Evaluating the Performance of Different Multiple Imputation Methods When Imputing Missingness in Time-Series-Cross-Sectional Data

Xiaochen Dai

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Public Health

University of Washington

2019

Reading Committee:

Kwun Chuen (Gary) Chan, Chair

Mauricio Sadinle

Program Authorized to Offer Degree:

Biostatistics

University of Washington

Abstract

Evaluating the Performance of Different Multiple Imputation Methods When Imputing Missingness in Time-Series-Cross-Sectional Data

Xiaochen Dai

Chair of the supervisory committee:

Kwun Chuen (Gary) Chan

Department of Biostatistics

This thesis evaluates the performance of different multiple imputation methods in imputing country-level proportions of key indicators that are missing in time-series-cross-sectional (TSCS) data. When imputing the country-level proportions missing in TSCS data due to questions not asked in the survey, we found that *Amelia* and Multiple Imputation by Chain Equation for two-level panel data (*mice.2l.pan*) performed best among seven methods being evaluated for both methods converged fast, produced reasonable and stable imputations and had small out-of-sample root mean squared error (RMSE) less than ±5% for proportions imputed and 95% coverage rate (CR_{95}) very close to 95%. In addition, we found that including incomplete auxiliary variables that are correlated with targeted incomplete variables improved the imputation performance regardless of the missing rate of the auxiliary variables. However, including the cluster means had little impact on the imputation performance.

The goal of the thesis is to produce empirical evidence on the performance of different multiple imputation methods in imputing missingness in TSCS data.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First of all, I would like to extend my heartfelt gratitude to my thesis committee members, Prof. Gary Chan and Mauricio Sadinle, who had provided tremendous support for my thesis along the way. As the chair of my dissertation committee, Gary helped me out during the most difficult time of the process and I am beyond grateful for his support and kindness.

In addition to my committee members, I would also like to thank the faculty and staff in the Department of Biostatistics. Especially, I would like to thank Prof. Barbra Ann Richardson, Prof. Lurdes Inoue and Gitana Garofalo for their enormous support along the journey. Not only did they help me academically, they also lifted me up spiritually when life knocked me down.

I would also like to thank all my friends and colleagues who had helped me along the journey. Without all your support and comfort, I wound not have finished this long journey.

Last but not the least, I cannot be more thankful to my parents who have been supporting me throughout my life. Without their unconditional love and selfless support, I would not be able to study in the US, let alone to finish this MPH degree. I am eternally grateful to them and will always love them.

Abstract

Background

Time series cross-sectional (TSCS) data are an important type of data for comparative global health research, such as the Global Burden of Disease (GBD) study. However, TSCS data often face serious missing data problem due to questions not asked in some surveys (a.k.a. missing variables). Multiple imputation (MI) is a principled method for imputing missing data across different research fields. There are two major families of MI, namely, MI using joint modeling (MIJM) and MI using chain equations (MICE). Although both MIJM and MICE methods have been developed to properly impute missing data for TSCS data, little is known about their comparative performance in imputing missing variables in real TSCS datasets.

Methods

To evaluate the performance of different MI methods, we systematically extracted survey data on HIV/AIDS knowledge and attitudes in 47 SSA countries in Global Health Data Exchange (GHDx) and created a real TSCS dataset with country-level estimates of 16 key indicators for HIV/AIDS knowledge and attitudes from 2000 to 2017. We used 3 MIJM and 4 MICE methods to impute the country-level proportions of key indicators that are missing in the dataset 1000 times and evaluated the performance of the 7 methods using 10-fold cross validation. We used root mean squared error (RMSE) and coverage rate of the 95% credible intervals (CR_{95}) to evaluate the average accuracy of the 1000 imputations. We further examined the impact of including in the imputation model the cluster means and incomplete auxiliary variables with different missing rate on the imputations.

Results

In the dataset, the overall missing rate was 11.8%, with *heard_aids* and *mtct_drug* having the smallest and the largest missing rate of 2.7% and 40.2% respectively. The overall RMSE and CR_{95} were 0.0391 (±3.91% for proportions imputed) and 95.2% for Amelia and 0.0378 (±3.78% for proportions imputed) and 94.75% for *mice.2l.pan*, respectively, indicating good performance for both methods. The diagnostic plots also showed that *Amelia* and *mice.2l.pan* converged faster and produced more stable imputations than the other methods. Lastly, the average running time of the two methods were among the smallest of the 7 methods as well. In addition, we found that including cluster means in the imputation model had little impact on the imputations. However, including incomplete auxiliary variables improved the imputations even if the missing rate of the incomplete auxiliary variables was high.

Conclusion

When imputing missingness in TSCS data, *Amelia* and *mice.2l.pan* performed best among the 7 methods. Both methods converged fast, produced reasonable and stable imputations and had small out-of-sample $RMSE$ less than ±5% for proportions imputed and CR_{95} very close to 95%. In addition, Amelia and MICE could also be implemented parallelly, which significantly reduced running time and made the two methods highly practical.

1. Introduction

As in many other fields such as political science and economy, the time series cross-sectional (TSCS) data are an important type of data for global health research, especially for comparative studies of health indicators across countries and over time, such as the Global Burden of Disease (GBD) Study.[1] Large-scale national health surveys, such as Demographic Health Survey (DHS), Multiple Indicators and Cluster Surveys (MICS) and other national surveys conducted regularly by each country, are important sources of TSCS data for global health research. Although these TSCS data provide numerous measurements that open doors to many research opportunities, when used together, they often face serious missing data problem simply because some questions are not asked in some surveys (a.k.a. missing variables).[2–5] Even for DHS or MICS which are designed to be as consistent as possible over time and across countries, the questionnaires used in different countries or in different rounds can be significantly different from each other due to local adaptation and changes in the health priorities over time.[6] When surveys from different sources are used, missing variables usually become more prevalent.

Multiple imputation (MI), a Bayesian model-based approach first introduced by Rubin[7], has become a major principled method for estimating missing data across different research fields.[3,8] The basic idea and intuition of MI is to estimate the missing values in a dataset by making use of all the observed data. The estimation of the missing values is usually repeated *m* times to produce *m* different complete datasets in which the observed data are the same but the imputed data are different across the *m* complete datasets. After obtaining the *m* complete datasets using MI, one can perform the identical analyses on each of the *m* complete datasets and combine the results (estimates and standard errors) using simple rules provided by Rubin[7] to produce an overall estimate and its standard error. The major benefits of MI are that it results in unbiased

estimates, increase statistical power by using all available data and account for the uncertainty due to missing data.[7–9] Nowadays, there are two major families of MI approaches, namely, MI using joint modeling (MIJM) and MI using chain equations (MICE).[8,10,11] MIJM draws missing values simultaneously for all incomplete variables using a multivariate distribution (e.g. multivariate normal)[9–11] while MICE imputes incomplete variables one at a time, drawing missing values sequentially from a series of univariate distributions (e.g. regression models).[8,11–13] When first introduced by Rubin and others, MI was mainly used for imputing missingness in single-level cross-sectional data[7,14]. However, after more than four decades, MI has been developed to be able to properly impute missing data for multiple-level data[10,15,16], longitudinal or panel data [17,18] and TSCS data.[19]

In our last study on people's knowledge and attitudes about HIV/AIDS in sub Saharan Africa, we estimated the trends of 16 key indicators of HIV/AIDS knowledge and attitudes across 47 SSA countries. Although we found 248 national surveys asking key indicators of HIV/AIDS knowledge and attitudes, only a few of them asked all the 16 key indicators and many missed one or more indicators. In addition, some surveys only collected women's but not men's data. Therefore, after we stacked all the country-level survey estimates together into one dataset, there were many missing values due to missing variables across surveys. To attenuate the impact of missing variables on our results, we used regression method to impute men's indicators using women's when men's indicators were not collected and we estimated the trends of the 16 key indicators separately to avoid missingness due to indicators not collected in some surveys.[20] However, both measures we took had limitations. First, although the indicators for women and men are highly correlated and the linear mixed model we used gave small in-sample prediction error, regression imputation has long been considered inappropriate for imputation due to its

inability to account for uncertainty of the imputed data.[9,21,22] Buuren even thought that regression imputation is the most dangerous of all imputation methods because it artificially strengthens the correlations in the data and thus leads to false positive and spurious relations.[21] Second, although we avoided the missing variables problem by estimating the 16 key indicators separately, we also lost the opportunity to improve the estimated trends of the indicators by borrowing information from available indicators in the survey. Given the scarcity of the data on most indicators, borrowing information from available indicators could significantly improve the estimated trends. Lastly, the missing variables in some surveys prohibited us from constructing more informative composite measurements of people's knowledge and attitudes.

In this thesis, we use MI approach to impute the missing country-level proportions of key indicators of knowledge and attitudes about HIV/AIDS in the TSCS survey data and evaluate the performance of different MI approaches by the accuracy of the imputed indicators. The goals of the paper are 1) to produce a comprehensive and complete dataset of people's knowledge and attitudes about HIV/AIDS in SSA countries, in which country-level missing proportions are fully imputed with the uncertainty of the imputation properly accounted for and 2) to provide some empirical evidence on the performance of different MI methods for imputation of TSCS data.

2. Literature Reviews

2.1 Types of Missing Data

The problem of missing data is ubiquitous in research of all fields including global health research. Missing data, if handled inappropriately, can pose great threat to the validity of research findings.[7,8,10] According to the literature, missing data can be categorized either by the causes of the missingness or by the mechanisms of the missingness.

2.1.1 based on the causes of the missingness

When thinking of missing data in surveys, people usually think of missingness due to nonresponse to certain questions asked in a survey. However, there are more causes of missingness, especially for time series cross-sectional (TSCS) data, which are data collected from multiple sections (e.g. countries) and across multiple time periods (e.g. years).[2,19,23]

In a study of multi-format and multi-wave surveys, He et al. categorized missing data into three types based on the causes of the missingness, namely, *unit nonresponse, block nonresponse* and *item nonresponse*.[3,24] According to He et al. *unit nonresponse* refers to missingness of all variables of patients due to sampled patients not participating in the study; *block nonresponse* refers to missingness of blocks of variables due to early drop out or use of different forms of survey asking different questions; and *item nonresponse* refers to missingness of certain variables of a patient due to skip patterns of the survey or questions being refused or answered "don't know" by the patient.

Using different terminology, Denk and Weber categorized missingness of TSCS data into 6 types, namely, *missing items*, *missing variables within sections* (e.g. countries), *missing periods (e.g. years) within sections, missing periods*, *missing variables and missing sections*.[2] According to Denk and Weber, *missing items* refers to missingness of one or multiple indicators in one or multiple periods (years) for one or multiple sections (countries), e.g. an indicator is not collected for one country in one yea; *missing variables within sections* refers to missingness of one or multiple indicators in one section across all periods, e.g. an indicator is not applicable for one country and thus has never been collected for this country; *missing periods within sections* refers to missingness of all indicators in one or multiple sections for one or multiple periods, e.g. no survey is conducted in one or multiple years in a country and thus there is no data at all for these

years in this country; *missing periods* refers to missingness of all indicators in one or multiple periods for all sections, e.g. no survey is conducted in one or multiple years for all countries and thus there is no data at all in these years for all countries; *missing variables* refers to missingness of one or multiple variables in all periods across all sections, e.g. some variables have never been collected for all countries; lastly, *missing sections* refers to missingness of all variables in all periods for one or multiple sections, e.g. there is no data at all for one or multiple countries of interest in the dataset. ***Figure 1*** visually represents the six different types of missingness described by Denk and Webe.[2]

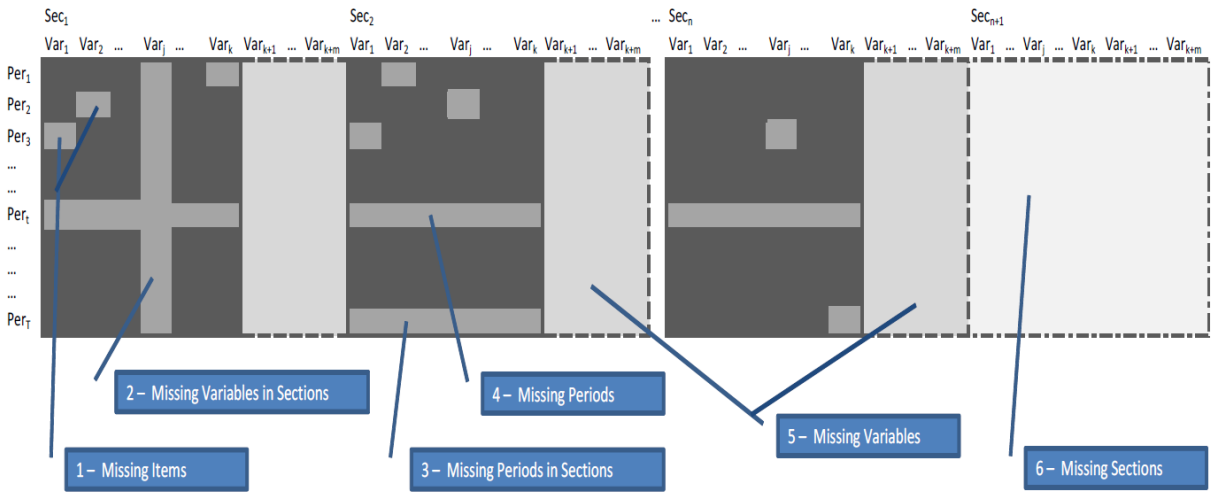


Figure 1 Missing data patterns for multivariate time series cross-sectional data [2]

2.1.2 based on the mechanisms of the missingness

In 1976, Rubin first described the mechanisms of missing data and categorized missing data into three types based on the reasons for the missingness.[25] According to Rubin, missing data can be missing completely at random (**MCAR**), missing at random (**MAR**) and missing not at random (**MNAR**).

The missing data are considered to be MCAR if the probability of missingness is not related to the data, either observed or unobserved. Mathematically, MCAR holds if

$$\Pr(M|D) = \Pr(M) \quad (1)$$

where M and D represent missingness and data (both observed and missing data) respectively. MCAR implies that there is no relationship at all between the data and missingness. If the missingness is by design or planned, or because the samples are lost in transit or due to equipment failure, the missingness is considered to be MACR.[11,26] Under MCAR, the analyses are not biased by the missingness and listwise deletion produces valid estimates and inferences though power may be lost due to exclusion of observations with missing data.[21,26]

MCAR is a strong assumption of missing data and is hardly met in reality. A less stringent and a more realistic assumption of missing data is MAR, under which the probability of missingness is related to the observed data but is independent of the unobserved data after conditioning on the observed data.

Mathematically, MAR holds if

$$\Pr(M|D) = \Pr(M|D_{obs}) \quad (2)$$

Under MAR, any relationship between missingness and the unobserved data disappears after conditioning on the observed data.[11] In statistics, the missingness is "ignorable" if the missing data are considered MCAR or MAR.[27] However, it is worth noting that the "ignorability" of missingness does not mean that one can ignore the missingness, especially if the missingness is MAR. Instead, one needs to handle the problem of missingness using appropriate methods so that the missingness does not bias the estimates and the inferences.

The last mechanism of missingness is MNAR, under which the probability of missingness is related to both observed and unobserved data. Mathematically, MNAR is expressed by the following equation:

$$\Pr(M|D) = \Pr(M|D_{obs}, D_{mis}) \quad (3)$$

which indicates that the missingness is related to the missing values themselves and the relationship between the missingness and the data remains even after conditioning on the observed data. For example, MNAR could occur if a respondent refuses to answer a question asking about his/her income because the respondent has very large or very small income.[11] MNAR is the real bane of missing data problem because one has to make untestable assumptions about the specific missing mechanisms to model the missing data in order to obtain unbiased estimates and inferences.[9,26]

2.2 Traditional Ad Hoc Methods for Missing Data Problem

Missing data is a very common issue in many studies across all fields. Many techniques have been developed to deal with missing data. Listwise deletion (LD), also referred to as complete-case analysis, is arguably the most commonly used ad hoc method to deal with missing data across fields. It is also the default way of handling missing data in many statistical packages, including R, Stata and SAS.[21] LD removes all the observations with missingness on the analysis variables and the following analyses are conducted using the complete cases only. The biggest advantage of LD is the convenience. Under MCAR, LD can produce unbiased estimates and inferences though at the cost of lower power due to exclusion of the incomplete observations.[21,28] However, the major limitation of LD is that it leads to biased estimates and inferences if the data are not MCAR, which is often the case.[22,29] In addition, LD often causes inconsistencies in the analyses. Since different analyses usually involve different subsets of variables and LD applies only to the active variables, different analyses are usually based on different subsamples of the dataset.[21]

Pairwise deletion (PD), a.k.a. available-case analysis, is a remedy to the data loss problem of LD by using the means, variances and covariances of all available data for the analyses. When calculating the means, variances and covariances of the data, PD uses all available cases, thus avoiding loss of data problem. After obtaining the moments of the data, one can use estimation method, such as method of moments (MOM), to estimate the coefficients of interest. Although being simple and avoiding the loss of data problem, PD still produces biased estimates and inferences if the data are not MCAR, which is the major shortcoming of the method. In addition, the covariance matrix may not be positive definite especially if the variables are highly correlated.[30] Moreover, due to missing data, using the average sample size for the estimates may yield over confident inferences. In short, PD only works well if the data are approximately multivariate normal, if the correlation between variables are low and if the missing data are MCAR.[21]

In addition to LD and PD, single imputation, including mean imputation, regression imputation (RI) and stochastic regression imputation (SRI), is another category of methods handling missing data. Mean imputation is to replace the missing data using the mean of the observed data. Although the method is simple, it seriously underestimates the variance, distorts the distribution of the data and produces biased estimates and inferences even under MCAR.[21] Compared with mean imputation, RI produces smarter imputations of missing data by incorporating information of the covariates in the imputation. However, RI artificially strengthens the relationships between the variables and systematically underestimates the variability of the imputed data.[21] SRI, which accounts for variability in the imputation by adding a random draw from the residual to the prediction, attempts to address correlation bias. However, the method still cannot fully capture the variability in the missing data and can produce implausible imputations.[21]

Another category of imputation method is donor-based imputation (DoBI), including hot-/cold-deck imputation and nearest neighbor methods. The general idea of DoBI is to impute the missing value of a "recipient" by finding a "donor" who is completely observed and has similar characteristics with the "recipient" and replacing the recipient's missing value with the donor's observed value. Hot-deck method groups the complete observations of a dataset into subsets which share the same values of some matching variables (e.g. age, sex, race etc.). Then, to each observation in the dataset with missing data, a donor is randomly assigned. The cold-deck method is only different from the hot-deck counterpart in that the donors are selected from other comparable data sources instead of the same dataset being filled in.[2] Nearest neighbor (NN) method measures the "distance" between complete and incomplete observations and matches the recipients with donors based on the distance between them. The distance can be calculated using multiple methods but is usually based on the metric matching variables. Usually, the nearest neighbor or one of the k nearest neighbors (KNN) randomly selected is used as the donor for the missing data. The benefits of NN/KNN are that it produces realistic imputations, better reflects the distributional property and can deal with missing data of any type.[2] However, the major limitation of NN/KNN method lies in its heuristic nature since the analyst needs to make many influential but subjective decisions such as the selection of matching variables and the choice of distance measures when using the method.[2,31]

The next category is distribution-based imputation (DtBI), which randomly draws imputation from the empirical (non-parametric) or probabilistic (parametric based on distributional assumptions) distribution of the observed data. Although univariate distributions are most often used for imputation of missing data, a multivariate approach may produce more reasonable combinations of imputed data for multiple incomplete variables.[2] DtBI is the foundation of more

sophisticated imputation methods such as multiple imputation, which we will discuss in detail in the next section.

