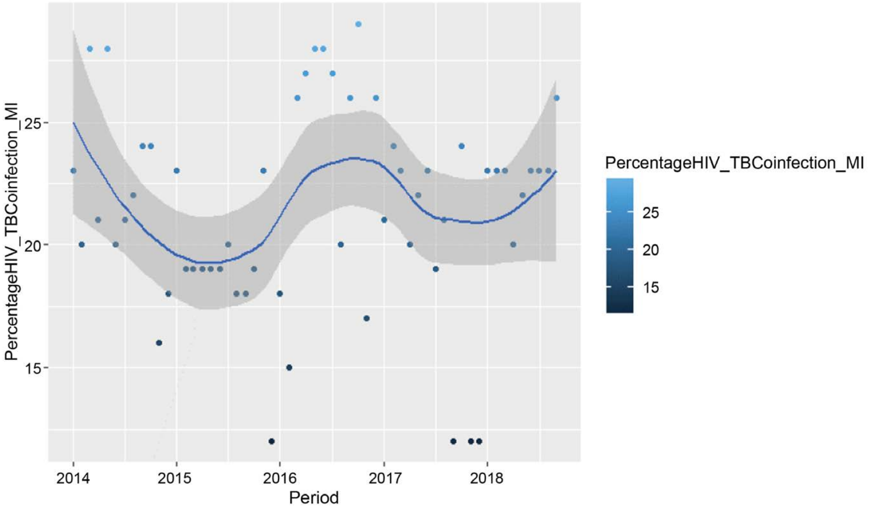


Figure 9. Plot of the Multiply Imputed data.

k-nearest neighbors with appropriate k) that are consistent, which means that, as one gets more and more data, the predictions of the algorithm are closer and closer to the optimal ones. So this seems to indicate that we can have some sort of universal algorithm.

Comparison of Four Approaches

Table 7 shows the comparison between Complete Case Analysis, Weighted method, Maximum Likelihood Based approach and Multiple imputation approach. Clearly, the complete case overestimates TB/HIV co-infection by a hooping 8%.

1) Complete Case Analysis

One major difficulty experienced in available case analysis, is that it can produce estimated covariance matrices that are implausible, such as estimating correlations outside of the range of Φ. Errors in estimation occur because of the differing numbers of observations used to estimate components of the covariance matrix. The relative performance of complete-case analysis and available case analysis, with MCAR data, depends on the correlation between the variables; available case analysis will provide consistent estimates only when variables are weakly correlated. The major difficulty with available case analysis lies in the fact that one cannot predict when available case analysis will provide adequate results and is thus not useful as a general method.

2) Weighted Method

In many scenarios, the mean imputation results in overall means that are equal to the complete case values, the variance of these same variables is underestimated. This underestimation derives from two sources. First, filling in the missing values with the same mean value does not account for the variation that would likely be present if the variables were observed. The true values probably vary from the mean. Second, the smaller standard errors due to the increased sample size do not adequately reflect the uncertainty that does exist in the data. A researcher does not have the same amount of information present when some cases are missing important variables as he or she would have with completely observed data. Bias in the estimation of variances and standard errors are compounded when estimating multivariate parameters such as regression coefficients. Under no circumstances does mean imputation produce unbiased results.

3) Likelihood Based approach

Maximum likelihood methods for missing multivariate normal data focus on

Table 7. Comparison of three approaches.

	Estimated HIV/TB Co-infection Rate	95% LCI	95% UCI
Complete Cases Only	29%	25%	33%
Weighted Method	27%	23%	31%
Likelihood Based approach	26%	24%	28%
Multiple Imputation Approach	21%	20%	22%

the estimation of the parameters of the observed data, namely the mean vector and variance-covariance matrix. Because we assume the data multivariate normal, we can utilize the well-known properties of conditional normal distributions to estimate the expected values of the sums and cross products of the variables. Using maximum likelihood with the EM algorithm does not result in values for individual missing variables. The estimates obtained for the means and the variance-covariance matrix of the variables of interest, and then uses of these parameter estimates to obtain model parameters such as the coefficients of a linear regression model (See **Tables 1-4**).

