University of Nairobi

College of Health Sciences

Master Project in Medical Statistics

MODEL-BASED FULLY CONDITIONAL SPECIFICATION AND PREDICTIVE MEAN MATCHING: APPLICATION TO HIV RISK FACTORS AMONG FEMALE SEX WORKERS IN KENYA

Research Report in Medical Statistics, W62/8248/2017, 2019

Kahindi Grace Kadzo

December 2019

Master Project in Medical Statistics **University of Nairobi**

December 2019

MODEL-BASED FULLY CONDITIONAL SPECIFICATION AND PREDICTIVE MEAN MATCHING: APPLICATION TO HIV RISK FACTORS AMONG FEMALE SEX WORKERS IN KENYA

Research Report in Medical Statistics, W62/8248/2017, 2019

Kahindi Grace Kadzo

College of Health Sciences
Institute of Tropical and Infectious Diseases
Old Mbagathi Road, off Ngong' Road
19676-00202 Nairobi, Kenya

Master Thesis

Submitted to the College of Health Sciences in partial fulfilment for a degree in Master of Science in Medical Statistics

Submitted to: The Graduate School, University of Nairobi, Kenya

ABSTRACT

Background: HIV disproportionately affects sex workers. It is important to continually evaluate sex work, given it's fluid and dynamic nature. Missing data is a common complication to HIV research, especially where accurate and complete collection of data is a challenge.

Aim: To study the missing data problem in the female sex workers' data and employ the multiple imputation technique.

Methods: Multiple imputation using the Fully Conditional Specification (FCS) was used to handle the missing data problem. For the target analysis, a binary logistic model was used to test association between HIV status and risk factors among female sex workers. We assessed the impact of missing data on the statistical significance of the risk factors of HIV. We further, compared the performance of model-based FCS and Predictive Mean Matching (PMM) by assessing distributional properties, convergence, adjusted odds ratios, interval width and relative efficiency.

Results: There were generally low proportions of missingness and missing data was not found to affect statistical significance of associations of HIV risk factors to HIV positivity of female sex workers. There was a reverse in the interpretation of results in the number of sex acts per week, though not statistically significant. Multiple imputation reduced standard errors of parameter estimates, giving more precise estimates and narrower confidence intervals. Distributional properties were also preserved by MI. Model-based FCS performed slightly better in convergence, interval width while PMM had better relative efficiency.

Conclusion: Multiple imputation results in more reliable estimates with lower standard errors. Performance of the model-based FCS was considerably better than PMM. These results are, however, not considered conclusive and may need validation using a large simulation study.

Master Thesis in Medical Statistics at the University of Nairobi, Kenya.
Research Report in Medical Statistics

DECLARATION AND APPROVAL

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.

Signature Date

Kahindi Grace Kadzo
Reg No. W62/8248/2017

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.

Signature Date

Dr Rachel Sarguta
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: rsarguta@uonbi.ac.ke

This project is dedicated to my family, for their continued support and constant encouragement.

ACKNOWLEDGEMENTS

First, I thank the University of Manitoba-University of Nairobi for availing the SWOP data and supporting me in the entire process. They made this project possible.

I'd also like to immensely appreciate my supervisor for her unwavering support and guidance throughout this project.

I wish to thank my classmates and family for their constant encouragement.

Finally, I thank God for making this happen.

Kahindi Grace Kadzo

Nairobi, 2019.

Contents

FIGURES AND TABLES

Figures

Tables

LIST OF ABBREVIATIONS

Abbreviations	Description of abbreviations
ACA	Available Case Analysis
AIDS	Acquired Immunodeficiency Syndrome
CCA	Complete Case Analysis
CI	Confidence Interval
FCS	Fully Conditional Specification
FSW	Female Sex Worker
HIV	Human Immunodeficiency Virus
IBBS	Integrated Behavior and Biological Survey
KP	Key Population
MAR	Missing At Random
MARP	Most At Risk Population
MCAR	Missing Completely At Random
MI	Multiple Imputation
MICE	Multivariate Imputation by Chained Equations
MNAR	Missing Not At Random
OR	Odds Ratio
PMM	Predictive Mean Matching
PROM	Patient Reported Outcome Measures
SD	Standard Deviation
SE	Standard Error
STI	Sexually Transmitted Infection
SW	Sex Worker
SWOP	Sex Worker Outreach Program
UNAIDS	Joint United Nations Programme on HIV/AIDS
WHO	World Health Organization

OPERATIONAL DEFINITIONS

Terms	Definition of Terms
Sex work	The negotiated exchange of money or other items of value for sexual services
Female Sex Worker	Woman who exchanges sex for money or other valuable items with men
Risk	The probability of a person getting HIV/STI.
Nonresponse	The failure to obtain a measurement on one or more study variables for one or more subjects in the study.

1 INTRODUCTION

This chapter gives an overview of the burden of HIV globally, regionally and locally; and HIV in sex workers.

1.1 Background

Human Immunodeficiency Virus (HIV) is a type of Sexually Transmitted Infection (STI), transmitted through contact with infected blood. It is a virus that attacks the CD4 cells, weakening the body's immune system.

1.1.1 HIV

According to World Health Organization (WHO), HIV is a major public health concern, globally. There were about 770,000 deaths from HIV-related causes in 2018 and approximately 37.9 million people living with HIV; with 1.7 million new infections in the same year (UNAIDS, 2019).

In 2016, the Global Health Estimates found HIV to be one of the leading causes of death in low-income countries. (WHO, 2018).

In 2018, Kenya had 1.6 million people living with HIV (AVERT, 2019). In the same year, Kenya had 46,000 new HIV infections and 25,000 AIDS-related deaths (AVERT, 2019). HIV prevalence in adults is estimated to have dropped from 10% in late 1990s to approximately 4.7% in 2018.

Table 1. HIV Global Estimates. Source: (UNAIDS, 2019)

WHO region	Estimated number of people living with HIV (2018) - all ages	Number of deaths due to HIV/AIDS (2018)
Africa	25 700 000 [22 200 000 - 29 500 000]	470 000 [340 000 - 630 000]
America	3 500 000 [3 000 000 - 4 200 000]	49 000 [36 000 - 65 000]
South-East Asia	3 800 000 [3 100 000 - 4 900 000]	150 000 [110 000 - 240 000]
Europe	2 500 000 [2 300 000 - 2 800 000]	38 000 [28 000 - 49 000]
Eastern Mediterranean	400 000 [290 000 - 570 000]	15 000 [10 000 - 23 000]
Western Pacific	1 900 000 [1 700 000 - 2 100 000]	48 000 [40 000 - 61 000]
(WHO) Global	37 900 000 [32 700 000 - 44 000 000]	770 000 [570 000 - 1 100 000]

Certain populations have a greater risk of contracting HIV due to some existing dynamics and drivers. These populations are also known as Most At Risk Populations (MARPs) or Key Populations (KPs). They include injecting drugs users, men who have sex with men, people in prison and/or other closed settings, transgender people and sex workers (SWs) and their clients. KPs and their clients play a big role in the prevalence of HIV. In 2018, they accounted for about 54% of all new infections, globally, and around 95% of new cases of HIV in central Asia, eastern Europe and North and Middle East Africa (UNAIDS, 2019). Legal and social barriers that KPs often face increase their vulnerability to HIV and impede their access to prevention services as well as testing and treatment programmes.