Lastly, last observation carried forward (LOCF) and baseline observation carried forward (BOCF) are two ad hoc imputation methods specific for missing data in time series and longitudinal data. The idea of these methods are simple—the last period/ baseline observed value is used to replace the missing data in the following period(s). LOCF and BOCF are commonly used in clinical trials due to its simple and convenient nature. However, LOCF and BOCF produce biased estimates and inferences even under MCAR[32] and thus are not recommended for handling missing data in longitudinal and panel data.[33]

2.3 Multiple Imputation

Since most traditional ad hoc methods handling missing data result in biased estimates and inferences, more appropriate methods have been developed to handle missing data, including full information maximum likelihood (FIML) and multiple imputation (MI).[26,34] Compared with LD and single imputation method, FIML is a more appropriate method coping with missing data as it incorporates information of both observed and missing data into the likelihood function and finds the estimates that maximize this likelihood function.[34–38] However, FIML is available only for certain models such as structure equation models (SEM) and can only be implemented by special software packages.[35] Also, FIML, like listwise deletion, does not impute missing values [34] and thus is not very useful for this study, where the missing country-level proportions of key indicators, along with their uncertainty, need to be imputed.

Among all the techniques handling missing data, MI is the most appropriate method for my purpose in that it imputes the missing values and also accounts for the uncertainty inherent in the imputation.[35,39,40] Multiple imputation was first proposed by Rubin[7] to deal with nonresponses in

surveys. It uses information (e.g. distribution, correlation etc.) of the observed data to estimate likely values of the missing data. MI estimates the missing values m times with each time incorporating a random component to account for uncertainty about the missing values. In the end, we obtain m completed datasets with the same observed data but different imputed missing data. Once m completed datasets are imputed, one can perform identical analyses he/she wants using each of the m completed datasets and then pools the estimates from each dataset together using Rubin's rule.[7,40] Under the assumption of MCAR or MAR, the pooled estimates have been proved to be unbiased and the standard errors are appropriately adjusted.[7,37,39–41] ***Figure 2*** presents a graphical flow chart of MI procedure.

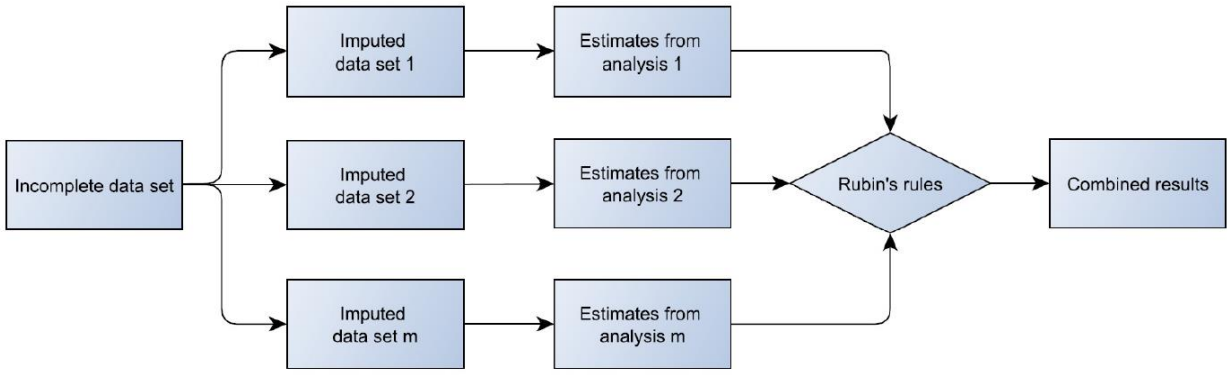


Figure 2 The prodecure of multiple imputation [7,40]

Although the MI process is always the same as shown in ***Figure 2***, there are different ways of imputing the missing data. Based on the imputation algorithm, there are two major families of MI, namely, MI by joint modeling (MIJM) and MI by Chained Equation (MICE).[11,12,15]

2.3.1 multiple imputation by joint modeling

MIJM assumes that the variables in the imputation model follow a joint distribution, such as a multivariate normal (MVN) distribution.[7,10,39] Under this assumption, the missing values are treated as random draws from the posterior predictive distribution given the observed data.[39]

However, drawing directly from this posterior joint distribution is difficult. Therefore, many algorithms have been developed to simulate the predictive posterior distribution.

Imputation-posterior (IP) is a full Bayesian algorithm based on Markov Chain Monte Carlo (MCMC) method. It is an iterative process involving two steps, imputation step (**I**) and posterior step (**P**). In imputation step, missing data are drawn from its conditional predictive distribution augmented by conditioning on estimates of distribution parameters,

$$\widetilde{D}_{mis} \sim P\left(D_{mis} \middle| D_{obs}, \tilde{\mu}, \tilde{\Sigma}\right) \quad (4)$$

Then, in the posterior step, new values of the parameters $\mu \; and \; \Sigma$ are drawn from their posterior conditioning on the observed and present imputed values for the missing data,

$$\tilde{\mu}, \tilde{\Sigma} \sim P(\mu, \Sigma | D_{obs}, \widetilde{D}_{mis}) \quad (5)$$

This process is iterated and when it is converged, the draws of $\widetilde{D}_{mis}, and \; \tilde{\mu} \; and \; \tilde{\Sigma}$ are from the true posterior independent of their starting values [39].

The advantage of IP is that the algorithm is theoretically justified, which means we are confident that once converged, the draws from the conditional posterior distribution are draws from the joint distribution. Therefore, the predictive posterior distributions are exact. However, the downside of IP is that this algorithm is computationally intensive. In many cases, MCMC can only converge after an infinite number of iterations, which means a long "burn-in period"[39] and the diagnosis of convergence needs expert assessment. In addition, the multiply imputed data needs to be independent in order to use Rubin's rule to pool them together.[7] However, the draws from MCMC are auto-correlated by nature. In practice, people reduce the dependence by using every r^{th} (e.g. 100th) random draws from IP, which further increases the computational burden of the algorithm.[39] Lastly, when the dataset contains different types of variables (e.g. continuous,

binary, categorical and count variables), the current classes of joint models (e.g. MVN model, log linear model, general location model),[42] may not be appropriate for the joint distribution of the data.[43,44]

Expectation Maximization (EM) is a deterministic version of IP. Instead of randomly drawing from the posterior distribution, EM calculates the posterior means and use them as the imputed values. In the E step, missing cells ($\widetilde{D}_{mis}$) are filled with their predicted values, and in the M step, the random draw of $\tilde{\mu}, \tilde{\Sigma}$ are replaced with the maximum posterior estimate.[39,45] The advantages of EM are that it is simple and fast, converges deterministically and can account for fundamental variability in the imputations. However, the serious shortcoming of EM is that it does not account for the uncertainty inherent in estimation of $\tilde{\mu}, \tilde{\Sigma}$.[39] To account for the uncertainty in estimation of $\tilde{\mu}, \tilde{\Sigma}$, King et al[39] propose EMs, EMis and EMB.

Expectation Maximization sampling (EMs) accounts for uncertainty in estimation of $\hat{\theta}$, (i.e. $\tilde{\mu}, \tilde{\Sigma}$) using the asymptotic approximation.[39] After running EM to find the maximum posterior estimates of $\hat{\theta}$, King et al use the outer product gradient or inverse of the negative Hessian to calculate the variance of $\hat{\theta}$, $V(\hat{\theta})$ and then draw $\hat{\theta}$ from a $MVN\left(\hat{\theta}, V(\hat{\theta})\right)$. They then use $\hat{\theta}$ to compute $\tilde{\beta}$ deterministically and use the equation

$$\widetilde{D}_{ij} = D_{i,-j}\tilde{\beta} + \tilde{\epsilon}_i, \qquad \tilde{\epsilon}_i \sim N(0,1) \quad (6)$$

to impute the missing values. In this process, they draw $\hat{\theta}$ m times and impute each missing values m times. The advantage of EMs is that it is fast, converges non-stochastically, does not reply on Markov chain and accounts for uncertainty in estimation of $\hat{\theta}$. However, although EMs works well in large sample, it leads to biased results when sample size is small, ratio of the

number of variables to that of observations is high or when there are highly skewed categorical data.[39]

Expectation Maximization importance resampling (EMis) builds upon the EMs but includes a round of importance resampling, which is a technique to improve small sample performance but not based on Markov chain.[7,39,46–48] In addition to all the EMs steps, EMis uses an acceptance-rejection algorithm by keeping draws of $\tilde{\theta}$ from EMs with probability proportional to the "importance ratio"

$$IR = \frac{L\left(\tilde{\theta}\middle|D_{obs}\right)}{N\left(\tilde{\theta}\middle|\tilde{\theta}, V\left(\tilde{\theta}\right)\right)} \quad (7)$$

and discarding the rest. The kept draws are considered independent draws from the posterior distribution. The EMis has all the benefits of EMs, such as easy and fast, dose not rely on Markov chain and thus produces full independent imputation. Besides, it also works well for small sample. The posterior produced by EMis also approximates that produced by IP, which is the gold standard for missing data imputation.[39] However, EMis does not work well for all likelihood functions, especially when the normal density is not a good approximation.[39]

Bootstrapped-based EM algorithm (EMB) draws m samples of size n with replacement from the data. In each sample, EM algorithm is implemented to produce estimate of $\tilde{\mu}, \tilde{\Sigma}$. Then for each set of estimates, use the original sample to impute the missing values in their original position. This process produces m multiply imputed datasets.[19] The benefits of EMB compared with EMis and IP are that, it is much faster (computation can be done in a parallel fashion), has better lower order asymptotics than the parametric approaches used by EMis and IP and is more

robust to distributional and small sample problems.[19] The bootstrapped estimates of $\tilde{\mu}, \tilde{\Sigma}$ are very close empirically to those from posterior distribution in large samples.[49]

2.3.2 multiple imputation by chain equation

Besides MIJM, **Multiple Imputation by Chained Equation (MICE)**, also called "fully conditional specification" (FCS) or "sequential regression multiple imputation",[13,35,50] is another major family of MI. Different from MIJM, where the missing values are treated as random draws from posterior predictive distribution given the observed data, MICE reduces the imputation problem to a series of estimations where each variable takes its turn to be estimated using the other variables. This procedure provides great flexibility as each variables can be assigned a suitable distribution, e.g. linear (for continuous variable), Poisson (for count variable), binomial (for binary variable) or multinomial (for categorical variable).[35,51] MICE runs through an iterative process, the detailed steps can be found in the paper of Azur et al.[35] In step 6, noted that a number of iterations (e.g. 10 iterations) are performed to make sure that the distribution of the parameters governing the model have converged and then one imputed dataset was obtained. The whole process is repeated m times producing m imputed complete datasets.[35] The biggest advantage of MICE is great flexibility. One can specify different models for different types of variables and can easily impose bounds and restrictions, such as skip pattern, upon some variables,[13,13,35] which makes it more suitable than joint modeling for some datasets [43]. However, although widely used in medical research, MICE still lacks theoretical justification,[35,40,51] i.e. the implicit joint distribution underlying the separate models may not always exist, that is, the conditional models may be incompatible.[43] When the conditional models are incompatible, the order in which the missing values are updated in the chained equations may seriously affects the results, which is referred to as "order effect".[43] Although Hughes, et al.[43] and Liu, et al.[52]

independently proves the sufficient conditions under which MICE equates MIJM, these conditions are so strict that they are hard to be met in practice.[43,52] Another major downside of MICE is that correctly specifying the model for each variable is very difficult if not impossible and enough auxiliary variables informing the missingness need to be included in the model.[35,40,53] Misspecification of the models can lead to failure of MICE[54] and is the major contributor to biased results.[40,55,56]

Due to the flexibility of MICE, the model for each variable needs not to be parametric, e.g. regression models. Instead, semi-parametric and non-parametric methods can also be used to estimate the missing values. When people use MICE to impute the missing data, two popular non-/semi-parametric methods, namely, random forest (RF) and predictive mean matching (PMM), are often used to estimate the missing values. **MICE with random forest (MICErf)** is a non-parametric technique well-suited for handling complex non-linear relationships. It reduces bias due to model misspecification when model includes complex interactions and polynomial terms.[40,57] Using MICErf, we only need to worry about including all variables (including auxiliary variables) that informs the missingness but do not need to worry about including non-linear terms, such as interactions and polynomial terms, in the model.[40,56] Compared with MICE, MICErf can handle high dimensional data and highly correlated predictors, and it also runs much faster.[40] However, similar to MICE, the biggest drawback of MICErf is that the conditional models may be incompatible.[40] **MICE with predictive mean matching (MICEPMM)** is a semi-parametric method and produces imputed values that resemble the observed values better than other methods because it uses the predicted value for a given observation to identify similar observations. Then the identified similar observations form a matching set and imputed values are randomly drawn from this matching set.[40,51] Therefore, one benefit of PMM method is that it

prevents unrealistic values.[40,58] The major disadvantage of PMM is also lacking of mathematical justification.[40]

2.3.3 multiple imputation for time series cross-sectional data

When first introduced by Rubin et al., MI could not properly handle missingness in multilevel or time series data because it could not account for the cluster structure or the autocorrelation of the data.[7] However, after decades of development, the state-of-the-art MI techniques can properly handle missingness in more complex data such as panel or even TSCS data.[18,59] Schafer et al. were the first to propose a multilevel imputation methods called PAN, which extends the MIJM for multilevel data by specifying a multivariate mixed model to predict the incomplete variables in level-1 variable using the complete level-1 and level-2 variables.[9,10,15] The multivariate mixed model is as follow:

$$\boldsymbol{Y_j} = \boldsymbol{X_j}\boldsymbol{\beta} + \boldsymbol{Z_j}\boldsymbol{B_j} + \boldsymbol{E_j} \quad (8)$$

where j denotes cluster j, $\boldsymbol{Y_j}$ contains all the incomplete level-1 variables in cluster j, $\boldsymbol{X_j}$ is the matrix of all the complete level-1 and level-2 variables including a unit vector for intercepts. $\boldsymbol{\beta}$ is the matrix of fixed effects of the complete variables which are the same for all clusters. $\boldsymbol{Z_j}$ is the matrix of a subset of complete level-1 variables which have random effects (slopes and intercepts) on the variables in $\boldsymbol{Y_j}$, as well as a vector of 1 for random intercepts. $\boldsymbol{B_j}$ is the matrix of random effects (level-2 residuals) of the variables in $\boldsymbol{Z_j}$ for cluster j. $\boldsymbol{E_j}$ is the matrix of level-1 residuals for cluster j.[11]

Then multilevel joint imputation draws imputed values from the following conditional multivariate normal distribution:

$$\boldsymbol{Y_j}|\boldsymbol{X_j} \sim MVN\left(\boldsymbol{X_j}\boldsymbol{\beta} + \boldsymbol{Z_j}\boldsymbol{B_j}, \boldsymbol{\Sigma}\right) \quad (9)$$

In equation (9), the multivariate normal distribution is centered at the predicted value from the imputation model with spread of $\boldsymbol{\Sigma}$, which is the level-1 residual covariance matrix. A scaler version of equation (9) is detailed by Mistler[11] and thus not described here. The parameters in equation (9) can be estimated using IP algorithm. The detailed estimation and sampling steps can be found in many literature[15,60–62] and thus are not described here. In equation (9), $\boldsymbol{\Sigma}$ captures the entire relationships between incomplete level-1 variables and is the same across all clusters, meaning that the random effects of the incomplete level-1 variables on other incomplete level-1 variables cannot be captured in the PAN method.[11,16] Therefore, if the analysis model contains the random effects between incomplete level-1 variables, PAN imputation method is said to be uncongenial with the analysis model.[11,42,63]

To allow the PAN method to account for the random effects between the incomplete level-1 variables, Yucel proposed an improved PAN method, called random-covariance and mixed-effect (RCME) model, which allows the residual covariance matrix $\boldsymbol{\Sigma}$ to randomly vary across different clusters and thus preserves the important relationships between the incomplete variables in the imputation model.[16,64] This RCME model can be implemented by R package JOMO[65] and thus we call it JOMO method.

Although the PAN and JOMO methods are not specifically developed for TSCS data, they can be used to handle missingness in TSCS data by including time (e.g. year) in the imputation model and allowing the effect of time on the incomplete variables to vary across the clusters (e.g. countries). In addition to PAN and JOMO, King et al.[19] developed another MIJM method called *Amelia* to handle missingness in TSCS data in particular. Different from PAN and JOMO, which employ a full Bayesian method, i.e. IP, to estimate the parameters of the posterior predictive joint distribution, Amelia uses EMB algorithm to estimate the parameters of the posterior.

Compared with IP, the EMB algorithm is much faster, can handle more variables, and produces empirically similar results to those produced by IP in large samples.[19] Under the assumption that time series variables often have smooth trends over time, Amelia included smooth basis functions, such as polynomials or splines, of time in the imputation model and allows the basis functions of time to interact with country indicators to account for heterogeneity of trends across countries.[19] Another advantage of Amelia is its ability to easily incorporate one's prior knowledge about the missing data into the imputation process. Instead of specifying the priors for the abstract model parameters, Amelia allows one to incorporate prior knowledge of the missing data using either (1) a mean plus a standard error of the missing value or (2) a confidence interval of the missing value.[66]

Besides MIJM, MICE can also deal with missingness in multilevel data. In fact, since MICE specifies univariate imputation model for each incomplete variables in the dataset, MICE is much more flexible to incorporate random effects in the imputation models than MIJM is. If the analysis model includes random effects of some incomplete variables on another incomplete variable, MICE can easily make the imputation models congenial by including the needed random effects in the imputation model. The mathematical expression of imputation model for cluster j and variable k in multilevel MICE is as follow:

$$\boldsymbol{y}_{jk} = \boldsymbol{X}_{jk}\boldsymbol{\beta}_k + \boldsymbol{Z}_{jk}\boldsymbol{b}_{jk} + \boldsymbol{e}_{jk} \quad (10)$$

where $\boldsymbol{y}_{jk}$ is the vector of incomplete level-1 variable k in cluster j ($\boldsymbol{y}_{jk}$ can represent each incomplete variable once). $\boldsymbol{X}_{jk}$ is the matrix of all the other complete and incomplete variables predicting $\boldsymbol{y}_{jk}$. $\boldsymbol{\beta}_k$ is the vector of fixed effects for variable k that are the same for all clusters. $\boldsymbol{Z}_{jk}$ is the matrix of a subset of $\boldsymbol{X}_{jk}$ that have random effects on $\boldsymbol{y}_{jk}$. $\boldsymbol{b}_{jk}$ is the vector of random

effects (i.e., level-2 residuals) of the variables in $\boldsymbol{Z}_{jk}$ on $\boldsymbol{y}_{jk}$. $\boldsymbol{e}_{jk}$ is the vector of level-1 residuals for cluster j, variable k.[11]

The multilevel MICE draws imputed values for each incomplete variable from a conditional normal distribution as follow:

$$\boldsymbol{y}_{jk}|X_{jk} \sim N\left(\boldsymbol{X}_{jk}\boldsymbol{\beta}_k + \boldsymbol{Z}_{jk}\boldsymbol{b}_{jk},\ \sigma_k^2\right) \quad (11)$$

In equation (11), the conditional univariate normal distribution is centered at the predicted value from the imputation model for variable k, with spread equal to the level-1 residual variance of k (i.e., σ_k^2).[11] Noted that the set of random effect variables predicting each incomplete variable k can be different, which allows for inclusion and exclusion of certain random effects when necessary. The parameters in equation (11) can be estimated by an MCMC algorithm which is described in many literature[15,21,60–62] and thus is not detailed here.

Built upon on the above model, Carpenter and Kenward[64] recommended to include the cluster means of level-1 variables as predictors in the univariate imputation models. Mistler and Enders[67] confirmed the benefits of such adaptation and found that inclusion of cluster means of level-1 variables greatly improved the performance of MICE in all scenarios. However, a recent study done by Resche-Rigon and White[68] found that including cluster means of level-1 variables to the imputation models had little impact on the performance of MICE though did not hurt either.

Similar to MIJM methods, such as PAN and JOMO, MICE can also handle missingness in TSCS data by including time (e.g., year) in the univariate imputation model for each of the incomplete variables. The MI methods mentioned above, namely, PAN[69], JOMO[65], Amelia[70] and MICE[71], can all be implemented in R.