The one major difficulty with treatment methods for missing data is the computation of the standard errors of estimates (such as the standard error of the mean). Testing whether a mean is significantly different from zero, for example, requires an estimate of how accurate our estimate is. In maximum likelihood theory, the negative second derivative of the observed data log likelihood is needed to obtain standard errors of the estimated mean vector and covariance matrix. This quantity requires algebraic analysis to compute, and is unique to every set of multivariate data.

4) Multiple Imputation Approach

Multiple imputation avoids two of the difficulties associated with maximum likelihood methods using the EM algorithm. From **Tables 1-4**, with multiple imputation, a researcher will use standard methods of analysis once imputations are computed, and can easily obtain standard errors of estimates. Though specialized computing is required in multiple imputation, the method provides much more flexibility than in the method described in the previous section. While multiple imputation appears the most promising of current missing data methods, some criticisms of the method center on the amount of computing and analysis time. Analyzing five sets of data is certainly more costly than one analysis, and the method does require specialized software. Whatever model the analyst fits using the imputed data sets must be congenial to (must include the same variables) as the model used by the person who originally imputed the data.

5. Conclusion

When few cases are missing values, complete case analysis methods can provide unbiased estimates. In other circumstances, as in the HIV/TB co-infection setting, the number of complete cases is a small fraction of the total. The expense and investment in the study warrant our using methods that utilize as much data as possible. There is need to assess and acknowledge the missingness mechanism limitations whenever there is missing data.

Acknowledgements

We thank the editors and the reviewers for their comments. This research did not receive any external funding.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Allison, P.D. (2001) Missing Data, Volume 136. Sage Publications, New York. https://doi.org/10.4135/9781412985079

[2] Dong, Y.R. and Peng, C.-Y.J. (2013) Principled Missing Data Methods for Researchers. *SpringerPlus*, **2**, 222. https://doi.org/10.1186/2193-1801-2-222

[3] Haukoos, J.S. and Newgard, C.D. (2007) Advanced Statistics: Missing Data in Clinical Research—Part 1: An Introduction and Conceptual Framework. *Academic Emergency Medicine*, **14**, 662-668. https://doi.org/10.1197/j.aem.2006.11.037

[4] Van Buuren, S. (2012) Flexible Imputation of Missing Data. CRC Press, Boca Raton. https://doi.org/10.1201/b11826

[5] Carter, R.L. (2006) Solutions for Missing Data in Structural Equation Modeling. *Research & Practice in Assessment*, **1**, 20-27.

[6] Rubin, D.B. (2004) Multiple Imputation for Nonresponse in Surveys, Volume 81. John Wiley & Sons, Hoboken.

[7] Chinomona, A. and Mwambi, H. (2015) Multiple Imputation for Non-Response When Estimating HIV Prevalence Using Survey Data. *BMC Public Health*, **15**, Article No. 1059. https://doi.org/10.1186/s12889-015-2390-1

[8] White, I.R., Royston, P. and Wood, A.M. (2011) Multiple Imputation Using Chained Equations: Issues and Guidance for Practice. *Statistics in Medicine*, **30**, 377-399. https://doi.org/10.1002/sim.4067

[9] van Buuren, S. and Groothuis-Oudshoorn, K. (2010) Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45**, 1-68. https://doi.org/10.18637/jss.v045.i03

[10] World Health Organization, *et al.* (2010) The Global Plan to Stop TB 2011-2015: Transforming the Fight towards Elimination of Tuberculosis.

[11] Papathakis, P. and Piwoz, E. (2008) Nutrition and Tuberculosis: A Review of the Literature and Considerations for TB Control Programs. United States Agency for International Development, Africa's Health 2010 Project, 1.

[12] Corbett, E.L., Watt, C.J., Walker, N., Maher, D., Williams, B.G., Raviglione, M.C. and Dye, C. (2003) The Growing Burden of Tuberculosis: Global Trends and Interactions with the HIV Epidemic. *Archives of Internal Medicine*, **163**, 1009-1021. https://doi.org/10.1001/archinte.163.9.1009

[13] World Health Organization (2013) Global Tuberculosis Report 2013.

[14] NTLD-P (2018) Annual Report 2018. https://www.nltp.co.ke/annual-reports/#

[15] NTLD-P (2017) Annual Report 2017. https://www.nltp.co.ke/annual-reports/#

[16] Haitovsky, Y. (1968) Missing Data in Regression Analysis. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **30**, 67-82. https://doi.org/10.1111/j.2517-6161.1968.tb01507.x

[17] Im, J. (2015) Some Methods for Handling Missing Data in Surveys.