Although there's a generalized epidemic of HIV in Kenya, in 2017, an estimated 47% of new infections occurred among KPs and their partners.

Complex sexual dynamics exist between the general population and MARPs. MARPs sexually mix with the general population and also with each other. The general population also sexually mixes with each other. As a result, the number of sex partners amongst minorities and consequently HIV risk is disproportionate (Muraguri, 2010). This affects the transmission dynamic as demonstrated in Figure 1.

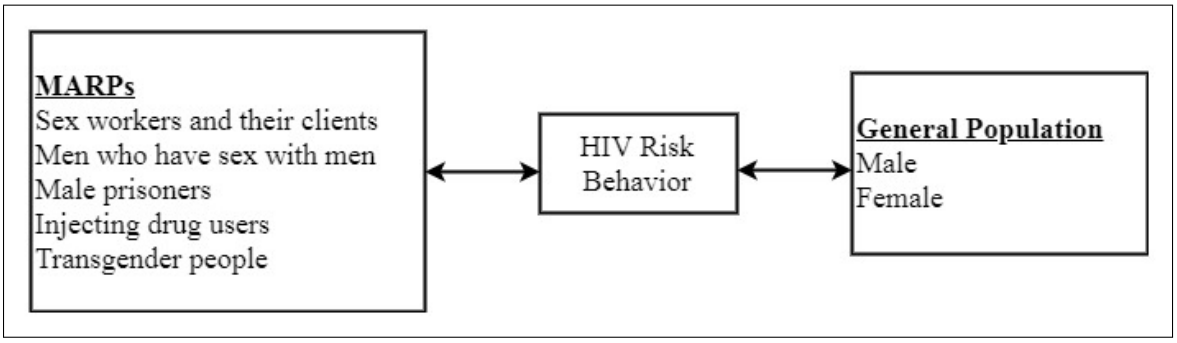


Figure 1. How high-risk populations relate with the general population. Source: (Muraguri, 2010)

1.1.2 Sex Workers

Sexual contact is the leading mode of transmission of HIV (more than 75% of new infections). Presence of other STIs and high rates of sexual partner change increase the efficiency and rate of sexual HIV transmission. UNAIDS (2019) estimates 21 times higher risk of acquiring HIV among sex workers. Sex workers also have the highest prevalence among key populations in Kenya, with approximately 29.3% infected (AVERT, 2019).

Sex workers are constantly at risk of HIV infection, especially where there's low condom use and health care services are not readily available. This can lead to high prevalence of HIV, up to (60-90%). Currently, an estimated 16% of sex workers have access to HIV prevention services. Effective interventions are needed to help lower transmission of HIV between clients, sex workers, regular partners and the general population.

1.1.3 Missing Data

Data collection and analysis forms the backbone of all empirical research (Raghunathan, 2015). All analyses aim to obtain unbiased estimates of population parameters (Schafer and Graham, 2002). However, this may not be possible especially when there's nonresponse, and as a result, the study has missing data. Missing data is inevitable in any research

Table 2. Sex worker numbers, Source: (AIDSinfo–UNAIDS, 2019)

Sex workers		
Population size estimate (#)	167 900	Region: 34 counties in Kenya; Method: Programmatic mapping
HIV prevalence (%)	29.3	Source: IBBS study conducted in Nairobi - 2011.
Knowledge of HIV status (%)	95.5	Source: Polling booth survey 2017
Antiretroviral therapy coverage (%)	73	Source: Polling booth survey 2017
Condom use (%)	92	Source: Polling booth survey 2017
Condoms distributed through prevention programmes (#)	459	Source: UNAIDS Special Analysis and Global AIDS Monitoring, 2019

(Pedersen et al., 2017), especially in longitudinal studies which tend to take long so attrition can be high.

Nonresponse can either be at unit level, item level or on a wave: Unit nonresponse, occurs when all data for a given unit is missing; Item nonresponse, occurs when partial data for a given unit is available; wave nonresponse occurs in longitudinal studies where attrition happens and therefore participants are present for data collection in some waves and missing for others (Schafer and Graham, 2002).

1.2 Statement of the Problem

HIV/AIDS disproportionately affects key and priority populations in Kenya as compared to the general population. This makes provision of HIV prevention services to these groups critical in the overall fight against HIV/AIDS.

Sex work is fluid and dynamic, changing constantly. For this reason, it is advisable to continually evaluate sex work and HIV in any given area to understand and respond to its changing nature (Muraguri, 2010).

Evaluation of the sex work becomes a challenge when there is missingness on analysis variables. Most studies done previously have taken the Complete Case Analysis (CCA), Available Case Analysis (ACA) or single imputation approaches. These methods often result in potentially biased estimates and reduced precision from overestimated or underestimated standard errors. They can also result in loss of distributional relationships between variables and more generally, loss of statistical information (Chinomona and Mwambi, 2015).

One flexible method of estimating missing data is Multiple Imputation (MI). It preserves uncertainty of missing values by creating several multiply imputed datasets and then pools results from each. This makes MI attractive. Unbiased estimates and standard errors from MI results in valid conclusions (Carpenter and Kenward, 2012). This research aims to highlight the impact of missing data in determining the association of risk factors of HIV among female sex workers. We also aim to illustrate the strength of MI as a method of handling missing data.

Unlike joint modeling approaches of MI, Fully Conditional Specification (FCS) does not assume multivariate normality of the variables, which is not plausible for categorical variables. Instead, FCS imputes on a variable by variable basis. A lot of comparisons have been made on multivariate normal imputation with FCS (Lee and Carlin, 2010), but very little on the FCS methods. In particular, very little has been done with regard to comparison on the performance of model-based FCS versus Predictive Mean Matching (PMM). In this study we highlight how the FCS methods work, as well as compare their performance.

1.3 Justification of Study

High-income countries have efficient systems set up to collect information on causes of death, like HIV. On the other hand, many low- and middle-income countries lack such infrastructure, and the numbers of deaths from HIV or numbers of people living with HIV have to be estimated from incomplete data (WHO, 2018).

In any research, missing data complicates interpretation of results and even small proportions of missingness can cause bias and inefficiency (Harel et al., 2012). Missing data are a major problem in HIV research, especially in sub-Saharan Africa where accurate and complete collection of data is a challenge (Chinomona and Mwambi, 2015). Efficient handling of missing data can help increase the validity of conclusions drawn. This study aims to illustrate the multiple imputation technique -using the fully conditional specification- for handling missing data in both covariates and the response variable to obtain unbiased estimates of HIV using risk behavioral factors and demographic variables among female sex workers.

1.4 Objectives of the Study

1.4.1 Overall Objective

To assess the impact of missing data on the association between HIV positivity and its covariates using female sex workers' data in Nairobi, Kenya.

1.4.2 Specific Objectives

1. To investigate the extent and patterns of missing data for risk factors of HIV and HIV positivity in the female sex worker data.

2. To investigate the missing data mechanism for risk factors of HIV and HIV positivity in the female sex worker data.

3. To assess the impact of missing data on the association between HIV positivity and its covariates.

4. To compare the performance of model-based FCS and PMM.

1.5 Scope

This study utilizes female sex worker data from Sex Worker Outreach Program (SWOP) of the University of Manitoba-University of Nairobi. Only female sex workers enrolled between June 2014 and June 2018 are considered. The geographical spread of the participants only covered those enrolled in the SWOP city center clinic of Nairobi, Kenya.