2.4 Evaluation of the Performance of MI

According to Chambers, there are four types of accuracy of imputations. Ranked from the hardest to the easiest to achieve, the four types of accuracy are predictive accuracy, ranking accuracy, distributional accuracy and estimation accuracy.[72] The strongest predictive accuracy requires the maximal preservation of true values, which implies the other three types of accuracy; ranking accuracy requires maximal preservation of the order of true values; distributional accuracy requires the maximal preservation of distributions of the true values (e.g. the marginal and higher order distributions); and the weakest estimation accuracy only requires the reproduction of lower order moments (e.g., mean and variance) of the true values.[2,72]

Evaluation of the performance of MI depends on one's objective of using the MI method. According to Barnard and Meng[73], the application of MI falls within two categories, the "outside" and "in-house" application. The traditional application of MI is mostly "in-house" where the imputer and the analyst is the same person and the imputation is done for one specific analysis, the so-called "one-analyst-one-goal" studies. However, in "outside" application, the imputer and the analyst are typically different persons and the main objective of MI is to accurately impute missing data and thus to produce multiple complete datasets to fit for the "many-analysts-many-goals".[5,24]

For "in-house" application, due to the nature of "one-analyst-one-goal", we only care about a single or a few parameters of the population, e.g., the mean and variance of some variables or the coefficients of a regression in the population. We only ask how well the MI method helps us estimate these population parameters. Therefore, we only require estimation accuracy of the MI method. Dr. Rubin[74] even stated that the objective of multiple imputation should rather be statistically valid inference (i.e. estimation accuracy) than the optimal point prediction (i.e.

predictive accuracy). As a results, for "in-house" application, the performance of MI is evaluated by how close the MI estimates and inferences are to the true population parameters.

However, different from "in-house" application, the "outside" application requires predictive accuracy of MI to achieve "many-analysts-many-goals". Since the predictive accuracy implies the other three types of accuracy, achieving predicative accuracy makes sure that the imputed datasets can be used for a variety of analyses.[72] To some extent, the "outside" application uses MI to accurately "predict" the missing values while accounts for uncertainty of the "predictions" by imputing the missing values multiple times.[5,24] As a results, for "outside" application, the performance is evaluated by how close the imputed data are to the true data.

Although the evaluative targets of MI are different under different applications, the methods used to evaluate the "closeness" of estimated parameters of the population (for "in-house" application) or imputed data (for "outside" application) to the true values are the same. Since the true values are unobserved, a simulation method is widely used to evaluate such "closeness" for imputation methods.[12,75–77] Simulation method artificially "generates" missingness by removing some of the observed data from the original dataset. The missingness can be generated under different assumptions, i.e., MCAR, MAR and MNAR.[12,67] The original dataset can be either artificially generated[11,67,75,76] or derived from a real world dataset.[12,77] Under "in-house" application, an MI method is considered good if the MI estimates and inferences of regression coefficients using the imputed datasets are close to the estimates and inferences using the original complete dataset.[11,12,67,75–77] Whereas, under the "outside" application, an MI method is considered good if the imputed data are close to the observed true data.[78,79] The common performance indicators are (1) bias, which is defined as the difference between an estimator's average value and its true value, (2) root mean squared error (RMSE), which is the square root of the mean squared

difference between average estimates and true values, and (3) the coverage rate of 95% confidence/credible interval, which is the proportion of the confidence/credible intervals covering the true values.[75] These three indicators can be used to evaluate the performance of MI methods under both applications.

3. Methods

3.1 Data Description

3.1.1 country-level estimates of the key indicators

The data used to evaluate the performance of different MI methods are extracted from national health surveys including DHS, AIS, MICS and other country specific national surveys. Surveys conducted from 2000 to 2017 are systematically searched in Global Health Data Exchange (GHDx) using 47 SSA countries (**Table 1**) as keywords. Among all the surveys found, only those having microdata of at least one of the 16 key indicators (**Table 2**) of HIV/AIDS knowledge and attitudes are included.

Table 1 The 47 Sub-Saharan African Countries of Interest

Subregion of SSA	Countries of Interest
Central SSA	Angola(AGO), Central African Republic(CAF), Congo(COG), Democratic Republic of the Congo(COD), Equatorial Guinea(GNQ) and Gabon(GAB)
Eastern SSA	Burundi(BDI), Comoros(COM), Djibouti(DJI), Eritrea(ERI), Ethiopia(ETH), Kenya(KEN), Madagascar(MDG), Malawi(MWI), Mozambique(MOZ), Rwanda(RWA), Somalia(SOM), South Sudan(SSD), Tanzania(TZA), Uganda(UGA) and Zambia(ZMB)
Western SSA	Benin(BEN), Burkina Faso(BFA), Cameroon(CMR), Cape Verde(CPV), Chad(TCD), Cote d'Ivoire(CIV), The Gambia(GMB), Ghana(GHA), Guinea(GIN), Guinea-Bissau(GNB), Liberia(LBR), Mali(MLI), Mauritania(MRT), Niger(NER), Nigeria(NGA), Sao Tome and Principe (STP), Senegal(SEN), Sierra Leone(SLE) and Togo(TGO)
Southern SSA	Botswana(BWA), Lesotho(LSO), Namibia(NAM), South Africa(ZAF), Swaziland(SWZ) and Zimbabwe(ZWE)
Northern SSA	Sudan(SDN)

Table 2 Key indicators of HIV/AIDS knowledge and attitudes

Category	Definition of indicator	Variable name	Values
Gateway question	Ever heard of HIV/AIDS	heard_aids	1: Yes 0: No/DK
Knowledge on HIV/AIDS prevention	Knowing that one can reduce chance of getting AIDS by having just one uninfected partner who has no other sex partners	kw_pv_one_partner	1: Yes 0: No/DK
	Knowing that one can reduce change of getting AIDS by using a condom every time they have sex	kw_pv_condom	1: Yes 0: No/DK
Knowledge on HIV/AIDS mother-to-child transmission	Knowing that HIV can be transmitted from mtc during pregnancy	kw_mtct_preg	1: Yes 0: No/DK
	Knowing that HIV can be transmitted from mtc during delivery	kw_mtct_delivery	1: Yes 0: No/DK
	Knowing that HIV can be transmitted from mtc through breasfeeding	kw_mtct_bf	1: Yes 0: No/DK
	Knowing that there is a drug to prevent mtct	kw_mtct_drug	1: Yes 0: No/DK
Knowledge on HIV/AIDS misconceptions	Believing that HIV can be transmitted by mosquito bites	mis_mosquito	1: No 0: Yes/DK
	Believing that HIV can be transmitted by sharing food with PLWHs	mis_food	1: No 0: Yes/DK
	Believing that one can get AIDS from witchcraft	mis_witchcraft	1: No 0: Yes/DK
Knowledge on HIV/AIDS testing	Knowing a place where people can get tested for HIV	kw_where_test	1: Yes 0: No/DK
Other knowledge on HIV/AIDS	Knowing that healthy-looking person can have HIV	kw_looking	1: Yes 0: No/DK
Attitudes towards people living with HIV/AIDS	Would buy fresh vegetable from a HIV-infected vendor	att_vegetable	1: Yes 0: No/DK
	Would want to remain a secret if a family member got infected with HIV	att_secret	1: No 0: Yes/DK
	Willing to care for an infected family member in his/her own household	att_willing_care	1: Yes 0: No/DK
	Believing that a female teacher with HIV should be allowed to continue teaching in the school	att_f_teacher	1: Yes 0: No/DK

To obtain the country-level estimates of these indicators, the individual-level microdata are systematically extracted and then aggregated into country-level estimates by taking weighted mean of the individual-level data. For each survey, the individual-level data are aggregated by sex and by age groups (e.g., 15-24, 25-49 and 15-49). We used the *survey* package in R to obtain the weighted means and their corresponding standard errors. As mentioned above, since not all indicators are asked in all the surveys, there are country-level proportions of key indicators that are missing in some surveys, which we try to impute latter. Since the aggregated country-level knowledge and attitudes variables are all proportions between 0 to 1, we logit-transform them before imputing the missing country-level proportions of key indicators to improve the imputation performance. Therefore, the imputed variables are also in the logit scale and thus need to be back transformed to its original scale afterwards.

3.1.2 country-level covariates of HIV/AIDS knowledge and attitudes

To improve the performance of imputation, important country-level covariates (**Table 3**) of HIV/AIDS knowledge and attitudes are included in the imputation model. The country-level covariates are extracted from IHME's GBD study 2017 and are complete over the period from 2000 to 2017. To improve the performance of imputation models, we logit-transform *contra_prev*, *ASFR*, *prop_urban*, *ANC4* and log-transform *GDP*.

Table 3 List of country-level covariates

Covariate	Description	Type
education	Mean years of education per capita (by sex)	Continuous
GDP	GDP per capita base 2010 international dollars	Continuous
ASFR	Age-specific fertility rate	Proportion
contra_prev	Modern contraception prevalence in women by age groups	Proportion
HAQI	Healthcare access and quality index [80]	Continuous
prop_urban	Proportion of population living in urban area	Proportion
Muslim	Binary indicator: value 1 if country is greater than 50% Muslim	Binary

HSA	Health system access: a composite score of immunization, measles immunization, hospital beds, in-facility delivery and skilled birth attendance	Continuous
ANC4	Proportion of pregnant women receiving 4 or more antenatal care from a skilled provider.	Proportion

3.1.3 covariates of survey and estimates

In addition to covariates of HIV/AIDS knowledge and attitudes, covariates of surveys and estimates, e.g., year and country of the survey, sex and gender of the estimates, are also important covariates to be included in the imputation model. **Table 4** describes these covariates in detail.

Table 4 List of covariates of the survey and the estimates

Covariate	**Description**	**Type**	**Values**
year_id	year of the estimates (centered at 1999)	Continuous	1-18
location_id	country indicator of the estimates	Categorical	47 countries
region_id	subregion indicator of the estimates	Categorical	5 subregions
sex_id	sex indicator of the estimates	Binary	1: male 2: female
age_group_id	age group indicator of the estimates	Categorical	1: 15-24 2: 25-49 3: 15-49
survey_type	Survey type indicator	Categorical	1: DHS/AIS 2: MICS 3: Other

3.2 Multiple Imputation Methods to be Evaluated

In this study, we examine and compare the performance of 7 different MI algorithms for TSCS missing data, namely, ***PAN***, ***JOMO***, ***Amelia*** and four ***MICE*** *algorithms* with different univariate modeling methods. As described in the literature review, PAN, JOMO and Amelia are all MIJM. PAN and JOMO use full Bayesian MCMC (the **IP**) algorithm to estimate the parameters of the posterior and to draw imputed values from the posterior[16,64,81] whereas Amelia uses EMB algorithm to estimate the parameters of the posterior.[59] Different from PAN, which assumes

fixed level-1 residual covariance matrix across different clusters, JOMO relaxes the assumption by allowing the covariance matrix of residuals at level 1 to vary across clusters.[16,64] The heterogeneous covariance matrix of residuals helps better capture the random effects between incomplete level-1 covariates across clusers.[11,16] However, the JOMO algorithm is very computationally intensive and is expected to take much longer time to run compared with the PAN algorithm.[11] To implement PAN, JOMO and Amelia, R functions "*panImpute*" and "*jomoImpute*" in package "mitml"[82] and function "amelia" in package "amelia"[70] are used.

As described in the literature review, MICE is more flexible than MIJM because each incomplete variable is modeled separately and in turn in MICE. Since the incomplete variables to be imputed are all two-level continuous variables, four univariate models for two-level continuous variable are chosen to impute the missing data, namely, ***mice.2l.pan***, ***mice.2l.norm***, ***mice.2l.lmer*** and ***mice.2l.pmm***. The methods *mice.2l.pan* and *mice.2l.norm* impute the univariate missing data using a two-level normal model with homogenous and heterogeneous within group variance respectively.[21] They both implement the Gibbs sampler to fit the two-level normal model (see details in ***section 2.3.3***). The *mice.2l.lmer* method uses univariate linear mixed model (using R function "lmer") to predict the univariate missing data. The predictions take into account the uncertainty of the model parameters, the random effects and the model residuals.[21,83] Based on *mice.2l.lmer* ,the *mice.2l.pmm* uses predictive mean matching based on predictions from the linear mixed model above. For each missing value, 5 donors are selected based on proximity to the predicted values and one of the 5 donors is randomly selected as the imputed value.[21,40] All the four MICE methods are implemented using R function "mice" in the "mice" package.[84]

3.3 Imputation Models

When comparing the performance of different MI methods, we use the same imputation model which includes all the 16 key indicators of HIV/AIDS knowledge and attitudes, the important country-level covariates (**Table 2**) and the covariates of the survey and the estimates (**Table 4**). Among these variables, the 16 key indicators are the target variables which contain missing values and the other variables are all auxiliary variables which are completely observed. Among the variables included in the imputation models, *location_id* is the cluster variable and *year_id* is the time variable whose slope is allowed to vary across clusters. Therefore, the primary imputation model is a two-level model with random slope and random intercept.

As discussed in the literature review, there are still discrepancies among researchers on whether including cluster means of variables improves imputation.[64,67,68] Therefore, building upon the model above, we further include the cluster means of key indicators and of covariates to examine the impact of adding these cluster means on the imputation performance. In addition, it is recommended that one can include as many auxiliary variables as possible to make the MAR assumption more plausible, which is the so-called inclusive strategy.[85] However, study also shows that including auxiliary variables that have too many missing values harms the efficiency of imputation.[86] Therefore, building upon the primary model, we further include auxiliary variables (other HIV/AIDS knowledge and attitudes indicators) with different proportion of missingness to examine the impact of adding these incomplete auxiliary variables on the imputation.

To examine the impact of two additional modeling strategies mentioned above, we first pick the best-performing MI methods for the primary imputation model and then use these methods to

conduct MI with two additional imputation models including cluster means and additional incomplete auxiliary variables respectively.

3.4 Implementation of the MI Methods

For all the MI methods, we use *location_id* and *year_id* as cluster and time variable respectively, model the time effect linearly, allow the time effect to vary across countries, and impute the missing values 1000 times. For **Pan** and **JOMO**, we set the burn-in to be 1000 iterations and draw one imputed value every 100 iterations to make sure that the draws are independent. For **Amelia**, we follow the advice by Honaker et al. and add a small ridge prior (1% of the total number of observations) to stabilize the imputation algorithm.[87] To reduce the running time, we use 1000 machines to run Amelia parallelly and combine the results afterwards. For the four **MICE** methods, we impute the variables in a monotone sequence and use 20 iterations for each imputation. The random seed is set to be 2019 for all the random processes.

3.5 Evaluation Methods

3.5.1 10-fold cross-validation

To evaluate the performance of different MI methods and imputation models, we employ a 10-fold cross-validation (CV) approach. In a 10-fold CV, the observed data are randomly divided into 10 mutually exclusive subsets (the folds) of approximately equal size. The imputation model is trained and tested 10 times; each time, one subset of observed data is left out and used as test set and the imputation model is trained using the remaining 9 subsets of observed data.[88]

3.5.2 simulation of missing data for 10-fold cross-validation

To simulate the missing data for 10-fold CV, we first randomly divide the observed data of each key indicator into 10 groups and then we randomly select one group from each indicator and remove the observed data in the selected groups. The selection of groups is repeated 10 times,

resulting in 10 datasets with different simulated missing data. Since the groups are selected without replacement, each group will only be selected once and no observed data will be removed twice. The missing data are MCAR.

3.5.3 indicators of performance

To evaluate the performance of different MI methods, we choose the root mean squared error ($RMSE$) and the percentage of 95% credible intervals covering the true data, a.k.a. the coverage rate of 95% CI (CR_{95}) as two major indicators of performance.

The $RMSE$ is the root of the mean of squared difference between the imputed values ($\hat{y}_i$) and the true values (y_i^{mis}) that are artificially removed. Mathematically,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n_{mis}} \left(y_i^{mis} - \hat{y}_i\right)^2}{n_{mis}}}$$

where y_i^{mis} is the removed true data value for unit i, $\hat{y}_i$ is the imputed value for unit i and n_{mis} is the total number of true values that are artificially removed. The $RMSE$ measures the average deviance of imputed values from true values and takes into account the trade-off between unbiasedness and efficiency of the estimator by combining information about both bias and variance of the estimator.[75] In general, an imputation method is considered to be better if the $RMSE$ of its imputed values is smaller. It is a commonly used performance indicator for imputation methods.[78,79] Although we impute the country-level missing proportions of key indicators in logit scale, we calculate the RMSE in the original scale of the country-level proportions, i.e. between 0 to 1. Therefore, in this study, a RMSE is considered good if it is smaller than 0.05 suggesting that the imputation method has less than ±5% imputation error on average.

The CR_{95} measures the relative frequency with which the 95% credible intervals of the imputed values covers the true values. By definition, the CR_{95} should be close to 95% if the imputation method is appropriate. Mathematically,

$$CR_{95} = \sum_{i=1}^{n_{mis}} \frac{I(y_i^{mis} \in [\hat{y}_{2.5th}, \hat{y}_{97.5th}])}{n_{mis}}$$

where y_i^{mis} is the removed true data value for unit i, n_{mis} is the total number of true values that are artificially removed, $\hat{y}_{2.5^{th}}$ and $\hat{y}_{97.5^{th}}$ are 2.5th and 97.5th percentile of the imputed values of unit i. As mentioned above, the CR_{95} should be close to 95% by definition. According to Grund et al., the MI method is considered suboptimal if CR_{95} is below 90% or very close 100%, suggesting that the distribution of imputed values are off or the variance of imputed data are too large.[75]

In addition to $RMSE$ and CR_{95}, the average running time (ART) of an MI method is used as a supplementary performance indicator. In 10-fold CV, each MI model will be run 10 times and the ART is the mean of the 10 running times. Mathematically,

$$ART = \frac{\sum_{k=1}^{10} RT_k}{10}$$

where RT_k is the running time of MI method for the k^{th}-fold CV. Although the running time of an MI method is practically important, it depends on computational power of the machine and on the selection of parameters of the MI model (e.g., number of iterations). Therefore, we only use ART as a practical guidance on model selection.

3.5.4 comparison of model performance

In each fold of the 10-fold CV, a different 10% of the observed data are purposefully removed (the testing set) and then imputed by an imputation method fitted by the remaining 90% of observed data (the training set). In each fold, all performance indicators are calculated using the testing set and the same performance indicators are averaged over the 10 folds to produce the final performance indicators. An MI method is considered better than another if it has a smaller $RMSE$ and a CR_{95} closer to 95%. To combine the two performance indicators, we calculate a performance score (PS) using the following formula

$$PS_m = 0.7 * \left(\frac{RMSE_m}{0.05}\right) + 0.3 * (\frac{|CR_{95} - 0.95|_m}{0.005})$$

where PS_m is the performance score for method m. We chose 0.05 and 0.005 because they are the cutoff values for good $RMSE$ and difference between CR_{95} and 0.95. Since we value small $RMSE$ more than small difference between CR_{95} and 0.95, we give 70% and 30% weight to $RMSE$ and CR_{95} respectively. We prefer MI method with smaller PS to the one with larger PS.

When taking ART of an MI method into consideration, we use the following formula

$$PS_m^t = \frac{7}{11} * \left(\frac{RMSE_m}{0.05}\right) + \frac{3}{11} * \left(\frac{|CR_{95} - 0.95|_m}{0.005}\right) + \frac{1}{11} * (\frac{ART_m}{60})$$

where 60 minutes is considered cutoff values for good ART. Since ART (in minutes) depends on computational power and utilization of the machine, it can be quite random. Therefore, we give a small weight to ART and we only use PS^t as a supplementary performance measurement.

4. Results

4.1 The Pattern of Missing Data

In the original dataset, the overall missing rate of the 16 key indicators of HIV/AIDS knowledge and attitudes is 11.8%. Among the 16 key indicators, *heard_aids* and *mtct_drug* have the smallest and the largest missing rate of 2.7% and 40.2% respectively. **Table 5** details the number and proportion of missingness by indicators. We see that 9 indicators have missing rate less than 10% and additional 5 indicators have missing rate less than 20%. ***Figure 3*** visualizes the missing rate by indicators and the pattern of missing data in the original dataset. We can see that there are 489 complete observations accounting for 49.7% of total observations and most missing patterns have very few observations.