[18] Lee, K.J. and Simpson, J.A. (2014) Introduction to Multiple Imputation for Dealing with Missing Data. *Respirology*, **19**, 162-167. https://doi.org/10.1111/resp.12226

[19] Schouten, R.M., Lugtig, P. and Vink, G. (2018) Generating Missing Values for Simulation Purposes: A Multivariate Amputation Procedure. *Journal of Statistical*

Computation and Simulation, **88**, 2909-2930. https://doi.org/10.1080/00949655.2018.1491577

[20] James, G., Witten, D., Hastie, T., Tibshirani, R., Hastie, T. and MASS Suggests (2017) Package "ISLR".

Appendices

1) Abbreviations

MCAR: Missing completely at random.

MAR: Missing at random.

MNAR: Missing not at random.

MI: Multiple Imputation.

TB: Tuberculosis

HIV: Human Immunodeficiency Virus

MCMC: Markov chain Monte Carlo

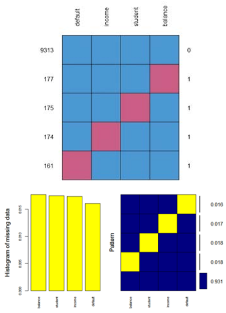
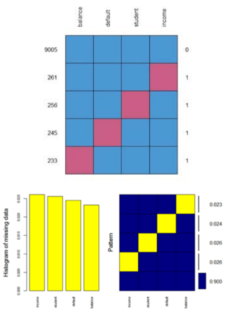
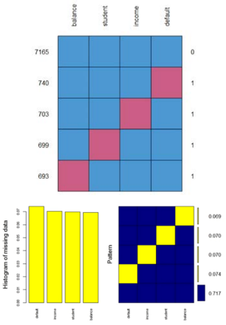
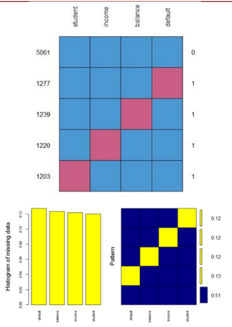
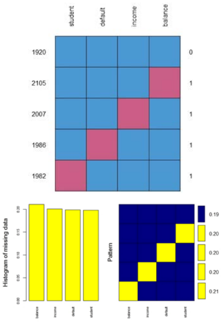
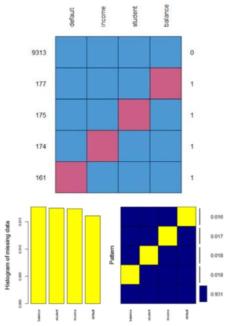
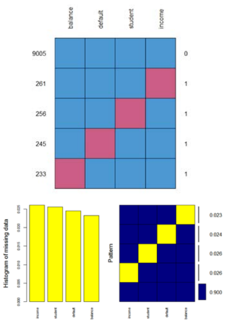
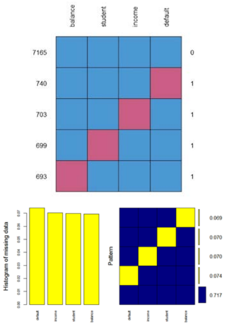
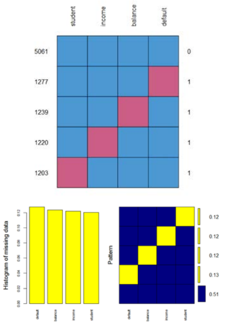
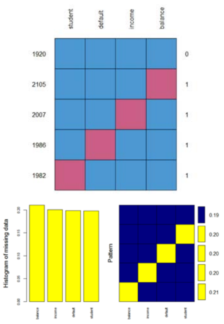
MICE: Multiple imputation by chained equations

NTLD-P: National Tuberculosis, Leprosy and Lung Disease Program

AUC: Area Under the receiver-operator Curve

2) Amputation results

Table A1. Graphical presentation of the amputations.

Missing Data Mechanism	7% Missingness	10% Missingness	30% Missingness	50% Missingness	80% Missingness
MNAR					
MAR					
MCAR