There's a wide range of risk factors of HIV among female sex workers. For this study, one demographic and three risk behavioral factors were considered.

2 LITERATURE REVIEW

2.1 Introduction

This chapter explores the existing relevant work that has been done by different scholars, researchers and authors. The review includes both international and local literature. The first two sections review risk factors of HIV among sex workers mentioning the methodologies used as well as highlighting how missing data was handled. The last section, reviews different missing data handling techniques, including MI, that have been used in past studies.

2.2 Risk factors of HIV Among Sex Workers

Muraguri (2010) and Israel et al. (2008) list the following as the factors increasing the risk of sex workers to HIV:

- Lack of HIV prevention
- Frequency of partner change
- Multiple sex partners
- High-risk (unprotected) sex
- Dry sex, douching/drying practices
- Higher levels of symptomatic or untreated STIs

- Drug and alcohol related-HIV risk behaviors - unprotected sex, sharing of drug injection equipment
- Limited access to health services
- Stigma and marginalization

Coetzee et al. (2017) examined the factors associated with HIV in FSW in South Soweto, South Africa, using a logistic model and chi-squared test of association. Migrancy, multiple clients, advancing age and incomplete secondary schooling, were found to increase the chances of acquiring HIV.

Heavy episodic drinking (taking ≥ 5 alcoholic drinks on a single occasion) among female sex workers was associated with higher risk of condom breaks, STIs, sexual violence and higher number of sex partners, in a study by Chersich et al. (2007).

In rural western Kenya, Amornkul et al. (2009) did a community-based cross-sectional survey to assess the prevalence of HIV and it's associated risk factors among young adults. Prevalence of HIV was found to be highest in females at 20.5%, while males had 10.2%. The researchers also found strong associations between HIV and widowhood, higher number of sex partners, age and Herpes Simplex Virus 2 (HSV-2) seropositivity.

2.3 Missing Data in Sex Workers' Studies

Campeau et al. (2018) examined and compared the risk factors of HIV positivity among the sexually active participants who injected drugs and practised sex work versus those who injected drugs but did not practice sex work. The researcher used a bio-behavioral survey in East Central Canada between 2004 and 2016. The study showed that HIV prevalence at baseline was higher among sex workers compared to non-sex workers, which agrees with

most literature. Results further showed that importance of sex work for HIV infection varied by gender with higher prevalence in women. Incarceration was associated with HIV among women, but interestingly, sexual risk behaviors were not positively associated with HIV among women. Missing data was present but was excluded from analysis.

Bui et al. (2001) did a cross-sectional study in urban, rural and minority residents in Viet nam, to estimate the proportion of people living there who engaged in HIV-related risk behavior, and to find out their level of knowledge on HIV/AIDS. Results showed low prevalence of premarital intercourse 4-7% for married women and (9-16%) for married men; 6-16% of single men had ever had sex and below 3% admitted to having ever had sex with a sex worker. Urban and rural dwellers were more knowledgeable on HIV/AIDS compared to the mountainous dwellers. Males in age group 20–29 years were associated with having multiple sex partners. There was, however, 89% response rate but the researchers did not mention how missing data was handled.

Beksinska et al. (2018) conducted a biological and behavioral assessment survey on female sex workers in Karnataka, south India to examine whether HIV/STI risk differs by perpetrator (workplace, community and domestic) of violence. Weighted, bivariate and multivariate analyses were used to determine associations between HIV/STI risk and violence by perpetrator. The researchers found that the risk of getting HIV/STI differed by the perpetrator and was highest among FSWs who experienced violence in the workplace/community and at home. The researchers used non-response weighting to handle missing data, the limitations of which are discussed in the next chapter.

2.4 Missing Data Approaches

For a long time CCA has been used, but a traditional transition has seen unit nonresponse handled by re-weighting, while item nonresponse handled by single imputation (Schafer and Graham, 2002). Presently, more efficient methods are preferred.

Performing both MI and CCA is informative as it bridges the gap between current and future practice, and is also reassuring if the results are similar (Welch, 2015). In case they give different results, it's best to understand why. This may require careful assessment of the missing data mechanisms (particularly if missing conditional on a covariate) to establish whether CCA or MI are valid. If plausible that both are valid, MI is preferred due to its greater efficiency. Welch (2015) further cautions that even though MI is robust to departures from assumptions, it may not be always guaranteed, and when covariates have substantial missingness, reporting should be appropriately cautious.

Audigier et al. (2018) studied and compared MI methods for binary and continuous data where variables are systematically (not measured or not defined consistently across clusters) and sporadically (missing data is specific to individual observation) missing. The comparisons showed that relative performances of MI methods may vary depending on the pattern of missing data, type of variable and the multilevel structure. The study further found that valid inferences are obtained when the dataset contains many clusters. The researchers further highlighted that heteroscedastic MI methods gave more accurate inferences compared to homoscedastic methods, which should when data contains few individuals per cluster.

Rombach et al. (2018) compared imputation of Patient Reported Outcome Measures (PROMs) at item level, subscale level and composite score level using PMM and Ordi-

nal logit methods of the Multivariate Imputation by Chained Equations (MICE) and also using CCA. Results showed that performance improved with lower proportions of missingness and increased sample size. For smaller samples ($n \leq 200$), MI at composite score and subscale levels did better than at item level, although, at high proportions of item nonresponse, imputation was found to be more accurate. For large sample sizes (≥ 500), both methods gave similar results. The study concluded that the choice of an imputation model should be guided by the proportion of missing data, number of PROM items, number of levels within individual items and sample size.

Nuwasiima (2018) used random survival forests to select highly predictive covariates of under-five child survival from the Demographic Health Survey data. Random Forests and MICE were also used to impute missing covariate data and analysis was done using random survival forests and Cox regression. Results showed that missingness in covariates was more related to the time-to-event than the event status. It was further shown that MI led to increase in variable importance scores. Random forest showed potential of producing more accurate estimates with high levels of missingness than MICE. The researcher recommended that validation of the study results may be necessary using a larger simulation study and other nonresponse models.

De Silva et al. (2019) did a simulation study to assess how varying proportions of missingness in maternal smoking associated with childhood obesity, using data from the longitudinal study of Australian children. Performance of several MI methods was assessed: FCS (PMM, ordinal and multinomial), two-fold FCS and multivariate normal imputation. These methods were compared under a restricted version - taking account of the restrictions of smoking status over time - and a standard version - without restrictions. Results showed that PMM generally performed best and there was reduced bias in all methods when

accounting for restrictions. The researcher recommended further research to explore implementation of restrictions within multilevel imputation methods.

3 METHODOLOGY

3.1 Introduction

This chapter gives an overview of missing data, multiple imputation, how data was obtained and handled. It also gives details of the study design, data management and the analysis process used.

3.2 Missing Data

3.2.1 Consequences of Missing Data

Missing data may cause the following major problems:

1. Loss of efficiency and information.
2. Biased inference when subset of observed data is not representative of the population.
3. Analysis needs an assumption about why data are missing - estimates are potentially biased if we make the wrong assumption.

It is of great importance to understand the pattern and mechanism of missing data. This ultimately determines how to deal with the missingness.

3.2.2 Patterns of Missing Data

A pattern of missing data shows the location of the missing values in a potential complete (100% response rate) data matrix (Raghunathan, 2015). The choice of imputation method must take into account the missing data pattern for valid imputations and efficient statistical inferences to be made.