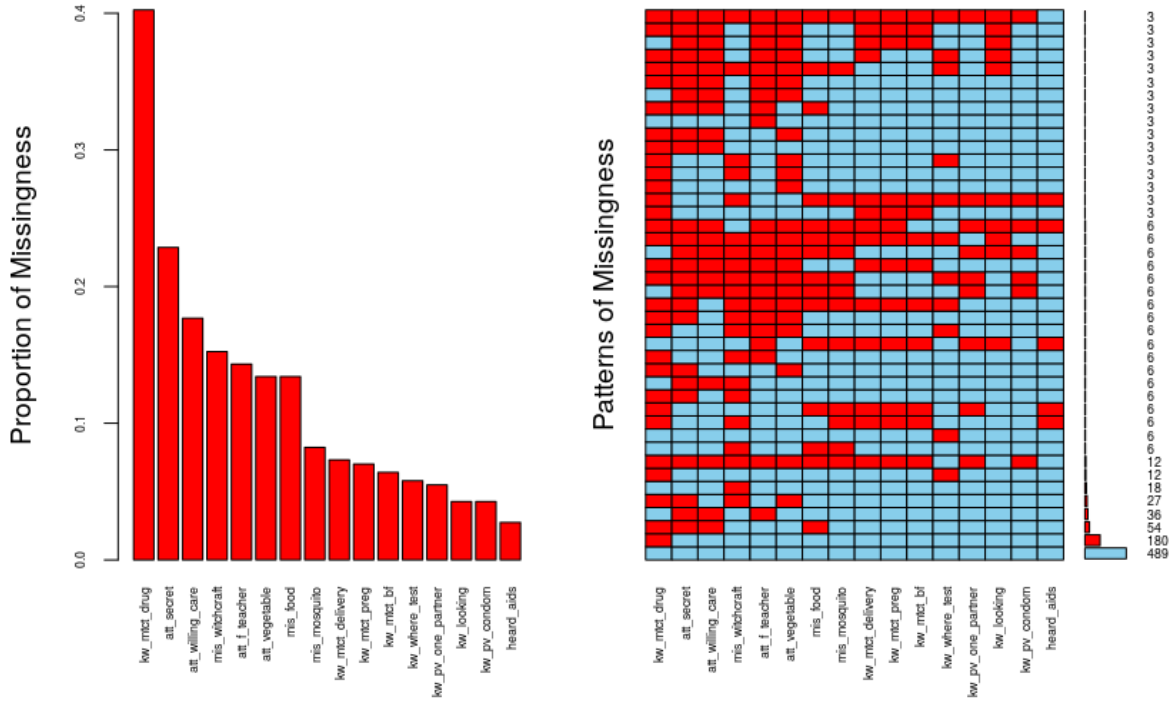


Figure 3 Proportion of missingness by variables and patterns of missingness

After removing 10% of observed data for each indicator, the overall missing rate of the simulated dataset becomes 20.7%. The number and proportion of missingness for each indicator in the simulated dataset are summarized in **Table 5**. In the simulated dataset, all indicators have more than 10% missingness and 7 indicators have more than 20% missingness.

Table 5 Number of missing data by variables

Key Indicators	Original dataset		Simulated dataset	
	Number of missingness	Proportion of missingness	Number of missingness	Proportion of missingness
heard_aids	27	2.7%	123	12.5%
kw_looking	42	4.3%	137	13.9%
kw_pv_condom	42	4.3%	137	13.9%
kw_pv_one_partner	54	5.5%	147	14.9%
kw_where_test	57	5.8%	150	15.2%
kw_mtct_bf	63	6.4%	156	15.9%
kw_mtct_preg	69	7.0%	161	16.4%
kw_mtct_delivery	72	7.3%	164	16.7%
mis_mosquito	81	8.2%	172	17.5%
att_vegetable	132	13.4%	218	22.2%
mis_food	132	13.4%	218	22.2%
att_f_teacher	141	14.3%	226	23.0%
mis_witchcraft	150	15.2%	234	23.8%
att_willing_care	174	17.7%	255	25.9%
att_secret	225	22.9%	301	30.6%
kw_mtct_drug	396	40.2%	455	46.2%

4.2 The Performance of MI Methods Using the Primary Imputation Model

The overall $RMSE$ and CR_{95} for each method are summarized in **Table 6** and the variable-specific $RMSE$ and CR_{95} are summarized in **Table 7** and **Table 8** respectively.

Table 6 The overall performance indicators for all MI methods

MI methods	RMSE	CR_{95}	ART (min)	PS	PS^t
Amelia	0.0391	0.9520	5	**0.6688**	0.6155
pan.100[1]	0.0378	0.9465	15	**0.7406**	0.6960
jomo.100	0.0204	0.9556	7053 (5.9 days)	**0.6226**	11.2523

[1] pan.100 and jomo.100 represent PAN and JOMO method with 1000 burn-ins and thinning factor of 100.

mice.2l.pan	0.0378	0.9475	2475 (1.7 days)	**0.6808**	4.3689
mice.2l.norm	0.0681	0.9583	5585 (3.9 days)	1.4489	9.7793
mice.2l.pmm	0.1050	0.9840	2939 (2.0 days)	3.5074	7.6415
mice.2l.lmer	0.1714	0.9467	1178 (0.8 days)	2.5985	4.1471

Table 7 The variable-specific $\boldsymbol{RMSE}$ for all MI methods

Indicators	**Amelia**	**Pan.100**	**Jomo.100**	**mice. 2l.pan**	**mice. 2l.norm**	**mice. 2l.pmm**	**mice. 2l.lmer**
heard_aids	0.024	0.023	0.012	0.023	0.040	0.078	0.099
kw_looking	0.033	0.032	0.017	0.032	0.071	0.078	0.175
kw_pv_condom	0.034	0.033	0.018	0.033	0.081	0.108	0.178
kw_pv_one_partner	0.038	0.036	0.018	0.036	0.079	0.105	0.156
kw_where_test	0.059	0.059	0.030	0.059	0.080	0.115	0.253
kw_mtct_bf	0.031	0.029	0.016	0.030	0.064	0.096	0.159
kw_mtct_preg	0.042	0.040	0.021	0.041	0.066	0.108	0.128
kw_mtct_delivery	0.028	0.027	0.014	0.027	0.068	0.084	0.146
mis_mosquito	0.032	0.030	0.017	0.030	0.056	0.097	0.160
att_vegetable	0.041	0.038	0.020	0.038	0.066	0.122	0.211
mis_food	0.028	0.027	0.015	0.026	0.063	0.087	0.150
att_f_teacher	0.036	0.034	0.017	0.034	0.057	0.104	0.183
mis_witchcraft	0.038	0.037	0.020	0.037	0.071	0.137	0.188
att_willing_care	0.039	0.037	0.019	0.037	0.059	0.112	0.134
att_secret	0.059	0.059	0.031	0.059	0.070	0.123	0.155
kw_mtct_drug	0.050	0.046	0.029	0.046	0.085	0.109	0.216

Table 8 The variable-specific $\boldsymbol{CR_{95}}$ for all MI methods

Indicators	**Amelia**	**Pan.100**	**Jomo.100**	**mice. 2l.pan**	**mice. 2l.norm**	**mice. 2l.pmm**	**mice. 2l.lmer**
heard_aids	0.927	0.930	0.948	0.933	0.951	0.959	0.936
kw_looking	0.954	0.949	0.963	0.948	0.962	0.989	0.948
kw_pv_condom	0.956	0.948	0.952	0.941	0.969	0.983	0.948
kw_pv_one_partner	0.941	0.944	0.951	0.945	0.953	0.980	0.944
kw_where_test	0.947	0.946	0.962	0.946	0.947	0.992	0.945
kw_mtct_bf	0.961	0.956	0.955	0.955	0.956	0.990	0.944
kw_mtct_preg	0.946	0.947	0.954	0.947	0.960	0.969	0.951
kw_mtct_delivery	0.957	0.950	0.953	0.951	0.962	0.995	0.948
mis_mosquito	0.959	0.952	0.948	0.950	0.945	0.981	0.951
att_vegetable	0.961	0.955	0.957	0.959	0.953	0.989	0.947
mis_food	0.961	0.950	0.958	0.957	0.957	0.993	0.946
att_f_teacher	0.961	0.953	0.956	0.954	0.968	0.989	0.949

mis_witchcraft	0.954	0.941	0.956	0.937	0.964	0.988	0.952
att_willing_care	0.949	0.952	0.957	0.953	0.967	0.975	0.942
att_secret	0.937	0.930	0.961	0.938	0.963	0.984	0.947
kw_mtct_drug	0.963	0.949	0.966	0.949	0.961	0.988	0.949

We can see from **Table 6** that among the 7 MI methods, JOMO has the smallest $RMSE$ (0.0204) and the smallest PS (0.537), suggesting that JOMO method performs the best among the 7 methods. However, the ART for JOMO method is 7053 minutes or 5.9 days, making this method immensely impractical. Based on the indicator PS^t, which takes into account the running time, JOMO is ranked the last among the 7 methods. Although having $RMSE$ higher than that of JOMO, Amelia, mice.2l.pan and PAN all have $RMSE$ smaller than 0.05 and CR_{95} very close to 95%. The PS scores for Amelia, mice.2l.pan and PAN are 0.642, 0.673 and 0.821 respectively, suggesting that these three methods also perform very well imputing the missing values in the dataset. However, the rest three methods, namely, mice.2l.norm, mice.2l.pmm and mice.2l.lmer, all have $RMSE$ greater than 0.05 and their PS scores are much higher than those of the other methods, suggesting that these three methods do not perform well imputing missing values in the dataset.

Regarding CR_{95}, except for mice.2l.pmm whose CR_{95} (0.984) is a bit far from 95%, all the other methods have CR_{95} quite close to 95%, which is assuring.

Among the 7 methods, Amelia has the smallest ART (5 min) and PAN comes next (15 min), making these two methods most practical. Based on PS^t, Amelia (0.584) and PAN (0.928) are much better than the other models due to their short running time.

Table 7 and **Table 8** summarize the variable-specific $RMSE$ and CR_{95}. The indicators are sorted based on proportion of missingness of the indicators, ranked from the top to the bottom from low

missingness to high missingness. Among the 16 key indicators, *heard_aids* has the smallest $RMSE$ and *att_secret, kw_where_test* and *kw_mtct_drug* have the largest $RMSE$. In general, the higher the proportion of missingness, the larger the $RMSE$ is because higher proportion of missingness usually leads to higher variance of the imputation. However, as shown in **Table 7**, this is not always the case because $RMSE$ not only depends on the variance but also depends on the bias. For instance, if the distribution of an indicator is far from normal, e.g. highly skewed, the bias of the imputations would be high. ***Figure 4*** compares the density of *kw_where_test* and *kw_mtct_bf.* Although the two indicators have similar missing rates of 15.2% and 15.9% respectively in the simulated dataset, *kw_where_test* has higher $RMSE$ than *kw_mtct_bf* does because the distribution of *kw_where_test* is more skewed.

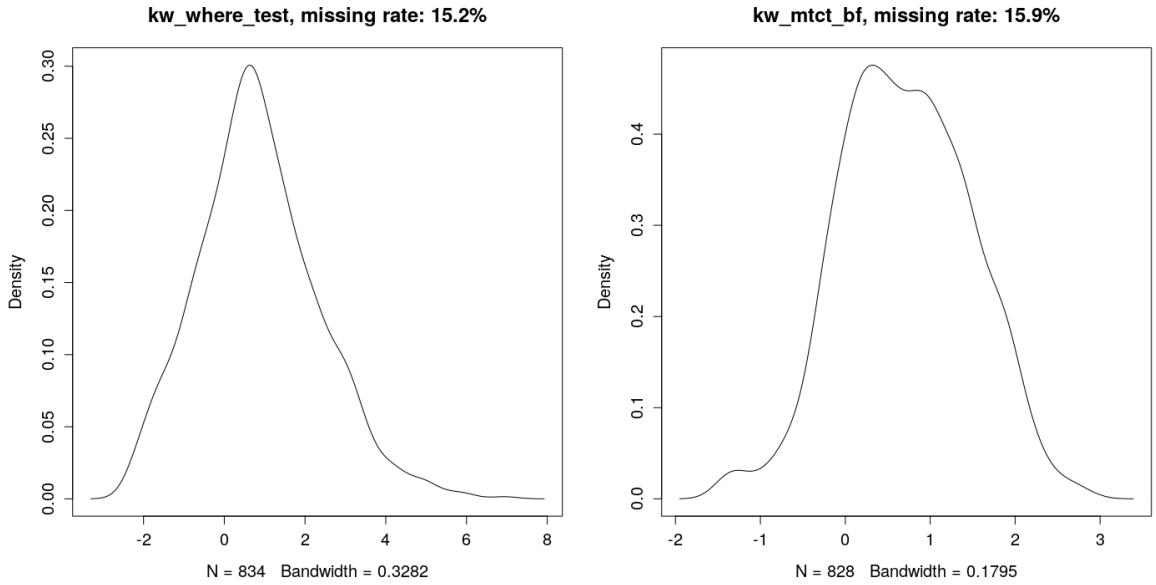


Figure 4 The density plots of kw_where_test and kw_mtct_bf

4.3 Diagnostics of the Imputation Methods

Although the $RMSE$ and CR_{95} are already important diagnostics for the imputation methods, other diagnostics are still informative of the performance of imputation methods. There are

common diagnostics for all imputation methods such as density plots of observed and imputed values. There are also method-specific diagnostics, such as potential scale reduction factor ($\hat{R}$) and trace plots of the MCMC chain to examine convergence for PAN and JOMO method. In this section, we provide important diagnostics for each imputation method.

4.3.1 diagnostics of Amelia methods

Figure 5 shows the density plots of observed (in black) and of imputed (in red) data for each indicator using Amelia. We can see that the distributions of the imputed data are very close to those of the observed data for nearly all the indicators, suggesting that Amelia produces valid imputations. For *kw_mtct_drug,* the mean of the observed and of the imputed data are slightly different, suggesting that imputations for this indicator are a bit off. This is probably due to the high missing rate (46.2%) of *kw_mtct_drug*.

Figure 6 shows the disperse plot which is a visual diagnostic of EM convergence. In the disperse plots, the EM chain is started at 5 different places and we can see that the Amelia EM algorithm converges well in both one and two dimensional spaces.

Figure 7 shows the overimputation plots for the key indicators in which each observed value is treated as missing and is imputed using the imputation model. In plots, the dots are the mean imputation and the vertical lines are the 90% confidence intervals. Ideally, 90% of the vertical lines should across the diagonal lines where the imputed values equal to the observed values. Based on the overimputation plots, we think the Amelia method works pretty well.

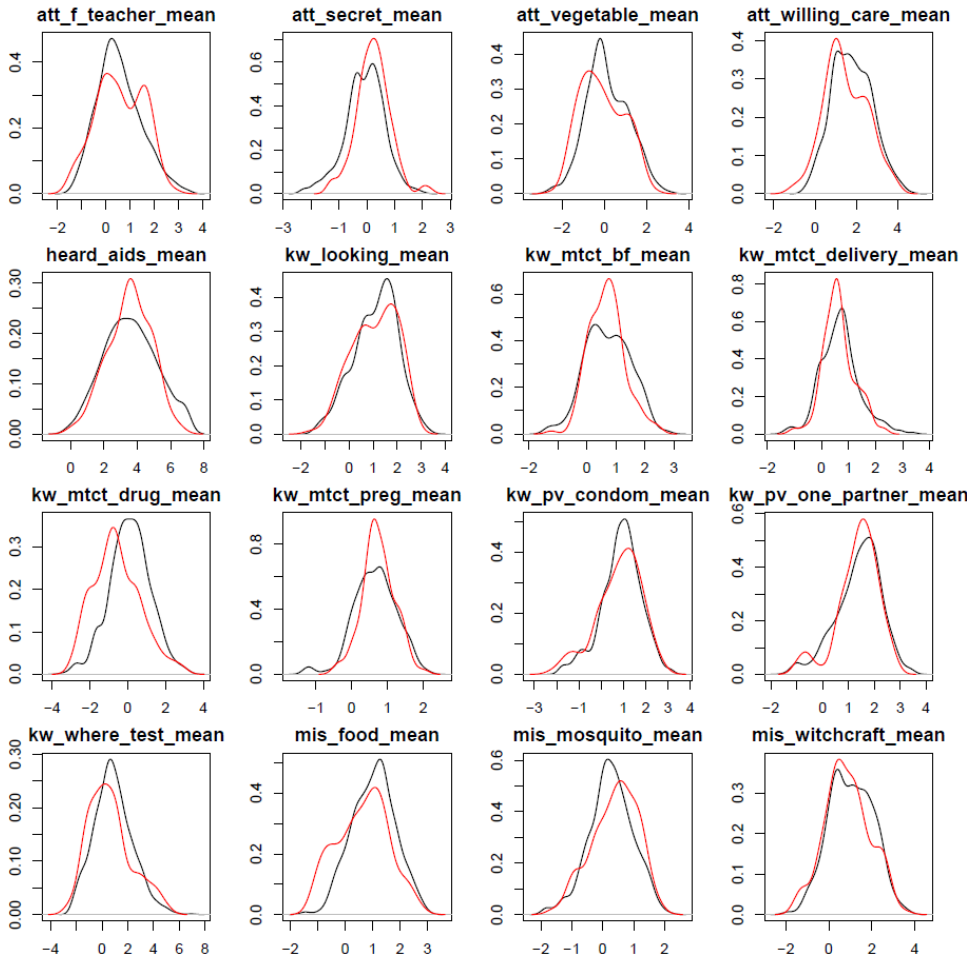


Figure 5 The density plots of observed (in black) and imputed (in red) values for each indicator using Amelia method

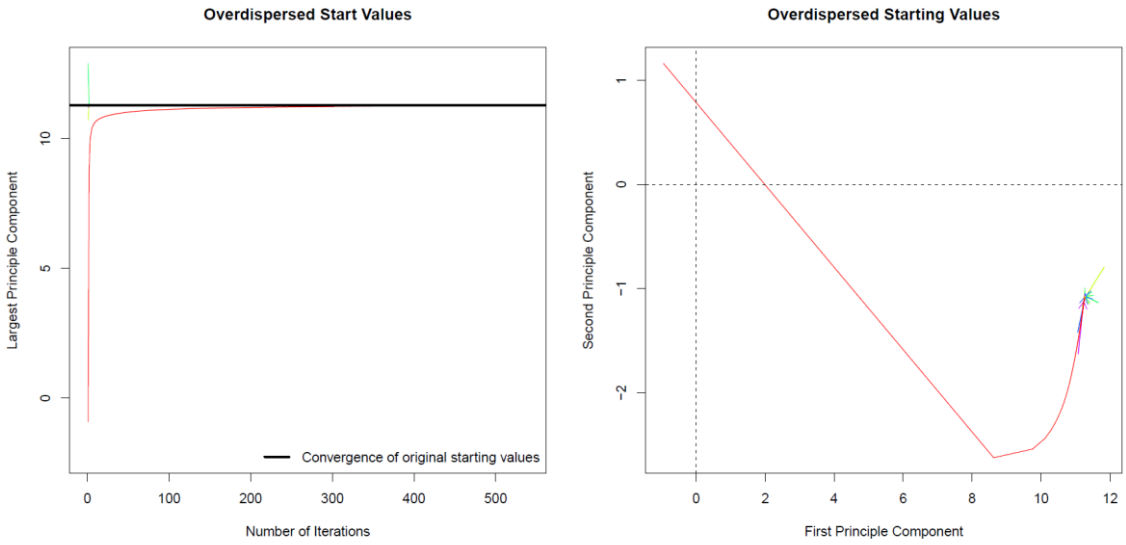
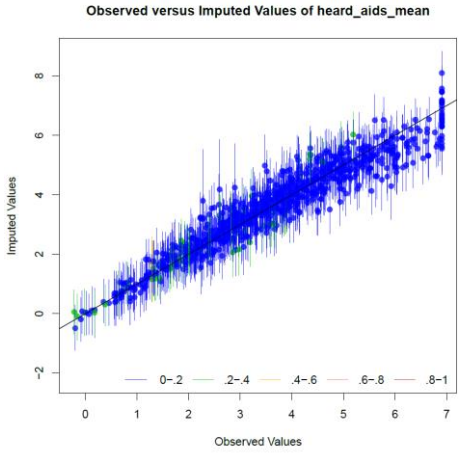
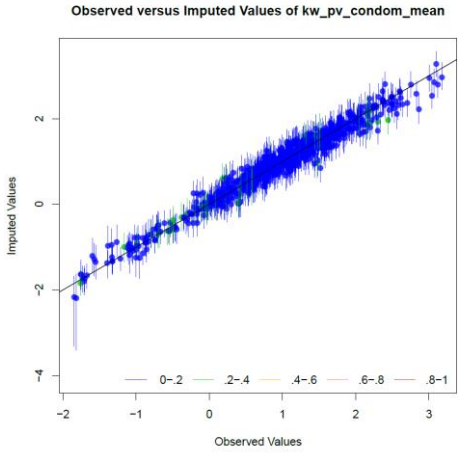
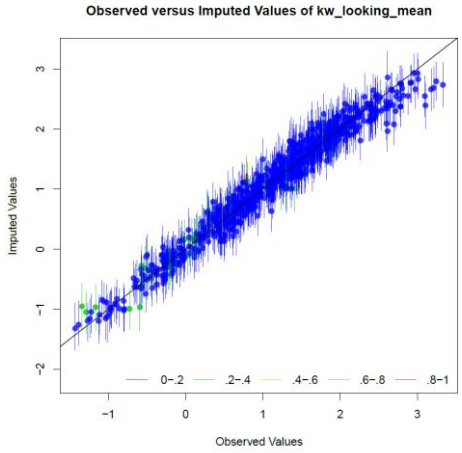
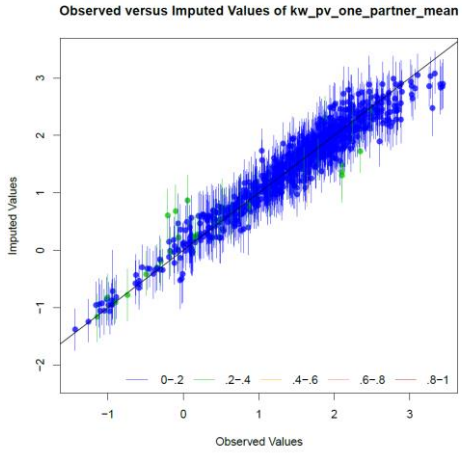


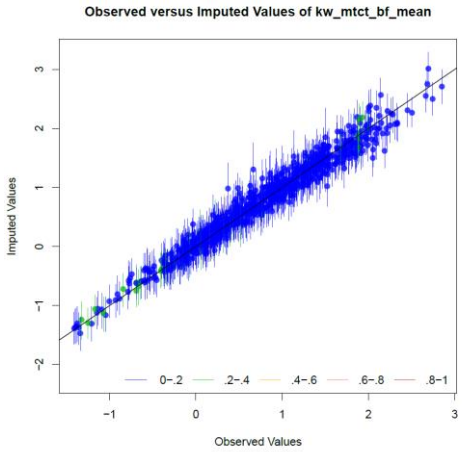
Figure 6 Disperse plots of one- and two-dimensional EM convergence

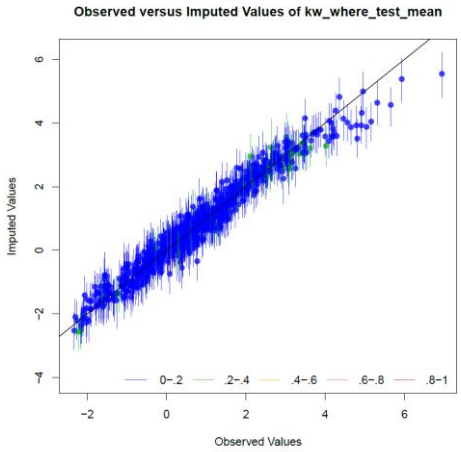

Observed versus Imputed Values of heard_aids_mean
Imputed Values
Observed Values
0−.2
.2−.4
.4−.6
.6−.8
.8−1

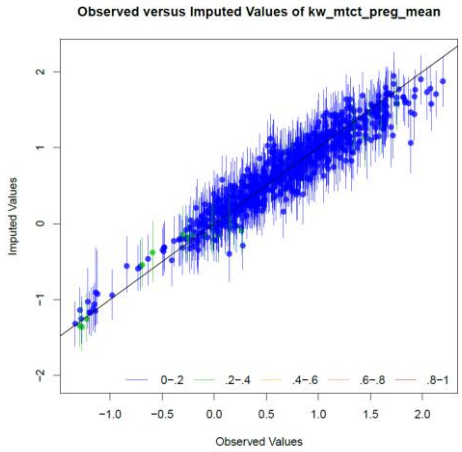

Observed versus Imputed Values of kw_pv_condom_mean
Imputed Values
Observed Values
0−.2
.2−.4
.4−.6
.6−.8
.8−1

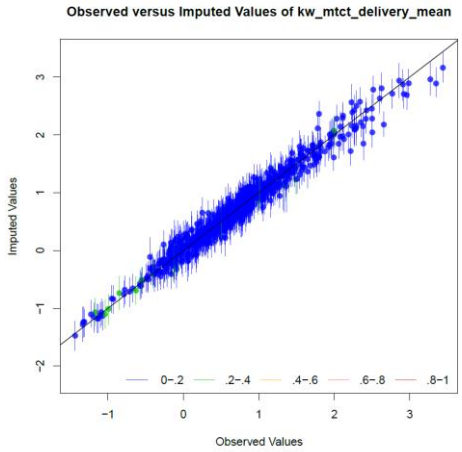

Observed versus Imputed Values of kw_looking_mean
Imputed Values
Observed Values
0−.2
.2−.4
.4−.6
.6−.8
.8−1


Observed versus Imputed Values of kw_pv_one_partner_mean
Imputed Values
Observed Values
0−.2
.2−.4
.4−.6
.6−.8
.8−1


Observed versus Imputed Values of kw_mtct_bf_mean
Imputed Values
Observed Values
0−.2
.2−.4
.4−.6
.6−.8
.8−1


Observed versus Imputed Values of kw_where_test_mean
Imputed Values
Observed Values
0−.2
.2−.4
.4−.6
.6−.8
.8−1


Observed versus Imputed Values of kw_mtct_preg_mean
Imputed Values
Observed Values
0−.2
.2−.4
.4−.6
.6−.8
.8−1


Observed versus Imputed Values of kw_mtct_delivery_mean
Imputed Values
Observed Values
0−.2
.2−.4
.4−.6
.6−.8
.8−1

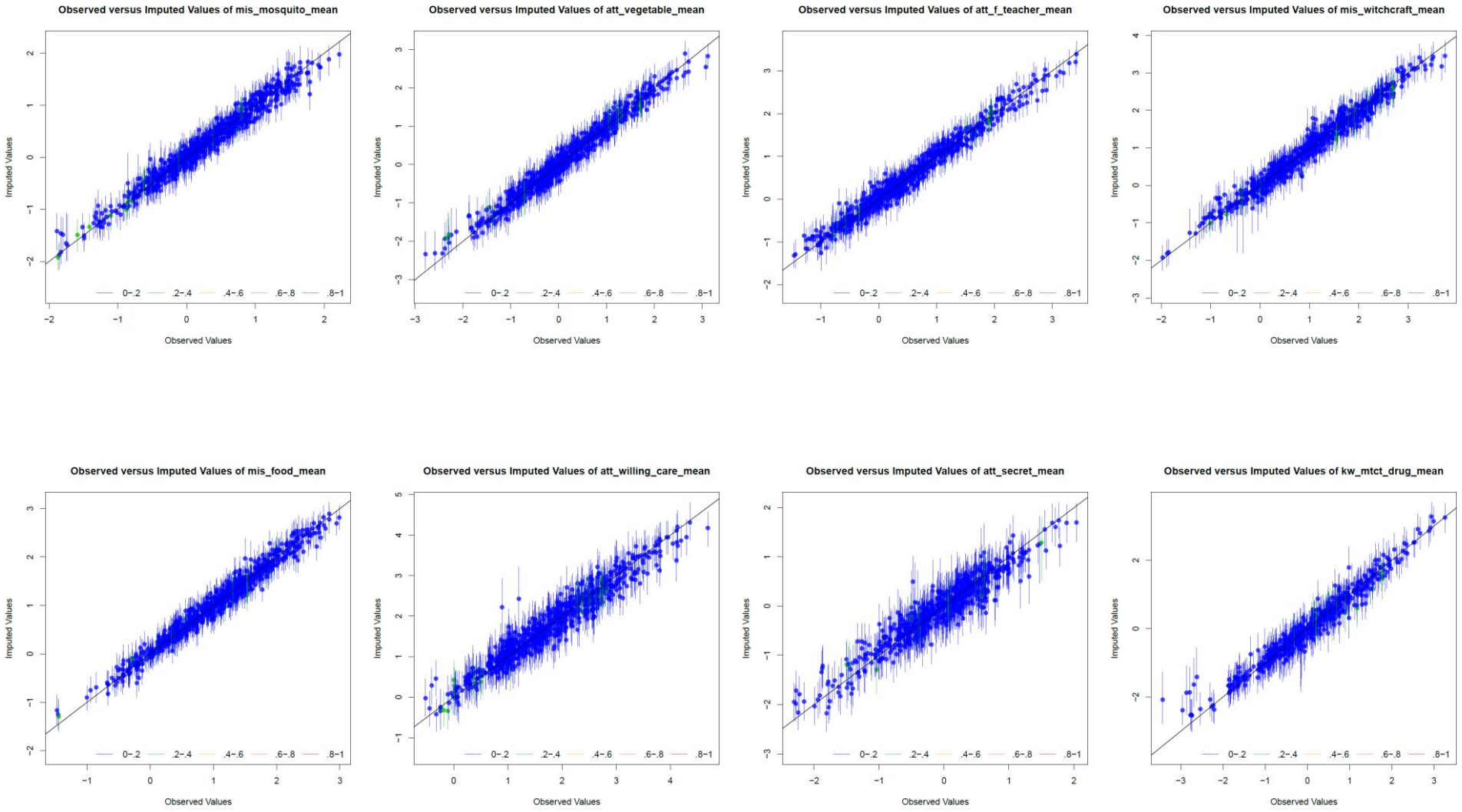


Figure 7 Overimputation plots for each key indicator

4.3.2 diagnostics of PAN and JOMO methods

Since PAN and JOMO use full Bayesian MCMC method to impute the missing data, the convergence of MCMC chains is important. In addition, draws from MCMC chain are inherently dependent especially for draws close to each other. However, for each missing value, the multiply imputed values should be independent draws.[39] To overcome the two problems, we discard the first 1000 draws hoping that the chains have converged after 1000 burn-ins and we draw an imputed value every 100 draws in the MCMC chain hoping that the imputed values are independent. After the imputation, we examine convergence of the chains using potential scale reduction factor ($\hat{R}$) and trace plots and examine independence of the imputed values using auto-correlation function (ACF) plots of the imputed values.

To calculate $\hat{R}$ we first discard the burn-in periods and divide the MCMC chain for each parameter into five segments. We then compare the variance within and between segments to detect shift of the chain. If the MCMC chain has converged, the $\hat{R}$ should be very close to 1. Practically, if $\hat{R} < 1.05$ for all parameters, we think the chains have converged.[81,89] **Table 9** shows the summary of $\hat{R}$ of different PAN and JOMO methods. In **Table 9**, pan.100 (jomo.100) and pan.200 (jomo.200) represent PAN (JOMO) with 1000 burn-ins and thinning factor of 100 and with 2000 burn-ins and thinning factor of 200 respectively; Beta, Psi and Sigma represent parameters of variables in the imputation model, variance and covariance of the random effects and variance of the residuals respectively. The maximum $\hat{R}$ of the parameters is way larger than 1.05 across all models, suggesting that the burn-ins are not enough and none of the models has converged. **Table 10** shows the summary of autocorrelation of the parameters in four models. In **Table 10**, k represents the number of iterations per imputation (the thinning factor). We can see that the autocorrelation between adjacent draws of the parameters are high. For variable with

maximum autocorrelation, the draws 400 iterations apart still have very high correlation, suggesting that thinning factors of 100 or even 200 are far from sufficient.

Table 9 Summary of $\widehat{\boldsymbol{R}}$ of different PAN and JOMO methods

		Min	25%	Mean	Median	75%	Max
pan.100	Beta:	1.000	1.001	1.025	1.003	1.009	2.170
	Psi:	1.000	1.001	1.002	1.001	1.002	1.021
	Sigma:	1.000	1.000	1.000	1.000	1.000	1.002
pan.200	Beta:	1.000	1.000	1.031	1.001	1.003	2.953
	Psi:	1.000	1.000	1.001	1.000	1.001	1.020
	Sigma:	1.000	1.000	1.000	1.000	1.000	1.002
jomo.100	Beta:	1.001	1.030	1.556	1.128	1.508	7.904
	Psi:	1.019	1.206	2.132	1.538	3.207	4.473
	Sigma:	1.008	1.033	1.471	1.074	1.520	4.577
jomo.200	Beta:	1.000	1.024	1.674	1.079	1.687	7.882
	Psi:	1.069	1.333	2.171	1.710	3.051	4.381
	Sigma:	1.006	1.018	1.325	1.048	1.361	3.407

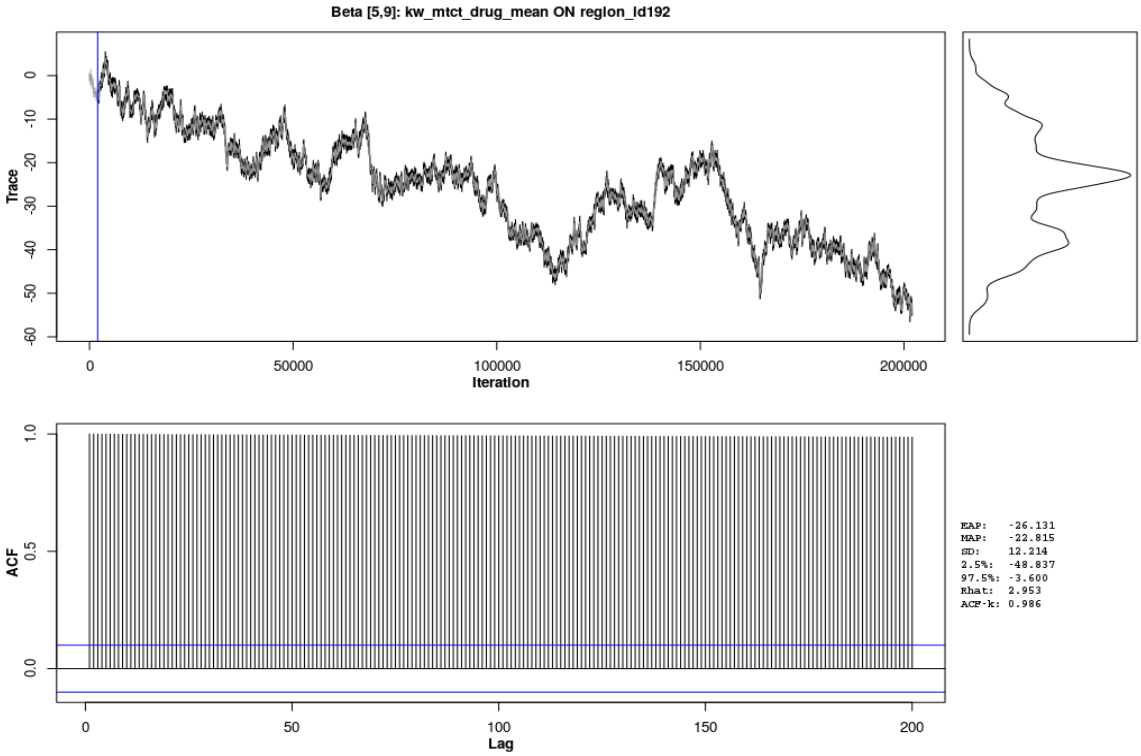
Table 10 Summary of ACF of different PAN and JOMO methods

		Mean			**Max**		
	Parameters	**Lag1**	**Lagk**	**Lag2k**	**Lag1**	**Lagk**	**Lag2k**
pan.100 (k=100)	Beta:	0.845	0.224	0.132	1.000	0.983	0.971
	Psi:	0.234	0.059	0.034	0.707	0.467	0.384
	Sigma:	0.353	0.015	0.008	0.627	0.052	0.044
pan.200 (k=200)	Beta:	0.845	0.132	0.067	1.000	0.989	0.979
	Psi:	0.230	0.035	0.017	0.709	0.377	0.259
	Sigma:	0.352	0.007	0.002	0.625	0.031	0.020
jomo.100 (k=100)	Beta:	0.955	0.665	0.583	1.000	0.998	0.996
	Psi:	0.568	0.530	0.523	0.935	0.933	0.932
	Sigma:	0.238	0.109	0.091	0.626	0.492	0.440
jomo.200 (k=200)	Beta:	0.956	0.601	0.525	1.000	0.997	0.995
	Psi:	0.606	0.587	0.582	0.943	0.937	0.932
	Sigma:	0.265	0.100	0.077	0.644	0.437	0.372

Although $\hat{R}$ is useful, Geyer argues that large $\hat{R}$ does not necessarily indicates poor convergence and examining the trace plots of the parameters is still important.[90] ***Figure 8*** shows trace and ACF plots of the parameters with the largest $\hat{R}$ for pan.200 (top) and jomo.200 (bottom)

respectively. For both parameters, we can see that the ACFs are so close to 1 that the chain almost becomes a "random walk" process making convergence almost impossible. We can see that even after 200000 iterations, the chain of the parameter still does not converge and the correlation between draws 200 iterations apart is still close to 1.

Based on $\hat{R}$, ACF and trace plots, compared with PAN, JOMO would require longer burn-ins and larger thinning factor to reach convergence and to produce independent imputations. Given the already long running time of JOMO (over 9 days for jomo.200), it seems impractical for JOMO to reach convergence and to produce independent imputations. Therefore, we run PAN again with larger burn-ins and thinning factor. Inspired by Grund et al. who implemented PAN with 50000 burn-ins and thinning factor of 5000 (pan.5k) and found that the model converged well and produced independent imputaitons,[81] we tried to run PAN with the same parameters hoping that the model will converge.