Monotone Missingness Pattern

Monotone missingness occurs when for a missing variable Y_k, then all Y_j variables ($j > k$) are also missing (Buuren, 2012). This is mostly seen in studies prone to attrition, like longitudinal studies.

Univariate Pattern

Univariate pattern is a situation where data is missing on only one variable.

Arbitrary Pattern

No particular pattern is seen in the missingness of data.

3.2.3 Missing Data Mechanisms

This is the process that governs the probability of a data-point being missing. Essentially, the missingness mechanism concerns the relationship existing between values and missingness in the data matrix (Little and Rubin, 2019).

Let:

- $\boldsymbol{Y}_i = (Y_{i,1}, Y_{i,2}, \cdots, Y_{i,p})^T$ for p ($j = 1, 2, \cdots, p$) variables from i units: $i = 1, 2, \cdots, n$.
- For each unit:
  - $\boldsymbol{Y}_{i,O}$ denotes a subset of p observed variables
  - $\boldsymbol{Y}_{i,M}$ denotes a subset of the unobserved (missing)
- For observed $Y_{i,j}$ then $R_{i,j} = 1$, otherwise $R_{i,j} = 0$.

$Pr(\boldsymbol{R}_i|\boldsymbol{Y}_i)$ therefore defines the missing data mechanism.

Missing Completely At Random (MCAR)

MCAR occurs when the probability of missingness is neither related to the observed nor the unobserved.

$$Pr(\boldsymbol{R}_i|\boldsymbol{Y}_i) = Pr(\boldsymbol{R}_i)$$

In this case the observed data is representative of the population (Carpenter and Kenward, 2012). MCAR implies that the distribution of the variable is the same for both groups.

Complexities arising when data is missing may then be ignored, apart from the obvious loss of information and statistical power. A reduced sample size in MCAR causes enlarged standard errors, but does not cause bias (Jakobsen et al., 2017). In practical instances, a strong MCAR assumption is rarely tenable (Buuren, 2012; Chinomona and Mwambi, 2015).

Missing At Random (MAR)

MAR occurs when the probability of missingness is related to the observed but is independent of the unobserved.

$$Pr(\boldsymbol{R}_i|\boldsymbol{Y}_i) = Pr(\boldsymbol{R}_i|\boldsymbol{Y}_{i,O})$$

MAR can be viewed as a much broader class than MCAR.

Missing Not At Random (MNAR)

MNAR occurs when both MCAR and MAR are not true. The missingness depends on the missing data, even given the observed data (Jakobsen et al., 2017; Sterne et al., 2009).

$$Pr(\boldsymbol{R}_i|\boldsymbol{Y}_i) \neq Pr(\boldsymbol{R}_i|\boldsymbol{Y}_{i,O})$$

In MNAR, the probability of being missing depends on an underlying value, which may not be known by the researcher. It is the most complex case and it's statistically impossible to take into account its potential bias with certainty (Sterne et al., 2009). It's advisable to examine the cause(s) of the missingness, and/or do what-if analyses to determine the sensitivity of the results in different scenarios.

3.2.4 Dealing with Missingness

A number of approaches are used in dealing with missing data. These include but are not limited to:

Complete Case Analysis (CCA)

CCA is a common approach that restricts the analysis to subjects/cases without missing values in the analysis variables. Though a convenient approach, it leads to loss of information (Yuan, 2010). Parameter estimates can be biased if the cases included (observed) and excluded (missing) from the analysis differ systematically (Yuan, 2010; Welch, 2015). Even if the included subjects are a random subset of the sampled subjects, the sampling error increases due to the reduced sample size.

Sterne et al. (2009) gives the following precautionary measures when opting for CCA:

1. If the proportion of missingness is below 5%, the potential impact of missing data is considered negligible.
2. If missingness is only in the dependent variable and there are no auxiliary variables.
3. If missingness mechanism is MCAR.

Available Case Analysis (ACA)

Just like CCA this is a simple approach. The mean of a variable is only based on the observed data of that variable. For covariance, data is used where both variables of interest are non-missing. Hence, different analyses are based on different subsets of the data leading to inconsistency with each other. Just like CCA, summaries are biased if missing data and observed data differ systematically.

ACA can also arise when a variable(s) is excluded from the analysis due to its high proportions of missing data (complete-variables analysis). However, this can lead to omitting variable(s) that are important in satisfying assumptions of desired causal interpretations.

Nonresponse Weighting

For every variable with missing data, a model can be built to predict the missingness in that variable using information from the rest of the variables. Survey weights could then be obtained by getting the inverse of predicted probabilities of response from the model. The complete-case sample is thus made representative of the full sample. However, the more the variables with missing data the more complicated this method becomes. Moreover, like any weighting criteria, if predicted probabilities approach 0 or 1, standard errors could potentially become erratic.

Full Information Maximization Likelihood (FIML)

Maximum likelihood estimation principally utilizes a joint distribution to estimate parameters of an outcome and its associated covariates that, if true, maximizes the probability of observing values that were actually observed (Jakobsen et al., 2017). FIML does not impute data, instead, it estimates a likelihood function for each observation on present variables so that all available data is used. FIML is an efficient approach which leads to unbiased estimates provided the data is MAR/MCAR.

Imputation

Missing values can be filled-in rather than getting removed. The imputation approach fills in a plausible set of values for the missing set which can be done through taking draws from the posterior distribution of the observed values. As a result, we preserve the sample size, which is good for precision and bias. Careful selection of the imputation method is necessary to avoid other kinds of bias.

Single Imputation Single imputation means a single draw is taken. The commonly used methods are: mean or median for continuous variables and mode for categorical variables. Common approaches in longitudinal data are Worst Observation Carried Forward (WOCF), Best Observation Carried Forward (BOCF) and Last Observation Carried Forward (LOCF).

Single imputation methods can yield underestimates of the standard deviation and consequently severely distort the distribution of the given variable. These methods, thus, often cause potentially biased estimates, loss of distributional relationships between variables and loss of statistical information (Chinomona and Mwambi, 2015).

Imputed values are treated no differently from known values in analyses. They don't reflect the uncertainty that comes from prediction. For any imputation we should appreciate that we are substantially uncertain of the unobserved values.

Multiple Imputation (MI) Multiple imputation is increasingly becoming the approach of choice for handling missing data. It involves specifying an imputation model, under which multiple values are drawn from their posterior predictive distribution given the observed values (Little and Rubin, 2019; Audigier et al., 2018). The goal of MI is not to

reproduce the data but to obtain valid inferences from partially observed data (Vink, 2016), and also reflect the uncertainty due to missing values (Yuan, 2010). The quality of the imputation method should thus be evaluated with respect to this goal (Buuren, 2012).

For a data matrix $\boldsymbol{Y}$: let $\boldsymbol{Y}_O$ denote observed data and $\boldsymbol{Y}_M$ missing data. The MI procedure involves three steps (Carpenter and Kenward, 2012):

1. Impute each missing value K times from the distribution $f(\boldsymbol{Y}_M|\boldsymbol{Y}_O)$, creating K multiply imputed datasets.
2. Analyze each dataset separately using appropriate methods. This yields K parameter estimates, $\boldsymbol{\beta}_k$, with $Var(\boldsymbol{\beta}_k)$ associated variance estimates.
3. Combine results from the K analyses, using Rubin's rules, to get one final result.

Rubin's rules for inference are as given below.