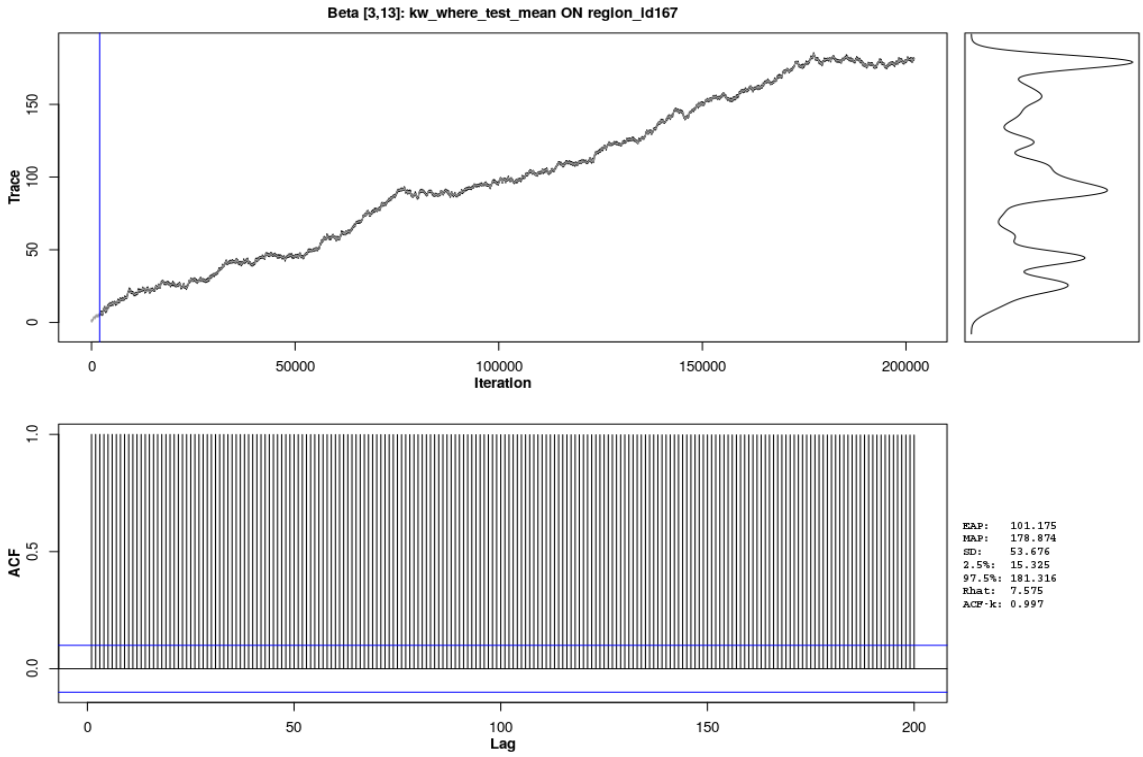


Figure 8 Trace and ACF plots of the parameters with the largest $\widehat{R}$ for pan.200 (top) and jomo.200 (bottom)

Table 11 Summary of $\widehat{R}$ and ACF of pan.5k

	$\widehat{R}$						ACF					
							Mean			Max		
	Min	**25%**	**Mean**	**Median**	**75%**	**Max**	**Lag1**	**Lagk**	**Lag2k**	**Lag1**	**Lagk**	**Lag2k**
Beta:	1.000	1.000	1.011	1.001	1.002	1.634	0.845	0.012	0.008	1.000	0.904	0.812
Psi:	1.000	1.000	1.000	1.000	1.000	1.004	0.231	-0.001	0.000	0.666	0.013	0.016
Sigma:	1.000	1.000	1.000	1.000	1.000	1.000	0.352	0.000	0.000	0.624	0.004	0.004

Table 11 and ***Figure 9*** show the summary of $\hat{R}$ and ACF and the trace and ACF plots for *pan.5k* respectively. We can see that although most parameters have converged, there are still a few that do not converge even with such long burn-ins and lag between imputations. Therefore, the usefulness of PAN and JOMO methods is limited due to the convergence issue when imputation model is complex.

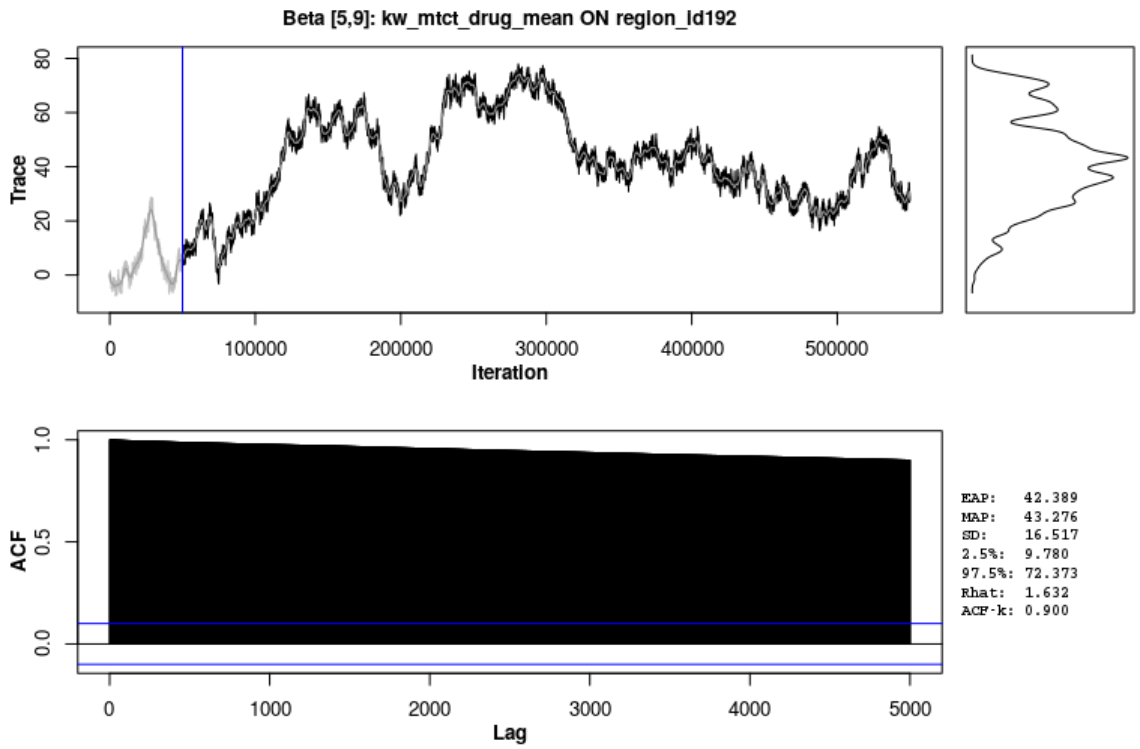


Figure 9 Trace and ACF plots of the parameter with the largest $\widehat{R}$ for pan.5k

Figure 10 shows the density plots of the observed and imputed values for each indicator for pan.200 method. Although the distributions of the imputed values approximate those of the observed values for many variables, the height of the density curve for a few variables, such as *heard_aids*, *kw_mtct_bf*, *mis_mosquito*, and *att_f_teacher*, is a bit off, suggesting suboptimal imputations for these indicators. More seriously, distribution of the imputed values for *kw_mtct_drug* has a very long tail to the right, suggesting that the imputation model does not converge and the imputations of this indicator are questionable.

Figure 11 shows the density plots of the observed and imputed values for each indicator for jomo.200 method. These plots show that jomo.200 has the same problems as the pan.200 method, suggesting that convergence of the full-Bayesian imputation methods (i.e., PAN and JOMO) can be crucial to the validity of imputations.

Figure 12 shows the density plots of observed and imputed values for *kw_mtct_drug* using pan.200 (left panel) and pan.5k (right panel), respectively. The distribution's tail of the imputed values shifts from the left to the right, further proving that imputations from full-Bayesian imputation methods can be unstable when the imputation model does not reach convergence.

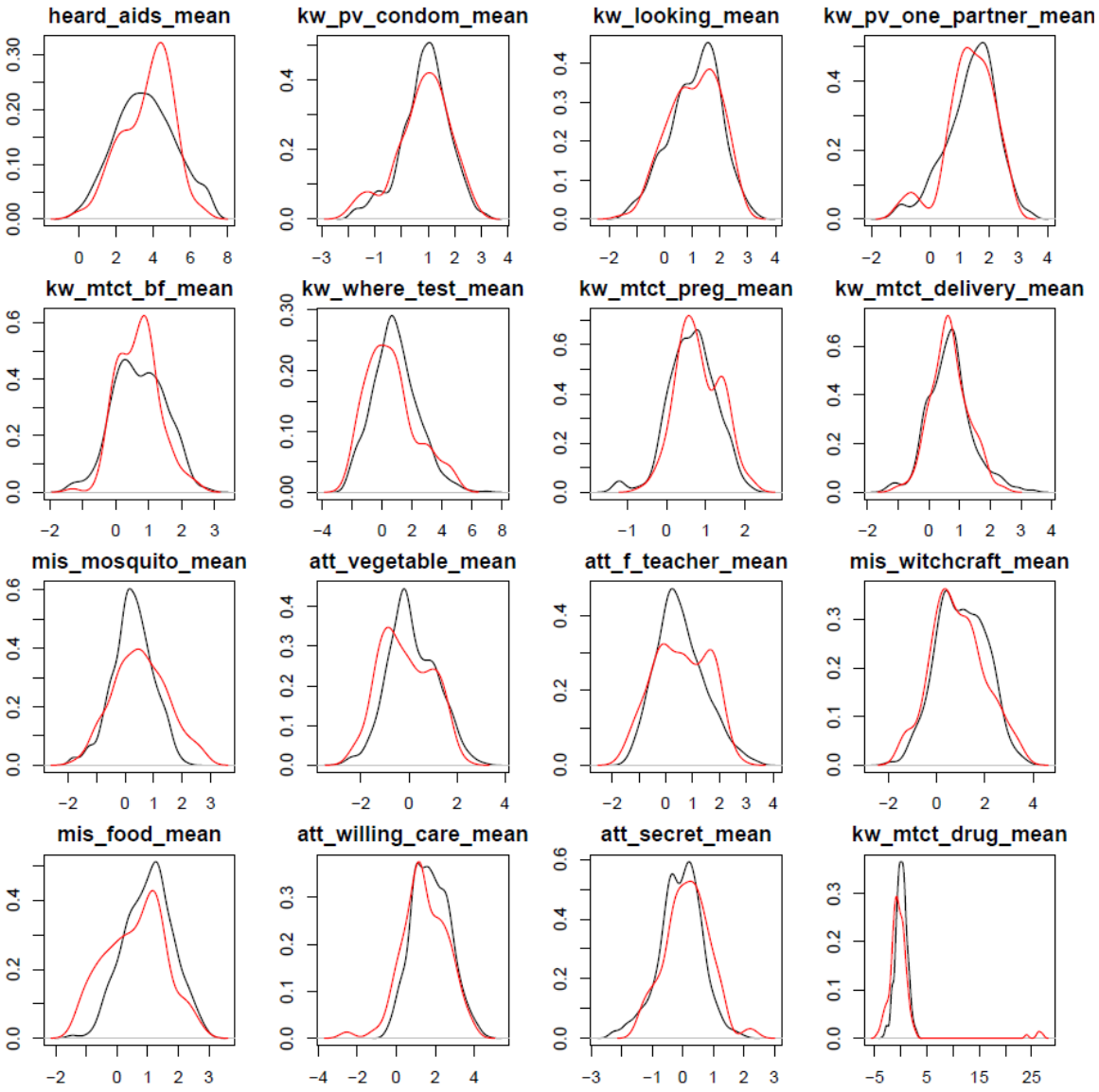


Figure 10 The density plots of observed (in black) and imputed (in red) values for each indicator using pan.200 method

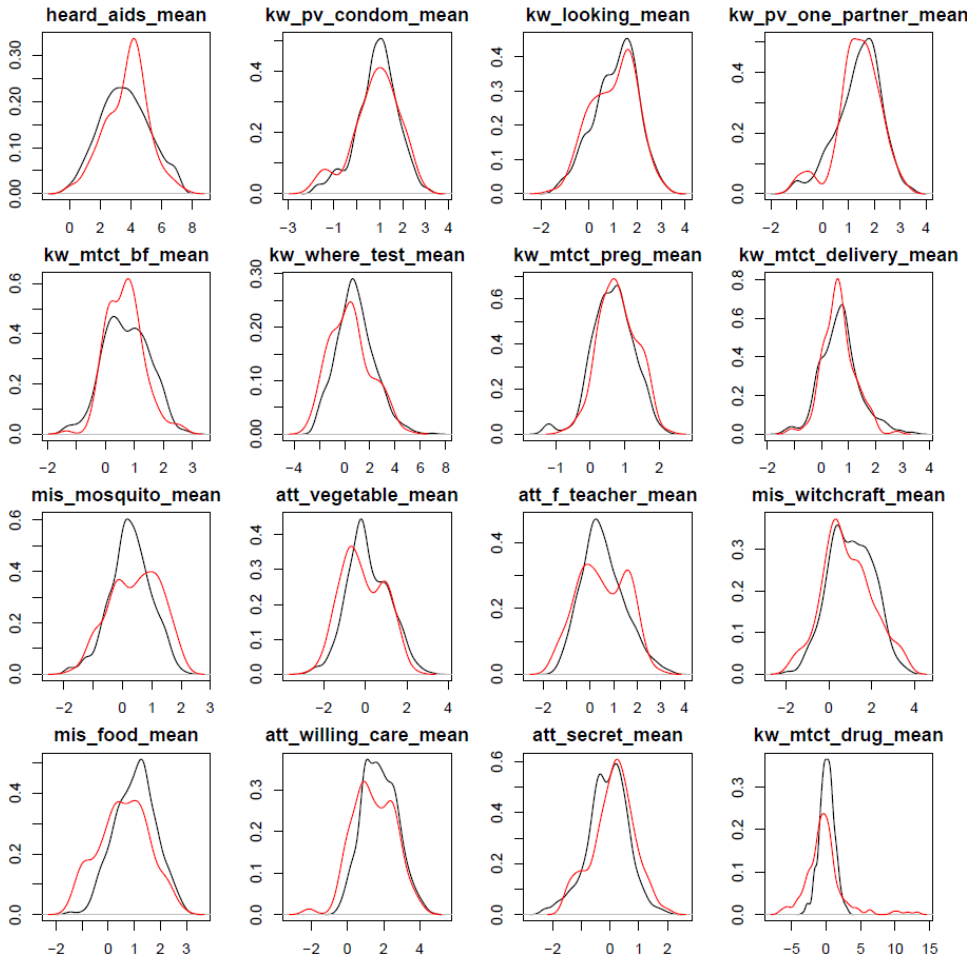


Figure 11 The density plots of observed (in black) and imputed (in red) values for each indicator using jomo.200 method

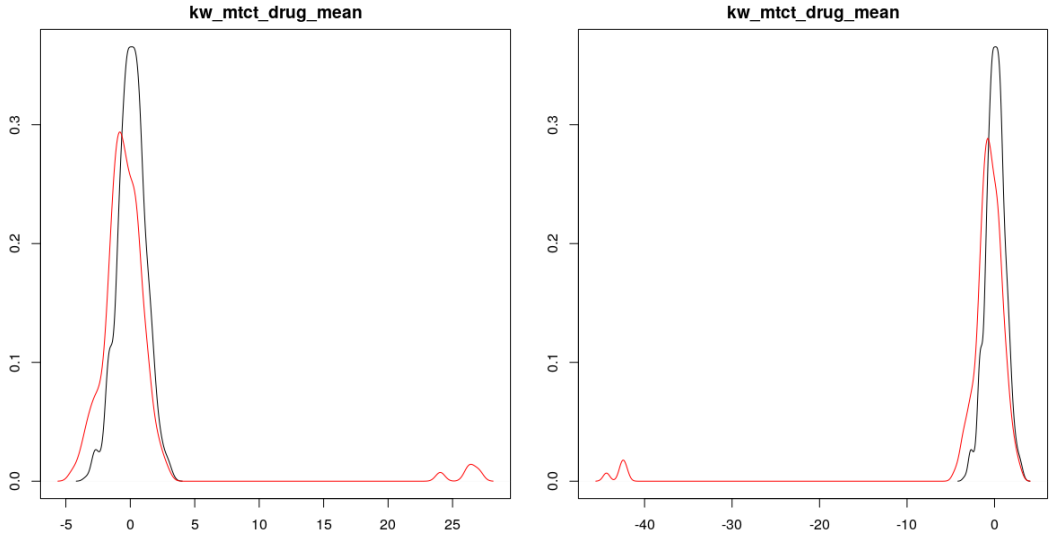


Figure 12 The density plots of observed (in black) and imputed (in red) values for kw_mtct_drug using pan.200 (left panel) and pan.5k (right panel)

4.3.3 diagnostics of MICE methods

Since *mice.2l.pan* is the only MICE method having good $RMSE$ and CR_{95}, we only provide diagnostics for *mice.2l.pan*. ***Figure 13*** shows the density plots of observed and imputed values for each indicator for *mice.2l.pan*. The plots are very similar to those of pan.200 except that the distribution of imputed values of *kw_mtct_drug* does not have long tail and is similar to the distribution of observed values. Compared with the PAN method, the mice.2l.pan method seems to perform better.

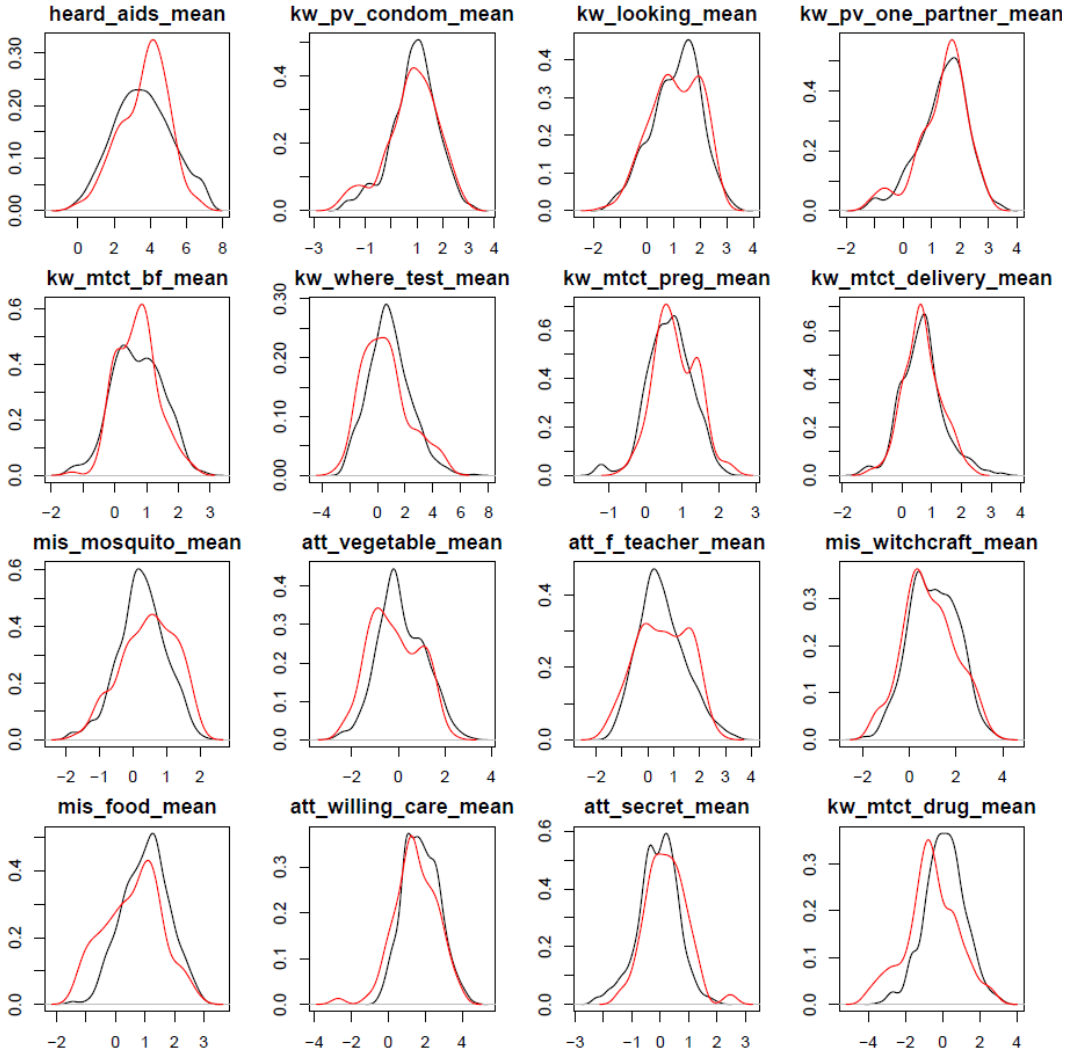
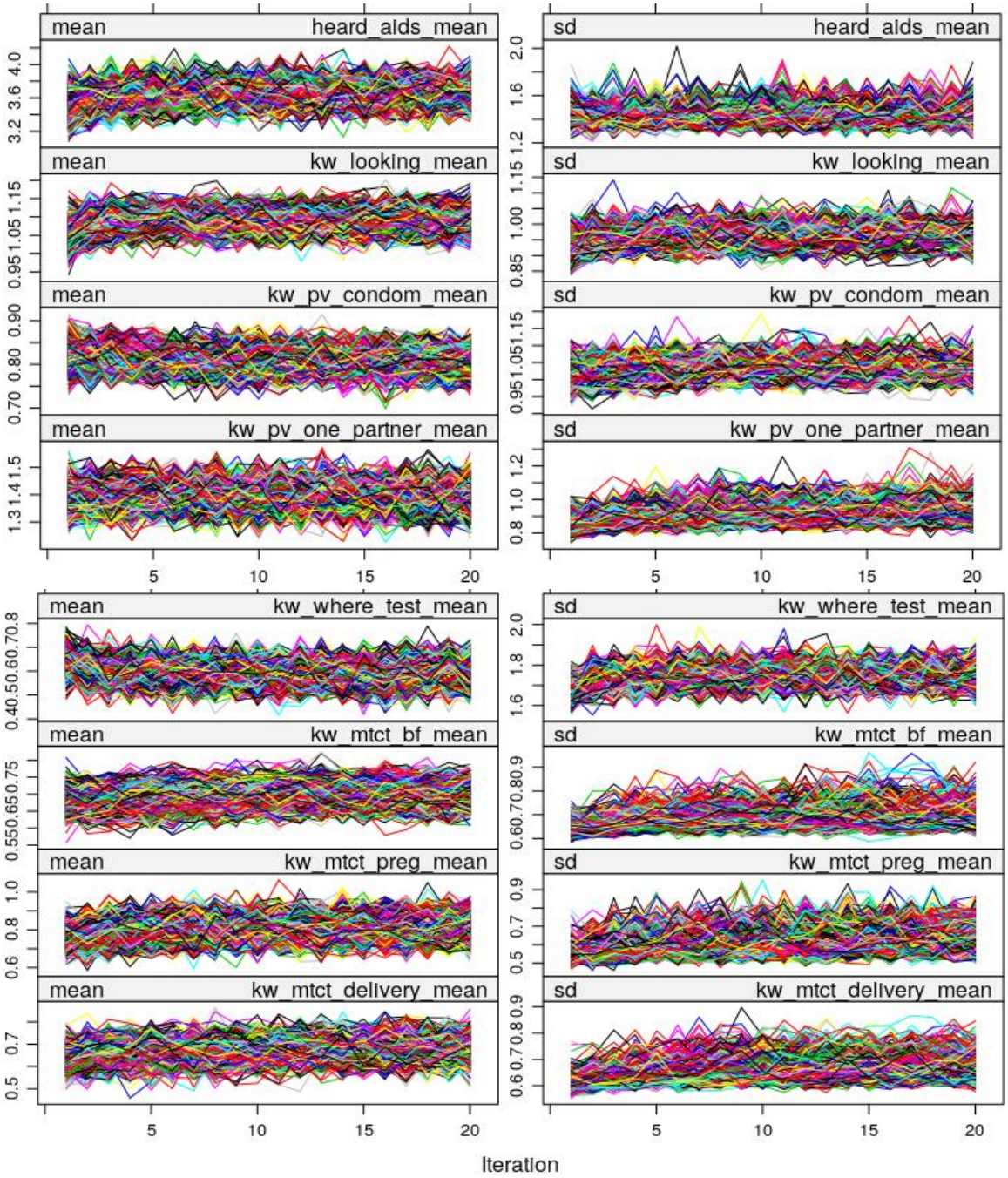


Figure 13 The density plots of observed (in black) and imputed (in red) values for each indicator using mice.2l.pan method

Since MICE implements an iterative MCMC algorithm, we need to examine the convergence of the model to make sure the imputed data are valid. ***Figure 14*** shows the trace plots of the mean and standard deviation of imputed values at each iteration for the incomplete variables. We can see that the traces of all variables are intermingle and free from any trend in the end, suggesting that the model has converged.[91]



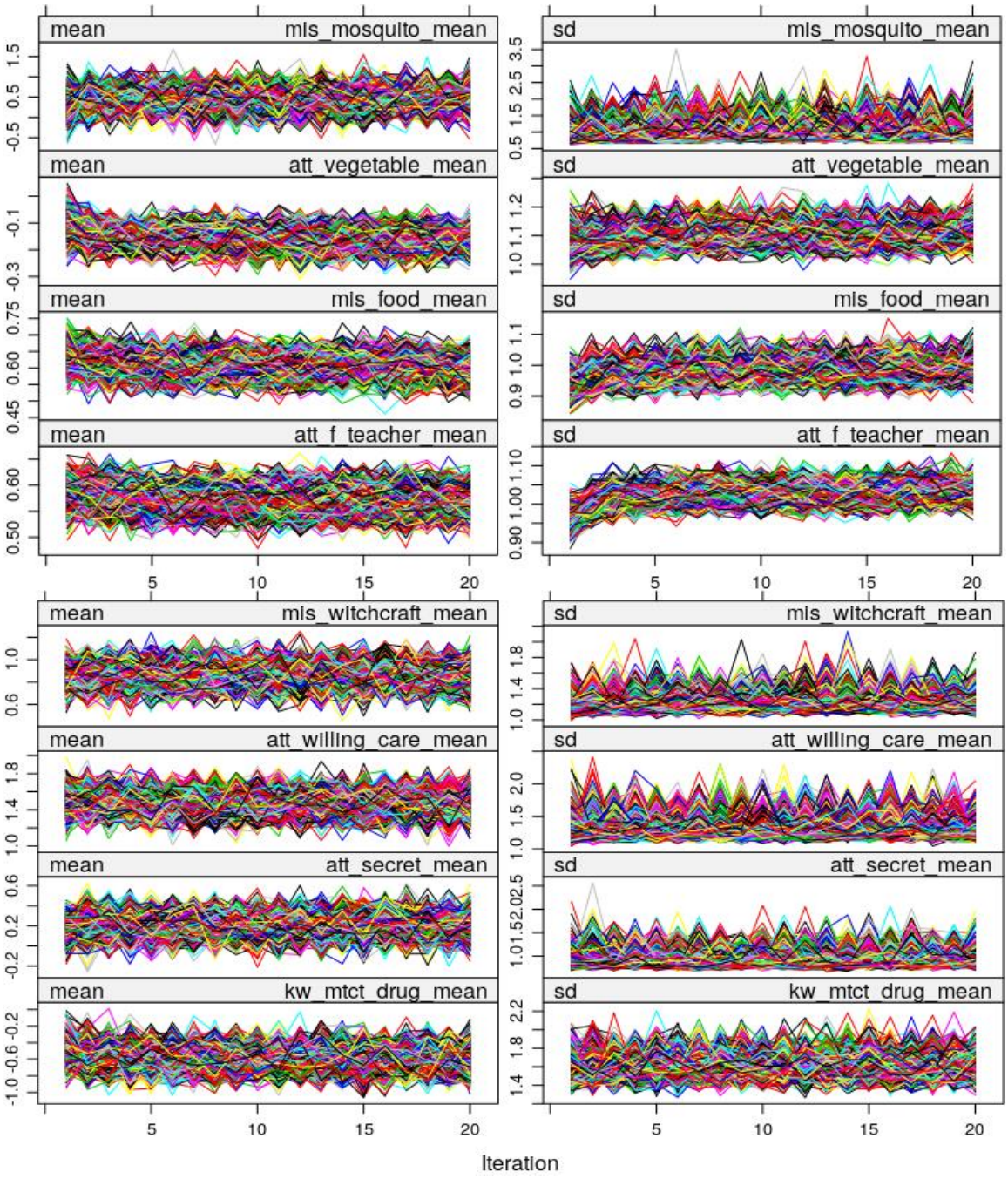


Figure 14 Trace plots of the mean and standard deviation of imputed values at each iteration for each indicator for mice.2l.pan method

4.4 The Impact of Including Cluster Means

When conducting MI using the primary model, we find that Amelia, pan.100, jomo.100 and mice.2l.pan have small $RMSE$ and CR_{95} close to 95%. We future exclude the JOMO method due

to its long running time and choose Amelia, pan.100 and mice.2l.pan to examine the impact of including cluster means into the imputation model. In the new imputation model, we include cluster means of the 16 key indicators and the complete country-level covariates of HIV/AIDS.

Table 12 compares $RMSE$, CR_{95}, PS and PS^t of the three MI methods using the primary imputation model versus the imputation model including the cluster means. From **Table 12** we can see that including cluster means in the imputation model has little impact on the $RMSE$ or the CR_{95} but does significantly increase the running time.

Table 12 Comparison of $\boldsymbol{RMSE}$, $\boldsymbol{CR_{95}}$, $\boldsymbol{PS}$ and $\boldsymbol{PS^t}$ of the three MI methods using primary imputation model and imputation model including cluster means

MI methods	RMSE	CR_{95}	ART (min)	PS	PS^t
Amelia	0.0391	0.9520	5	0.6688	0.6155
Amelia.gp	0.0390	0.9493	7	0.5890	0.5467
pan.100	0.0378	0.9465	15	0.7406	0.6960
pan.100.gp	0.0386	0.9447	76	0.8603	0.8980
mice.2l.pan	0.0378	0.9475	2475 (1.7 days)	0.6808	4.3689
mice.2l.pan.gp	0.0377	0.9499	5916 (4.1 days)	0.5326	9.4483

4.5 The Impact of Including Incomplete Auxiliary Variables

To examine the impact of including incomplete auxiliary variables with different missing rate, we include other indicators of HIV/AIDS knowledge and attitudes with different missing rate by steps. Building upon the primary model, we include, in sequence, the indicators with missing rate less than 60%, 70%, 80%, 90% and lastly include all the indicators. For each model, we impute the missing data 1000 times using Amelia and PAN with the same configurations as before. **Table 13** summarizes the additional incomplete indicators of HIV/AIDS knowledge and attitudes and **Table 14** summarizes the $RMSE$, CR_{95} and PS scores of models including indicators with different missing rate.

Table 13 additional HIV/AIDS knowledge and attitudes indicators with various missing rates

Indicators	Definition	Missing rate	Group
ever_tested	have ever been tested for hiv	0.052	< 60%
test_at_anc	tested for the AIDS virus as part of anc	0.328	< 60%
kw_pv_no_sex	knowing that one can reduce change of getting aids by not having sex at all	0.495	< 60%
att_ch_condom	believing that children should be taught condom to avoid aids	0.557	< 60%
kw_pv_any	knowing anything (else) a person can do to avoid or reduce the chances of getting aids	0.615	< 70%
test_at_delivery	tested for hiv between the time went for delivery but before the baby was born	0.635	< 70%
kw_someone_aids	Knowing someone who has or died of aids	0.677	< 70%
kw_mtct	knowing mother-to-child transmission	0.698	< 70%
kw_pv_condom_s	mentioning that using condom reduces hiv risk	0.708	< 80%
kw_pv_one_partner_s	mentioning that being faithful to one partner reduces hiv risk	0.714	< 80%
kw_pv_inj_s	mentioning that avoiding injection reduces hiv risk	0.724	< 80%
kw_pv_no_sex_s	mentioning that abstaining from sex reduces hiv risk	0.724	< 80%
kw_pv_fus_s	mentioning that avoiding tranfusion reduces hiv risk	0.729	< 80%
kw_pv_mp_s	mentioning that avoidng multiple parters/ having fewer partners reduces hiv risk	0.740	< 80%
kw_pv_pros_s	mentioning that avoiding prostitutes reduces hiv risk	0.745	< 80%
kw_pv_homo_s	mentioning that avoiding homosexuals reduces hiv risk	0.776	< 80%
kw_pv_sexmp_s	mentioning that avoiding sex with partner having many partners reduces hiv risk	0.781	< 80%
kw_pv_trheal_s	mentioning that seeking protection from a traditional healer reduces hiv risk	0.781	< 80%
kw_pv_kis_s	mentioning that avoiding kissing reduces hiv risk	0.786	< 80%
kw_pv_mos_s	mentioning that avoiding mosquito bites reduces hiv risk	0.786	< 80%
kw_pv_razor_s	mentioning that avoiding sharing razor reduces hiv risk	0.792	< 80%
kw_pv_sexinj_s	mentioning that avoiding sex with IDUs reduces hiv risk	0.797	< 80%
att_ashamed	agreeing that people with aids should be ashamed of themselves	0.865	< 90%

att_blamed	agreeing that people with aids should be blamed for bringing disease to the community	0.880	< 90%
att_denied_hserv	Knowing someone who has been denied health services b/c of aids in the past 12 months	0.880	< 90%
att_denied_social	Knowing someone who has been denied social event b/c of aids in the past 12 months	0.880	< 90%
att_verbal_abused	knowing someone who has been verbally abused b/c of aids in the past 12 months	0.880	< 90%
mis_aids_cured	Believing that aids can be cured	0.927	ALL
kw_trans_fus	mentioning that hiv can be transmitted by transfusion	0.932	ALL
kw_trans_inj	mentioning that hiv can be transmitted by injections	0.932	ALL
mis_mosquito_s	not mentioning mosquito as a way of HIV transmission	0.932	ALL
kw_trans_com	mentioning that hiv can be transmitted by sex without condom	0.938	ALL
kw_trans_sex	mentioning that hiv can be transmitted by sex	0.938	ALL
att_m_teacher	believing that a male teacher with hiv should be allowed to continue teaching in the school	0.943	ALL
att_allow_secret	allowing a person to keep it a secret if got infected with hiv	0.953	ALL
kw_trans_mp	mentioning that hiv can be transmitted by sex with multiple partners	0.958	ALL
mis_trans_kiss_s	not mentioning that hiv can be transmitted through kissing	0.958	ALL
kw_trans_razor	mentioning that hiv can be transmitted by contaminated razor or blade or other instruments	0.964	ALL
kw_trans_homo	mentioning that hiv can be transmitted by sex with homosexuals	0.969	ALL
kw_trans_pros	mentioning that hiv can be transmitted by sex with prostitutes	0.969	ALL
mis_food_s	not mentioning sharing food as a way of HIV transmission	0.969	ALL
mis_witchcraft_s	not mentioning witchcraft as a way of HIV transmission	0.969	ALL
kw_aids_fatal	Knowing that AIDS is a fatal disease	0.974	ALL
kw_mtct_s	mentioning that hiv can be transmitted from mother to child	0.979	ALL

From **Table 14** we can see that including more auxiliary variables, regardless of the missing rate of the variables, always decrease the $RMSE$.

Table 14 $\boldsymbol{RMSE}$, $\boldsymbol{CR_{95}}$ and $\boldsymbol{PS}$ scores of models including indicators with different missing rates

MI methods	RMSE	CR_{95}	ART (min)	PS	PS^t	No. of additional variables
Amelia	0.03915	0.95201	4.76	0.66876	0.61518	0
Amelia_60	0.03792	0.95309	6.53	0.71649	0.66124	4
Amelia_70	0.03754	0.95309	8.49	0.71114	0.65936	8
Amelia_80	0.03722	0.95382	13.88	0.75014	0.70298	22
Amelia_90	0.03720	0.95339	15.52	0.72387	0.68157	27
Amelia_100	0.03705	0.95194	69.91	0.63538	0.68355	44
pan.200	0.03776	0.94697	30.63	0.71038	0.69221	0
Pan.200_60	0.03694	0.94633	39.31	0.73723	0.72976	4
Pan.200_70	0.03627	0.94727	56.04	0.67161	0.69546	8
Pan.200_80	0.03581	0.94669	184.22	0.69973	0.91524	22
Pan.200_90	0.03558	0.94619	255.07	0.72689	1.04728	27
Pan.200_100	0.03533	0.94604	681.38	0.73204	1.69789	44

4.6 The Impact of Including Random Effects between Incomplete Variables

When imputing the country-level missing proportions of key indicators using a 2-level imputation model, it is likely that the effects between the incomplete indicators are random across clusters. For example, people's knowledge on drug to prevent mother-to-child transmission (MTCT) of HIV may have different effects on their knowledge on ways of MTCT transmission across different countries. However, including the random effects between incomplete variables is not straightforward for MIJM methods.[11] PAN and Amelia cannot account for random effects between incomplete variables whereas JOMO has such flexibility by allowing the variance-covariance matrix of the level-1 error to randomly vary across the clusters to mimic the underlying random effect.[11,16] However, the computation of JOMO is very complex and it takes very long time to run. For instance, the jomo.200 method takes over 9 days to finish

one-fold imputation. Moreover, the trace and ACF plots suggest that JOMO method is very hard to converge especially for complex imputation model. In our study, the convergence of JOMO method seems too impractical to be possible.

Compared with MIJM methods, MICE methods account for random effects between incomplete variables more easily because MICE methods model each incomplete variable separately. To examine the impact of including random effects between incomplete variables for MICE, we build upon *mice.2l.pan* and add random effects between the incomplete variables on the basis of the primary imputation model, i.e., *mice.2l.pan.re*. **Table 15** and **Table 16** compare the overall and indicator-specific $RMSE$ and CR_{95} for pan.200, jomo.200, mice.2l.pan and mice.2l.pan.re, respectively.

Based on **Table 15**and **Table 16**, including random effects between incomplete variables significantly reduces $RMSE$ but increases CR_{95} especially for *mice.2l.pan.re*. In addition, the results of PAN and JOMO methods are very similar to the results of *mice.2l.pan* and *mice.2l.pan.re* respectively.

Table 15 $\boldsymbol{RMSE}$, $\boldsymbol{CR_{95}}$ and $\boldsymbol{PS}$ scores of pan.200, jomo.200, mice.2l.pan and mice.2l.pan.re

MI methods	RMSE	CR_{95}	ART (min)	PS	PS^t
pan.200	0.0378	0.9470	31	0.7104	0.6928
jomo.200	0.0203	0.9550	13145	0.5870	20.4503
mice.2l.pan	0.0378	0.9475	2475	0.6808	4.3689
mice.2l.pan.re	0.0199	0.9796	4064	2.0572	8.0278

Table 16 indicators specific $\boldsymbol{RMSE}$ and $\boldsymbol{CR_{95}}$ of pan.200, jomo.200, mice.2l.pan and mice.2l.pan.re

Indicators	pan.200		jomo.200		mice.2l.pan		mice.2l.pan.re	
	RMSE	CR	RMSE	CR	RMSE	CR	RMSE	CR
heard_aids	0.034	0.953	0.017	0.954	0.034	0.954	0.017	0.983
kw_looking	0.059	0.930	0.031	0.961	0.059	0.938	0.030	0.978
kw_pv_condom	0.038	0.955	0.020	0.958	0.038	0.959	0.019	0.987

kw_pv_one_partner	0.037	0.952	0.019	0.960	0.037	0.953	0.018	0.969
kw_where_test	0.023	0.930	0.012	0.945	0.023	0.933	0.012	0.956
kw_mtct_bf	0.032	0.949	0.017	0.963	0.032	0.948	0.017	0.987
kw_mtct_preg	0.029	0.956	0.016	0.959	0.030	0.955	0.017	0.992
kw_mtct_delivery	0.027	0.950	0.014	0.954	0.027	0.951	0.014	0.987
mis_mosquito	0.046	0.949	0.029	0.966	0.046	0.949	0.025	0.980
att_vegetable	0.040	0.947	0.021	0.952	0.041	0.947	0.019	0.985
mis_food	0.033	0.948	0.018	0.953	0.033	0.941	0.019	0.981
att_f_teacher	0.036	0.944	0.018	0.944	0.036	0.945	0.018	0.969
mis_witchcraft	0.059	0.946	0.030	0.959	0.059	0.946	0.032	0.963
att_willing_care	0.027	0.950	0.015	0.955	0.026	0.957	0.013	0.988
att_secret	0.030	0.952	0.017	0.946	0.030	0.950	0.017	0.992
kw_mtct_drug	0.037	0.941	0.020	0.958	0.037	0.937	0.019	0.977
Overall	**0.038**	**0.947**	**0.020**	**0.955**	**0.038**	**0.947**	**0.020**	**0.980**

5. Discussion

The primary goal of the study is to evaluate the performance of different MI methods to impute missingness in TSCS data. Among all the methods, mice.2l.norm, mice.2l.pmm and mice.2l.lmer have $RMSE$ greater than 0.05 suggesting poor performance. The other 4 methods, namely Amelia, PAN, JOMO and mice.2l.pan, all have small $RMSE$ less than 0.05 and CR_{95} close to 95%. However, JOMO and PAN, which are full Bayesians JM methods, do not fully converge and thus produce unstable imputations for indicator *kw_mtct_drug*. Furthermore, although JOMO has the best PS which incorporates both $RMSE$ and CR_{95}, the ART of JOMO is too long for this method to be practical. On the other hand, Amelia and mice.2l.pan converge fast and produce stable imputations. In addition, both methods can be implemented parallelly which greatly reduces running time. Therefore, based on the results of this study, we think that Amelia and mice.2l.pan are the best for imputation of incomplete continuous variables in TSCS data.