For a single imputed variable, the estimate β from MI is the mean of the estimates from K imputations:

$$\hat{\beta}_{MI} = \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k, \tag{1}$$

with a corresponding overall variance given by:

$$\hat{V}_{MI} = \hat{W} + \left(1 + \frac{1}{K}\right)\hat{B}, \tag{2}$$

where $\hat{W} = \frac{1}{K}\sum_{k=1}^{K}\hat{\sigma}_k^2$ is the within imputation variance and $\hat{B} = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{\beta}_k - \hat{\beta}_{MI})^2$ is the between imputation variance.

The hypothesis $H_0 : \beta = \beta^0$ is then tested by the t statistic

$$T = \frac{\hat{\beta}_{MI} - \beta^0}{\sqrt{\hat{V}_{MI}}}, \tag{3}$$

with v degrees of freedom, where $v = (K-1)\left(1 + \frac{\hat{W}}{(1+\frac{1}{K})\hat{B}}\right)^2$

Choosing the Imputation Model Choosing the correct model can be the most challenging part of multiple imputation. For proper imputations, and consequently valid statistical inferences, it is crucial that the imputation model (Molenberghs et al., 2014):

- Takes the cause of missingness into account
- Preserves the relations in the data and their uncertainty

The greatest scope for error in using multiple imputation is inappropriate specification of the imputation model (Carpenter and Kenward, 2012). It is important to include all variables of the target analysis to preserve the relationships amongst variables (Welch, 2015).

MI only utilizes information from the observed data. More precise estimates can be obtained by including auxiliary variables. These are variables which can predict the missing values and/or their probability of being unobserved. Therefore, if computationally feasible, it is advisable to include as many auxiliary variables as possible (Welch, 2015). However, this should be done with caution to avoid over-fitting.

For continuous variables, auxiliary variables which have a strong correlation with analysis variables ($r \approx 0.7$–0.9) can reduce bias and standard errors (Johnson and Young, 2011).

Categorical auxiliary variables can be informed by literature and/or good knowledge of the data.

When to Impute If the bigger proportion of observations in a covariate is missing one can choose to drop the variable from analysis. However, this modifies the substantive analysis and a different scientific question is being answered in the new analysis (Raghunathan et al., 2018). If the variable is important for analysis then it's better to impute the missing data using some auxiliary predictors.

Planning for adjustment for missing data in a study should be done at the design stage and not as post-hoc thinking. As such, if a variable is expected to have substantial missingness, then it's better to collect correlates of that variable which have good predictive power for the missing data (Raghunathan et al., 2018).

MI Methods The choice of MI method is dependent on the missingness pattern, missingness mechanism and type of variable. Table 3 summarizes this.

Table 3. MI Methods Available in SAS (SAS, 2017)

Missingness Pattern	Type of Imputed Variable	Type of Covariates	Available Methods in SAS
Monotone	Continuous	Arbitrary	Monotone regression
			Monotone predicted mean matching
			Monotone propensity score
Monotone	Classification (ordinal)	Arbitrary	Monotone logistic regression
	Classification (nominal)	Arbitrary	Monotone discriminant function
Arbitrary	Continuous	Continuous	MCMC full data-imputation
			MCMC monotone data-imputation
Arbitrary	Continuous	Arbitrary	FCS regression
			FCS predicted mean matching
Arbitrary	Classification (ordinal)	Arbitrary	FCS logistic regression
Arbitrary	Classification (nominal)	Arbitrary	FCS discriminant function
			FCS logistic regression

3.3 Fully Conditional Specification (FCS)

FCS is also known as imputation using Chained Equations. It assumes a joint distribution exists for the imputed variables (Buuren, 2007), works for both continuous and categorical data and does not assume normality. Imputation of multivariate missing data is done per variable in an iterative fashion (Buuren, 2012). It requires an imputation model to be specified for every incomplete variable. Every variable is then imputed using its full conditional distribution on all the other specified variables.

Carpenter and Kenward (2012) states that in FCS, to fill in the missing values, draws are made with replacement from the observed.

FCS involves two phases (SAS, 2017):

1. Fill-in phase - Missing values are filled in sequentially over the variables with preceding variables used as the covariates. For p variables in the order $Y_1, Y_2, \cdots, Y_p$, missing values are replaced using the sequence given below.

$$
\begin{aligned}
\theta_1^{(0)} &\sim P(\theta_1 | Y_{1(obs)}) \\
Y_{1(*)}^{(0)} &\sim P(Y_1 | \theta_1^{(0)}) \\
Y_1^{(0)} &= (Y_{1(obs)}, Y_{1(*)}^{(0)}) \\
&\vdots \\
\theta_p^{(0)} &\sim P(\theta_p | Y_1^{(0)}, \cdots, Y_{p-1}^{(0)}, Y_{p(obs)}) \\
Y_{p(*)}^{(0)} &\sim P(Y_p | \theta_p^{(0)}) \\
Y_p^{(0)} &= (Y_{p(obs)}, Y_{p(*)}^{(0)})
\end{aligned}
\tag{4}
$$

where

$Y_{j(\text{obs})}$ is a set of observed Y_j values, $Y^{(0)}_{j(*)}$ is the set of filled-in Y_j values, $Y^{(0)}_j$ is the set of both observed and filled-in Y_j values, and $\theta^{(0)}_j$ is the set of simulated parameters for $(Y_j|Y_1, Y_2, \cdots, Y_{j-1})$.

2. Imputation phase - Filled-in values provide the starting values for the missing values. A number of iterations are performed and in each variable, the imputed values are used for imputation. At each iteration t, imputation is done sequentially over the variables, and these imputed values replace the filled-in values.

$$
\begin{aligned}
\theta^{(t)}_1 &\sim P(\theta_1 \,|\, Y_{1(obs)}, Y^{(t-1)}_2, \cdots, Y^{(t-1)}_p) \\
Y^{(t)}_{1(*)} &\sim P(Y_1 \,|\, \theta^{(t)}_1) \\
Y^{(t)}_1 &= (Y_{1(obs)}, Y^{(t)}_{1(*)}) \\
&\vdots \\
\theta^{(t)}_p &\sim P(\theta_p \,|\, Y^{(t)}_1, \cdots, Y^{(t)}_{p-1}, Y_{p(obs)}) \\
Y^{(t)}_{p(*)} &\sim P(Y_p \,|\, \theta^{(t)}_p) \\
Y^{(t)}_p &= (Y_{p(obs)}, Y^{(t)}_{p(*)})
\end{aligned}
\tag{5}
$$

where

$Y^{(t)}_{j(*)}$ is the set of imputed Y_j values at iteration t, $Y^{(t-1)}_{j(*)}$ is the set of filled-in Y_j values ($t = 1$) or the set of imputed Y_j values at iteration $t (t > 1)$, $Y^{(t)}_j$ is the set of both observed and imputed Y_j values at iteration t, and $\theta^{(t)}_j$ is the set of simulated parameters for the conditional distribution of Y_j given covariates constructed from $Y_1, \cdots, Y_{j-1}, Y_{j+1}, \cdots, Y_p$.

This forms a cycle and several cycles are performed so that the algorithm converges to a stationary distribution. The current values of Y_i, for $i = 1, \cdots, n$, then form the first imputed dataset. More cycles are performed to give K imputed dataset which are stochastically independent of each other.