Existing literatures show discrepancies among researchers on whether to include cluster means to improve imputations. Based on the results of this study, including cluster means makes little improvement to the imputations but does not hurt either. This finding provides empirical

evidence for Resche-Rigon and White's recent simulation study.[68] However, as Buuren[21] pointed out, the results may be dataset and/or model specific. For a different dataset or model, the cluster means may have huge effects on the missingness and thus on the imputations.

It is a consensus that including complete auxiliary variables improves the imputations.[85] However, incomplete auxiliary variables with high missing rate are usually not recommended to be included in the imputation model.[21] However, the results of our empirical study suggest that including more auxiliary variables is always beneficial regardless of the missing rate of the auxiliary variables. Of course, there is always a practical limit on how many auxiliary variables to include due to computational capacity and practical running time. This finding provides new insight on the inclusive strategy of auxiliary variables. However, since this is an empirical study, the finding can be dataset or model dependent. For instance, the auxiliary variables tested in this study are highly correlated with the targeted indicators. Therefore, future simulation evidence is still needed to fully understand the impact of different missing rates of auxiliary variables on the imputations.

Based on our results, two models which account for random effects between incomplete variables, namely JOMO and mice.2l.pan.re, have the smallest $RMSE$ among all methods. However, JOMO runs too long, hardly converges and thus produces unstable imputations, making this MI method less useful. For mice.2l.pan.re, although the method converges fast and produces reasonable imputations for all indicators, the CR_{95} of the imputed values are too high (~98%), especially for *kw_mtct_preg* and att_secret whose CR_{95} are close to 100%, suggesting large variance of the imputed values. We think if the primary interest is low bias of the mean of imputed values, mice.2l.pan.re may be preferred over mice.2l.pan.

Previous studies examining the performance of MI methods often use parameters of the analytical model as the evaluative targets.[12,85] However, our study, along with Mandel[79] and Ahmat Zainuri et al.'s study,[78] examines the performance of MI methods by assessing the accuracy of the imputed values. In other words, we examine the predictive rather than estimation accuracy of the imputation methods.[72] Another reason why we examine the predictive accuracy of imputations is that the imputed values are of interest and will be used to estimate the trends of the key indicators in a following study. To account for uncertainty of the imputations, we calculate the variance of the 1000 imputations for each missing value and will incorporate this variance in estimating trends of the key indicators. Therefore, although Rubin[74] and Buuren[21] think that the objective of MI is not to produce accurate imputations, we believe that the predictive accuracy of MI methods is justified and preferred in our study.

At IHME, simple regression method is often used to predict missing variable of a gender or of an age group using the observed indicator of the other gender or of the other age groups. The procedure is called cross-walking (CW). The results of our study suggest that MI, particularly, Amelia and mice.2l.pan, can be a much better choice than regression method for CW for three reasons. First, traditional CW often uses observed variable of one reference group to impute the same variable of other groups. MI, on the other hand, can utilize information of many variables from multiple groups to impute the missing variables. Second, traditional CW can only predict one missing variable for one specific group at a time. MI, however, can impute missingness in all indicators and of all gender or age groups in one shot, making the imputations more consistent and convincing. Third, compared with the traditional CW method, MI can naturally account for uncertainty of the imputations by imputing each missingness 1000 times. The variance of the 1000 imputations captures the uncertainty of imputations for the missingness. In addition, the

results of our study provide empirical evidence that MI methods can work well in imputing missingness in TCSC data. Therefore, we think that MI should be used and preferred in CW of TSCS data.

Similar to Castellacci et al.'s[5] and He et al.'s[24] studies, our study is an example of the "outside" application of MI, in which MI is used to produce multiply imputed datasets which can be used for many different analyses. A goal of this study is to make the multiply imputed datasets on HIV/AIDS knowledge and attitudes in the 47 SSA countries public available and researchers can use them to conduct different analyses. Therefore, the predictive accuracy of MI is highly preferred in this study.

Although carefully conducted, this study is not without limitations. First, according to Gelman et.al.[4] and Rendall et.al.,[92] missingness due to survey design is more likely to be MAR but the simulated missingness in this study are MCAR, which may affect the performance of MI. However, MI can also be used to handle MCAR and since all MI methods are evaluated using the same dataset, the performance of different MI methods is still comparable. Second, this study uses a real world dataset to evaluate the performance of MI methods. There may be uncontrolled and complex factors specific to this dataset that affect the performance of MI methods differentially. However, since all methods are evaluated using the same dataset and in the same way, we believe that our study provides useful empirical evidence on performance of different MI methods when imputing missingness in TSCS data. Lastly, since we do not extract all variables in the surveys, the multiply imputed datasets produced in this study can definitely be improved further by including more auxiliary variables from the surveys. However, based on the out-of-sample $RMSE$ and CR_{95}, we believe that our imputed datasets are good enough for a wide range of future analyses.

6. Conclusion

When imputing missingness in TSCS continuous data due to questions not asked in the survey, we find that Amelia and mice.2l.pan perform best among all the 7 multiple imputation methods. Both methods converge fast, produce reasonable and stable imputations and have small out-of-sample $RMSE$ less than 0.05 and CR_{95} very close to 95%. Amelia and MICE can also be implemented parallelly, which greatly reduces running time and makes the two methods more practical.

Based on the results of our study, including cluster means of variables in the imputation model has little impact on the imputations but significantly increases running time. However, including incomplete auxiliary variables that are correlated with targeted incomplete variables improves the imputation performance regardless of the missing rate of the auxiliary variables. In other words, even if the auxiliary variables have missing rate over 90%, including them in the imputation model still improves imputation of the targeted incomplete variables.

Regarding random effects between incomplete variables, JOMO and MICE are the only two methods that allow random effects between incomplete variables. However, JOMO converges poorly and runs slowly, which makes the method less useful in practice. MICE, on the other hand, works well. However, although allowing random effects between incomplete variables significantly reduces out-of-sample $RMSE$, it increases out-of-sample CR_{95} of the imputed values, suggesting larger variance/uncertainty of imputed values. Therefore, the usefulness of the method depends on whether the uncertainty of imputations is of primary concern.

Reference

1. IHME. About GBD. *Institute for Health Metrics and Evaluation* http://www.healthdata.org/gbd/about (2014).

2. Denk, M. & Weber, M. Avoid Filling Swiss Cheese with Whipped Cream : Imputation Techniques and Evaluation Procedures for Cross-Country Time Series. in (2011).

3. He, Y., Zaslavsky, A., Landrum, M., Harrington, D. & Catalano, P. Multiple imputation in a large-scale complex survey: a practical guide. *Stat. Methods Med. Res.* 19, 653–670 (2010).

4. Gelman, A., King, G. & Liu, C. Not Asked and Not Answered: Multiple Imputation for Multiple Surveys. *J. Am. Stat. Assoc.* 93, 846–857 (1999).

5. Castellacci, F. & Natera, J. M. A new panel dataset for cross-country analyses of national systems, growth and development (CANA). *Innov. Dev.* 1, 205–226 (2011).

6. The DHS Program - DHS Questionnaires. https://www.dhsprogram.com/What-We-Do/Survey-Types/DHS-Questionnaires.cfm.

7. Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. (John Wiley & Sons, 1987).

8. van Buuren, S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat. Methods Med. Res.* 16, 219–242 (2007).

9. Schafer, J. L. & Olsen, M. K. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivar. Behav. Res.* 33, 545–571 (1998).

10. Schafer, J. L. *Imputation of missing covariates under a multivariate linear mixed model*. https://cran.r-project.org/web/packages/pan/vignettes/pan-tr.pdf (1997).

11. Mistler, S. A. Multilevel Multiple Imputation: An Examination of Competing Methods. (Arizona State University, 2015).

12. Buuren, S. V., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M. & Rubin, D. B. Fully conditional specification in multivariate imputation. *J. Stat. Comput. Simul.* 76, 1049–1064 (2006).

13. Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. & Solenberger, P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.* 27, 85–96 (2001).

14. Rubin, D. B. Multiple Imputations in Sample Surveys--A Phenome- nological Bayesian Approach to Nonrespon. *Proc. Surv. Res. Methods Sect. Am. Stat. Assoc. 20-34* 9 (1978).

15. Schafer, J. L. & Yucel, R. M. Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values. *J. Comput. Graph. Stat.* 11, 437–457 (2002).

16. Yucel, R. M. Random-covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Stat. Model.* 11, 351–370 (2011).

17. He, Y., Yucel, R. & Raghunathan, T. E. A functional multiple imputation approach to incomplete longitudinal data. *Stat. Med.* 30, 1137–1156 (2011).

18. Schafer, J. L. & Zhao, M. J. Multiple Imputation for Multivariate Panel or Clustered Data. (2016).

19. Honaker, J. & King, G. What to do About Missing Values in Time Series Cross-Section Data. *Am. J. Polit. Sci.* 54, 561–581 (2010).

20. Dai, X. & Wang, H. Change in knowledge and attitude about HIV/AIDS in sub-Saharan Africa, 1990–2017: an analysis of national survey data. *Lancet Glob. Health* 7, S4 (2019).

21. Buuren, S. van. *Flexible Imputation of Missing Data, Second Edition*. (Chapman and Hall/CRC, 2018).

22. Little, R. J. A. & Rubin, D. B. *Statistical Analysis with Missing Data(Hardback) - 2002 Edition*. (John Wiley & Sons Inc, 2002).

23. Beck, N. & Katz, J. N. What to do (and not to do) with Time-Series Cross-Section Data. *Am. Polit. Sci. Rev.* 89, 634–647 (1995).

24. He, Y., Zaslavsky, A. M., Harrington, D. P., Catalano, P. & Landrum, M. B. Imputation in a Multiformat and Multiwave Survey of Cancer Care. 9 (2010).

25. Rubin, D. B. Inference and missing data. *Biometrika* 63, 581–592 (1976).

26. Kang, H. The prevention and handling of the missing data. *Korean J. Anesthesiol.* 64, 402–406 (2013).

27. *Ecological Statistics: Contemporary theory and application*. vol. Chapter 4: Missing data: mechanisms, methods, and messages (Oxford University Press, 2015).

28. King, G., Honaker, J., Joseph, A. & Scheve, K. Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *Am. Polit. Sci. Rev.* 95, 49–69 (2001).

29. Schafer, J. L. & Graham, J. W. Missing data: our view of the state of the art. *Psychol. Methods* 7, 147–177 (2002).

30. Little, R. J. A. Regression With Missing X's: A Review. *J. Am. Stat. Assoc.* 87, 1227–1237 (1992).

31. Teknomo, K. K Nearest Neighbors Tutorial: Strength and Weakness. https://people.revoledu.com/kardi/tutorial/KNN/Strength%20and%20Weakness.htm (2017).

32. Ebrahim, G. J. Missing Data in Clinical Studies Molenberghs G. and Kenward M. G. *J. Trop. Pediatr.* 53, 294–294 (2007).

33. Little, R. J. *et al.* The Prevention and Treatment of Missing Data in Clinical Trials. *N. Engl. J. Med.* 367, 1355–1360 (2012).

34. Dong, Y. & Peng, C.-Y. J. Principled missing data methods for researchers. *SpringerPlus* 2, (2013).

35. Azur, M. J., Stuart, E. A., Frangakis, C. & Leaf, P. J. Multiple Imputation by Chained Equations: What is it and how does it work? *Int. J. Methods Psychiatr. Res.* 20, 40–49 (2011).

36. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B Methodol.* 39, 1–38 (1977).

37. Graham, J. W. Missing data analysis: making it work in the real world. *Annu. Rev. Psychol.* 60, 549–576 (2009).

38. Barnard, J. & Rubin, D. B. Miscellanea. Small-sample degrees of freedom with multiple imputation. *Biometrika* 86, 948–955 (1999).

39. King, G., Honaker, J., Joseph, A. & Scheve, K. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *Am. Polit. Sci. Rev.* 95, 49–69 (2001).

40. Wulff, J. N. & Ejlskov, L. Multiple Imputation by Chained Equations in Praxis: Guidelines and Review. *Electron. J. Bus. Res. Methods* 15, (2017).

41. Newman, D. A. Missing Data: Five Practical Guidelines. *Organ. Res. Methods* 17, 372–411 (2014).

42. Schafer, J. L. *Analysis of Incomplete Multivariate Data*. (CRC Press, 1997).

43. Hughes, R. A. *et al.* Joint modelling rationale for chained equations. *BMC Med. Res. Methodol.* 14, 28 (2014).

44. van Buuren, S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat. Methods Med. Res.* 16, 219–242 (2007).

45. McLachlan, G. & Krishnan, T. *The EM Algorithm and Extensions*. (John Wiley & Sons, 2007).

46. Gelfand, A. E. & Smith, A. F. M. Sampling-Based Approaches to Calculating Marginal Densities. *J. Am. Stat. Assoc.* 85, 398 (1990).

47. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. *Bayesian Data Analysis*. (Chapman & Hall/CRC, 2003).

48. Wei, G. C. G. & Tanner, M. A. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *J. Am. Stat. Assoc.* 85, 699–704 (1990).

49. Efron, B. Missing Data, Imputation, and the Bootstrap. *J. Am. Stat. Assoc.* 89, 463–475 (1994).

50. van Buuren, S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat. Methods Med. Res.* 16, 219–242 (2007).

51. White, I. R., Royston, P. & Wood, A. M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* 30, 377–399 (2011).

52. Liu, J., Gelman, A., Hill, J., Su, Y.-S. & Kropko, J. On the stationary distribution of iterative imputations. *Biometrika* 101, 155–173 (2014).

53. Hardt, J., Herke, M. & Leonhart, R. Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC Med. Res. Methodol.* 12, 184 (2012).

54. Multiple Imputation in Stata. https://www.ssc.wisc.edu/sscc/pubs/stata_mi_models.htm (2018).

55. Murray, J. S. Multiple Imputation: A Review of Practical and Theoretical Findings. *ArXiv180104058 Stat* (2018).

56. Seaman, S. R., Bartlett, J. W. & White, I. R. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Med. Res. Methodol.* 12, 46 (2012).

57. Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O. & Hemingway, H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *Am. J. Epidemiol.* 179, 764–774 (2014).

58. Little, R. J. A. Missing-Data Adjustments in Large Surveys. *J. Bus. Econ. Stat.* 6, 287–296 (1988).

59. Honaker, J. & King, G. What to do about missing values in time-series cross-section data. *Am. J. Polit. Sci.* 54, 561–581 (2010).

60. Browne, W. J. & Draper, D. Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Comput. Stat.* 15, 391–420 (2000).

61. Goldstein, H., Carpenter, J., Kenward, M. G. & Levin, K. A. Multilevel models with multivariate mixed response types. *Stat. Model.* 9, 173–197 (2009).

62. Yucel, R. M. Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philos. Transact. A Math. Phys. Eng. Sci.* 366, 2389–2403 (2008).

63. Meng, X.-L. Multiple-Imputation Inferences with Uncongenial Sources of Input. *Stat. Sci.* 9, 538–558 (1994).

64. Carpenter, J. R. & Kenward, M. G. *Multiple imputation and its application*. (John Wiley & Sons, 2013).

65. Quartagno, M. & Carpenter, J. Package 'jomo': Multilevel Joint Modelling Multiple Imputation. (2019).

66. Honaker, J., King, G. & Blackwell, M. Amelia II: A Program for Missing Data. *J. Stat. Softw.* 45, 1–47 (2011).

67. Mistler, S. A. & Enders, C. K. A Comparison of Joint Model and Fully Conditional Specification Imputation for Multilevel Missing Data. *J. Educ. Behav. Stat.* 42, 432–466 (2017).

68. Resche-Rigon, M. & White, I. R. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Stat. Methods Med. Res.* 27, 1634–1649 (2018).

69. Schafer, J. L. Package 'pan': Multiple Imputation for Multivariate Panel or Clustered Data. (2018).

70. Honaker, J., King, G. & Blackwell, M. A Program for Missing Data: Package 'Amelia'. *J. Stat. Softw.* 45, (2018).

71. Buuren, S. van & Groothuis-Oudshoorn, K. MICE: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* 45, 1–67 (2011).

72. Chambers, R. Evaluation Criteria for Statistical Editing and Imputation. *EUREDIT Deliv. D33* (2000).

73. Barnard, J. J. & Meng, X. L. Applications of multiple imputation in medical studies: from AIDS to NHANES. *Stat. Methods Med. Res.* 8, 17–36 (1999).

74. Rubin, D. B. Multiple Imputation After 18+ Years. *J. Am. Stat. Assoc.* 91, 473–489 (1996).

75. Grund, S., Lüdtke, O. & Robitzsch, A. Multiple Imputation of Missing Data for Multilevel Models: Simulations and Recommendations. *Organ. Res. Methods* 21, 111–149 (2018).

76. Grund, S., Lüdtke, O. & Robitzsch, A. Multiple Imputation of Missing Data at Level 2: A Comparison of Fully Conditional and Joint Modeling in Multilevel Designs. *J. Educ. Behav. Stat.* 43, 316–353 (2018).

77. Schafer, J. L. *et al.* THE NHANES III MULTIPLE IMPUTATION PROJECT. 10 (1996).

78. Ahmat Zainuri, N., Jemain, A. A. & Muda, N. A Comparison of Various Imputation Methods for Missing Values in Air Quality Data. *Sains Malays.* 44, 449–456 (2015).

79. Mandel J, S. P. A Comparison of Six Methods for Missing Data Imputation. *J. Biom. Biostat.* 06, (2015).

80. Barber, R. M. *et al.* Healthcare Access and Quality Index based on mortality from causes amenable to personal health care in 195 countries and territories, 1990–2015: a novel analysis from the Global Burden of Disease Study 2015. *The Lancet* 390, 231–266 (2017).

81. Grund, S., Lüdtke, O. & Robitzsch, A. Multiple Imputation of Multilevel Missing Data: An Introduction to the R Package pan. *SAGE Open* 6, 2158244016668220 (2016).

82. Grund, S., Robitzsch, A. & Luedtke, O. Tools for Multiple Imputation in Multilevel Modeling: Package 'mitml'. (2019).

83. Jolani, S. Hierarchical imputation of systematically and sporadically missing data: An approximate Bayesian approach using chained equations. *Biom. J. Biom. Z.* 60, 333–351 (2018).

84. Buuren, S. van & Groothuis-Oudshoorn, K. Multivariate Imputation by Chained Equations: Package 'mice'. (2019).

85. Collins, L. M., Schafer, J. L. & Kam, C. M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol. Methods* 6, 330–351 (2001).

86. Madley-Dowd, P., Hughes, R., Tilling, K. & Heron, J. The proportion of missing data should not be used to guide decisions on multiple imputation. *J. Clin. Epidemiol.* 110, 63–73 (2019).

87. Honaker, J., King, G. & Blackwell, M. Amelia II: A Program for Missing Data. *J. Stat. Softw.* 45, (2012).

88. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. in *Ijcai* vol. 14 1137–1145 (Montreal, Canada, 1995).

89. Gelman, A. & Rubin, D. B. Inference from Iterative Simulation Using Multiple Sequences. *Stat. Sci.* 7, 457–472 (1992).

90. Geyer, C. J. Practical Markov Chain Monte Carlo. *Stat. Sci.* 7, 473–483 (1992).

91. White, I. R., Royston, P. & Wood, A. M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* 30, 377–399 (2011).

92. Rendall, M. S., Ghosh-Dastidar, B., Weden, M. M., Baker, E. H. & Nazarov, Z. Multiple Imputation For Combined-Survey Estimation With Incomplete Regressors In One But Not Both Surveys. *Sociol. Methods Res.* 42, (2013).