3.3.1 Continuous Variables

FCS Regression Method

A regression model is fitted with the specified set of effects as the covariates. Missing values are imputed for each variable by simulating a new regression model, based on the fitted model, from the posterior predictive distribution of the parameters. For a partially observed variable Y_j, a model $Y_j = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ is fitted using observed values in Y_j and its covariates $X_1, X_2, \cdots, X_k$.

The fitted model includes a regression with the parameter estimates $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)$ and associated covariance matrix $\hat{\sigma}^2 \boldsymbol{V}_j$, where $\boldsymbol{V}_j$ is $\boldsymbol{X}^T \boldsymbol{X}$ inverse matrix derived from the intercept and covariates $x_1, x_2, \cdots, x_k$.

The imputed values are obtained as follows:

1. New parameters $\boldsymbol{\beta}_* = (\beta_0, \beta_1, \cdots, \beta_k)$ and $\hat{\sigma}^2_{*j}$ are drawn from the posterior predictive distribution of the parameters; that is, they are simulated from $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)$, $\hat{\sigma}^2_{*j}$ and V_j. The variance is then drawn as $\hat{\sigma}^2_{*j} = \hat{\sigma}^2(nj\text{–}k\text{–}1)/g$, where g is a χ^2_{nj-k-1} random variate and n_j is the number of non-missing observations for Y_j. The regression coefficients are drawn as $\boldsymbol{\beta}_* = \hat{\boldsymbol{\beta}} + \sigma_{*j} \boldsymbol{V}^T_{hj} \boldsymbol{Z}$, where $\boldsymbol{V}^T_{hj} \boldsymbol{Z}$ is the upper triangular matrix

in the Cholesky decomposition, $\boldsymbol{V}_j = \boldsymbol{V}_{hj}^T \boldsymbol{V}_{hj}$, and $\boldsymbol{Z}$ is a vector of $k+1$ independent random normal variates.

2. The missing values are then replaced by $\beta_{*0} + \beta_{*1} x_1 + \cdots + \beta_{*k} x_k + z_i \sigma_{*j}$, where $x_1, x_2, \cdots, x_k$ are the values of the covariates and z_i is a simulated normal deviate.

FCS Predictive Mean Matching (PMM)

This is a popular approach for creating imputed values based on the predicted value of the variable with missing values. PMM is used to impute missing data for any missing data pattern in quantitative variables, especially when the normality assumption is either not met, not plausible or variables are non-linearly related. In such cases it may be considered more appropriate compared to regression method (Horton and Lipsitz, 2001). PMM is robust to transformations of target variable (Buuren, 2012), making it attractive. It also produces values within the range of the observed data. It is known to perform well when missingness is less than 50% and the missing data are not MNAR.

The imputed values are generally obtained in a similar fashion as the regression method, but imputation is done by randomly drawing from a set of observed values using the following algorithm:

1. Similar to the first step of the FCS regression method.

2. For every missing value, a predicted value $y_{*i} = \beta_{*0} + \beta_{*1} x_1 + \cdots + \beta_{*k} x_k$, is computed with covariates $x_1, x_2, \cdots, x_k$.

3. For $i = 1, \cdots, n_1$, y_{*i} is a predicted value of observed y_i cases.

   For $j = 1, \cdots, n_0$, y_{*j} is a predicted value of missing y_j cases.

(a) Select a donor, which is the closest candidate for which $|y_{*i}-y_{*j}|$ is minimal.

(b) Obtain a set of d candidate donors for which $|y_{*i}-y_{*j}|$ is minimal.

(c) Sample one donor with a probability that depends on $|y_{*i}-y_{*j}|$.

(d) Make a random draw from the d observed values (y_i) to replace the missing value.

The underlying assumption of PMM is that donors and receivers have the same distribution as the target variable, within the group of candidate donors (Buuren, 2012). The variability from imputations over repeated draws is another reflection of the uncertainty of the unobserved value.

3.3.2 Categorical Data

The logistic method is useful for classification variables. Classification variables can be either binary (with only two levels), multinomial (with more than two un-ordered levels) or ordinal (with more than two ordered levels).

For a variable p,

$$logit(p) = log\left(\frac{p}{1-p}\right)$$

Binary Logistic Method

It is used when the dependent variable has only two categories. To impute the missing values for each variable, a logistic regression model is simulated from the posterior predictive distribution of the parameters.

Let Y contain categories 1 and 2. A logistic model is fitted as Y: $logit(p_1) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ for $X_1, X_2, \cdots, X_p$ covariates of $p_1 = Pr(Y = 1|X_1, X_2, \cdots, X_p)$.

Parameter estimates $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)$ with associated covariance matrix $\boldsymbol{V}$, are then obtained from the fitted model.

The imputation algorithm is as follows:

1. New parameters $\hat{\boldsymbol{\beta}}_* = (\hat{\beta}_{*0}, \hat{\beta}_{*1}, \cdots, \hat{\beta}_{*p})$ are drawn from the posterior predictive distribution of the parameters. $\boldsymbol{\beta}_* = \hat{\boldsymbol{\beta}} + \boldsymbol{V}_h^T \boldsymbol{Z}$ where $\boldsymbol{V}_h$ is the upper triangular matrix in the Cholesky decomposition, $\boldsymbol{V} = \boldsymbol{V}_h^T \boldsymbol{V}_h$, and $\boldsymbol{Z}$ is a vector of $p+1$ independent random normal variates.

2. For an observation with missing Y_j and covariates $x_1, x_2, \cdots, x_P$, compute the predicted probability that $Y = 1$ using:

$$p_1 = \frac{exp(\mu_1)}{1 + exp(\mu_1)}, \text{ where } \mu_1 = \beta_{*0} + \beta_{*1} x_1 + \cdots + \beta_{*p} x_p$$

3. Draw a random uniform variate, $0 \leq \nu \leq 1$. If the value of $\nu < p_1$, impute $Y = 1$; else impute $Y = 2$.

Multinomial Logistic Method

It is used when the dependent variable has more than two un-ordered categories. The generalized logit model is of the form:

$$log\left(\frac{Pr(Y = j|\boldsymbol{x})}{1 - Pr(Y = j|\boldsymbol{x})}\right) = \alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{X}, j = 1, 2, \cdots, K-1$$

where $\alpha_1, \cdots, \alpha_{K-1}$ are $K-1$ intercept parameters, and $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_{K-1}$ are $K-1$ the vector of slope parameters.

When imputing, a logistic model is fitted as Y

$$log\left(\frac{p_j}{p_K}\right) = \alpha_j + \beta_{j1}X_1 + \cdots + \beta_{jp}X_p$$

for $X_1, X_2, \cdots, X_p$ covariates of $p_j = Pr(Y \leq j | X_1, X_2, \cdots, X_k)$.

Parameter estimates $\hat{\alpha} = (\hat{\alpha}_0, \cdots, \hat{\alpha}_K - 1)$ and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)$ with associated covariance matrix $\boldsymbol{V}$, where $\hat{\boldsymbol{\beta}}_j = (\hat{\beta}_{j0}, \hat{\beta}_{j1}, \cdots, \hat{\beta}_{jp})$, are then obtained from the fitted model.

The imputation algorithm is as follows:

1. New parameters γ are drawn from the posterior predictive distribution of the parameters. $\gamma = \hat{\gamma} + \boldsymbol{V}_h^T \boldsymbol{Z}$ where $\hat{\gamma} = (\hat{\alpha}, \hat{\boldsymbol{\beta}})$, $\boldsymbol{V}_h$ is the upper triangular matrix in the Cholesky decomposition, $\boldsymbol{V} = \boldsymbol{V}_h^T \boldsymbol{V}_h$, and $\boldsymbol{Z}$ is a vector of $p + K - 1$ independent random normal variates.

2. For an observation with missing Y and covariates $x_1, x_2, \cdots, x_k$ compute the predicted probability that $Y = j, j = 1, 2, \cdots, K - 1$

$$Pr(Y = j) = \frac{exp(\alpha_j + \boldsymbol{x}^T \boldsymbol{\beta})}{\sum_{k=1}^{K-1} exp(\alpha_j +, x^T \boldsymbol{\beta}) + 1}$$

   and

$$Pr(Y = K) = \frac{1}{\sum_{k=1}^{K-1} exp(\alpha_j + \boldsymbol{x}^T \boldsymbol{\beta}) + 1}$$

3. Compute the cumulative probability for $Y \leq j$ as $P_j = \sum_{k=1}^{j} Pr(Y = k)$

4. Draw a random uniform variate, $0 \leq \nu \leq 1$ and impute:

$$Y = \begin{cases} 1 & \text{if } \nu < p_1, \\ k & \text{if } p_{k-1} \leq \nu < p_k, \\ K & \text{if } p_{K-1} \leq \nu \end{cases}$$

Ordinal Logistic Method

It is used when the dependent variable has more than two ordered categories. It is used to model cumulative probabilities.

For a response variable Y with values $1, 2, \cdots, K$, the cumulative model is of the form:

$$logit(Pr(Y \leq j|\boldsymbol{x})) = log\left(\frac{Pr(Y \leq j|\boldsymbol{x})}{1-Pr(Y \leq j|\boldsymbol{x})}\right) = \alpha_j + \boldsymbol{\beta}^T \boldsymbol{X}, j = 1, 2, \cdots, K-1$$

where $\alpha_1, \cdots, \alpha_{K-1}$ are $K-1$ intercept parameters, and $\boldsymbol{\beta}$ is the vector of slope parameters.

When imputing, a logistic model is fitted as Y: $logit(p_j) = \alpha_j + \beta_1 X_1 + \cdots + \beta_p X_p$ for $X_1, X_2, \cdots, X_p$ covariates of Y and $p_j = Pr(Y \leq j|X_1, X_2, \cdots, X_k)$.

Parameter estimates $\hat{\alpha} = (\hat{\alpha}_0, \cdots, \hat{\alpha}_K - 1)$ and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)$ with associated covariance matrix $\boldsymbol{V}$, are then obtained from the fitted model.

The imputation algorithm is as follows:

1. New parameters γ are drawn from the posterior predictive distribution of the parameters. $\gamma = \hat{\gamma} + \boldsymbol{V}_h^T \boldsymbol{Z}$ where $\hat{\gamma} = (\hat{\alpha}, \hat{\boldsymbol{\beta}})$, $\boldsymbol{V}_h$ is the upper triangular matrix in the Cholesky

decomposition, $\boldsymbol{V} = \boldsymbol{V}_h^T \boldsymbol{V}_h$, and $\boldsymbol{Z}$ is a vector of $p + K - 1$ independent random normal variates.

2. For an observation with missing Y and covariates $x_1, x_2, \cdots, x_k$ compute the predicted probability that $Y \leq j$:

$$p_j = Pr(Y \leq j) = \frac{exp(\alpha_j + \boldsymbol{x}^T \boldsymbol{\beta})}{exp(\alpha_j + \boldsymbol{x}^T \boldsymbol{\beta}) + 1}$$

3. Draw a random uniform variate, $0 \leq \nu \leq 1$ and impute

$$Y = \begin{cases} 1 & \text{if } \nu < p_1, \\ k & \text{if } p_{k-1} \leq \nu < p_k, \\ K & \text{if } p_{K-1} \leq \nu \end{cases}$$

3.4 Study design

This study used secondary data obtained from SWOP which adopted a cross-sectional single group design.

3.5 Study Area

The study adopted the same study area used in the primary study which covered Starehe sub-county of Nairobi county.

3.6 Study population

For this study, respondents were comprised of self-identified female sex workers enrolled at SWOP-Kenya clinic at city center, Nairobi clinic, over a four-year period, from June 2014 to June 2018.

Inclusion criteria:

- Female sex worker.
- Enrolled between June 2014 and June 2018.

Exclusion criteria:

- Males.

All respondents that met the inclusion/exclusion criteria were used for analysis.

3.7 Sampling Procedure

In the SWOP program, respondents were recruited through self identification, snow-balling and through peer educators on the ground. There was no pre-determined sample size because the program continually enrols female sex workers and other KPs to provide customized health care services for their needs.

3.8 Analysis Variables

The dependent variable was HIV status of FSWs at the time of enrolment into the SWOP facility. Independent variables were demographic and risk behavioral factors:

Table 4. Analysis Variables

Demographic Factors	Education
Risk Behavioral Factors	Use of condoms
	Douching
	Number of sex acts per week

3.9 Conceptual Framework

In this framework, there were certain factors considered to influence HIV positivity among FSWs which we broadly grouped into demographic and behavioral factors. It is important to mention that factors influencing HIV positivity among the FSWs are not just limited to these, but the other factors were beyond the scope of this study.

The HIV status upon enrollment was the response variable that was associated with a set of independent variables as illustrated in Figure 2:

3.9.1 Analysis Model

The target analysis was examining the association between HIV status at enrolment into the SWOP clinic and the risk factors of HIV positivity among FSWs adjusting for other source of income, engaging in sex under influence of alcohol, use of drugs, number of casual clients per week, number of regular clients per week and duration of sex work.

Table 5 describes the analysis variables.

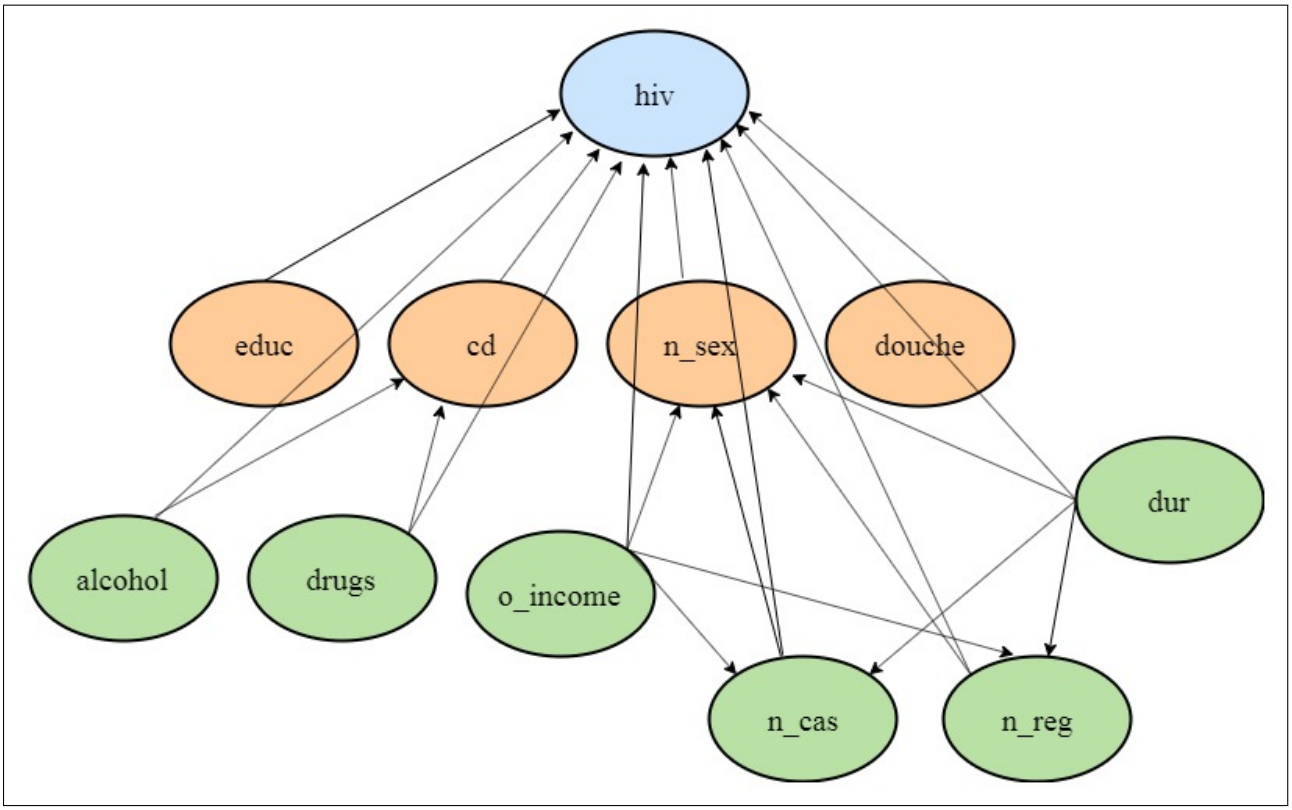


Figure 2. Causal diagram for the association between Education Level (educ), Use of condoms consistently (cd), Number of sex acts per week (n_sex), Douching (douche) and HIV status (hiv) adjusting for Engaging in sex under influence of alcohol (alcohol), Use of drugs (drugs), Other source of income (o_income), Number of casual clients per week (n_cas), Number of regular clients per week (n_reg) and Duration of sex work (dur).

Table 5. Description of variables: Outcome variable, Risk factors of HIV and Confounders

Role	Variable Description	Type	Grouping/Unit	Variable Name
Response	HIV Status	Categorical	1=Positive 2=Negative 3=I do not want to share	hiv
Covariate	Education level	Categorical	1=Completed primary 2=Did not complete primary 3=Completed secondary 4=Did not complete secondary 5=Completed tertiary 6=Did not complete tertiary 7=Never attended school 8=Other	educ
Covariate	Use of condoms consistently	Categorical	1=Yes 2=No	cd
Covariate	Number of sex acts per week	Continuous	Count	n_sex
Covariate	Douching (Vaginal)	Categorical	1=Yes 2=No	douche
Confounder	Other source of income	Categorical	1=Yes 2=No	o_income
Confounder	Engage in sex under influence of alcohol	Categorical	1=Never 2=Sometimes 3=Most times 4=Always	alcohol
Confounder	Use of drugs	Categorical	1=Yes 2=No	drugs
Confounder	Number of casual clients per week	Continuous	Count	n_cas
Confounder	Number of regular clients per week	Continuous	Count	n_reg
Confounder	Duration of sex work	Continuous	Years	dur

3.9.2 Imputation Model

For this study all variables in the analysis model were used and auxiliary variables based on literature. Training on harm reduction regarding alcohol/drugs use and training on condom negotiation were used as auxiliary variables for HIV Status and are described in Table 6. Imputation was performed in SAS version 9.4 and for both methods of MI, 10 imputations were performed with a seed of 1305. For PMM, number of donors used were 10.

Table 6. Description of auxiliary variables

Role	Variable Description	Type	Grouping	Variable Name
Auxiliary	Trained on harm reduction regarding alcohol/drugs use	Categorical	1=Yes 2=No	harm_red
Auxiliary	Trained on condom negotiation	Categorical	1=Yes 2=No	cd_neg

3.10 Quality Assurance Procedures

Quality assurance and quality control was implemented and maintained with adequate ethical considerations, adherence to the objectives and scope of this study and principles of Good Clinical Practices. All data was documented, recorded and reported in compliance with the analysis plan.

3.11 Ethical Considerations

Ethical approval was sought from the SWOP team through the University of Manitoba-University of Nairobi. For confidentiality and anonymity of subjects, all personal identifiers

were removed by the SWOP data manager during file conversion from the SWOP database to an Excel spreadsheet. Only unique alphanumeric subject identifiers were used.

3.12 Data Management Plan

3.12.1 Data Source

Secondary data was used.

The source of data was a sex workers' research database focused on HIV prevention, care and treatment, behavioral interventions, biomedical interventions and structural inventions on gender-based violence. The database is maintained by the University of Manitoba in partnership with the University of Nairobi through the Sex Worker Outreach Program (SWOP) - Kenya. SWOP's focus is on the vulnerable groups that face social and structural barriers preventing them from accessing health care. Their programs deliver a whole range of sexual and reproductive health interventions aimed at meeting KP specific needs in order to optimize uptake.

3.12.2 Data Collection

The bio-data was obtained from the SWOP database which was originally collected using a FSW questionnaire through face-to-face interviews. Verbal informed consent was obtained by peer educators. Upon collection, data was entered into the research database.

3.12.3 Data Entry

The data manager at the SWOP clinic ensured that all data in the questionnaires and queries were accurate and complete and that all entries were verifiable with source documents.

These documents are maintained by the SWOP clinic. The data manager also verified the data in the questionnaires with that in the database and confirmed that there were no inconsistencies between them.

3.12.4 Software and statistical considerations

Data cleaning and variable categorization was done using the SAS software version 9.4. The tests were done in SAS software version 9.4 and/or R version 3.6.1 and were considered statistically significant at 5% level of significance.

3.12.5 Analysis Plan

General Considerations: Exploratory Analysis Descriptive statistics were presented for continuous variables, and included n (number of subjects), mean, standard deviation (SD), median, minimum and maximum values. For categorical data, frequencies and percentages were displayed.

Statistical comparisons were made using two-sided tests at $\alpha = 0.05$ level of significance.

All data processing, summarizing and analyses were performed using SAS version 9.4 and/or R version 3.6.1.

Checking the Extent and Pattern of Missing data

Proportions,in percentage, of missingness in all affected variables were presented graphically using aggregate plots.

Further the R VIM package was used to graphically show missingness patterns using R version 3.6.1, as suggested by Templ et al. (2019); Prantner (2011).

Checking the Missing Data Mechanism

For MCAR the null hypothesis H_0: Data is MCAR was tested using the LittleMCAR test under the BaylorEdPsych R library (Beaujean, 2012). A few tests have been proposed to test MAR versus MCAR but due to their unclear practicality, they are not widely used (Buuren, 2012). Generally, it's impossible to test MAR versus MNAR because the information needed for the test is missing (Buuren, 2012). It needs to be ascertained whether or not missing values are considered MNAR from a researcher's understanding of the variables under investigation. In this study we assumed MAR assumption for imputation, when the test for MCAR was significant.

Assessing impact of Missing data

Standard errors of the parameter estimates and statistical significance of p-values in CCA were compared with those of MI.

Assessing Performance of MI Methods

Evaluation of the MI methods was done by assessing:

1. Adjusted Odds Ratios and corresponding Confidence Interval width in tabular comparison.

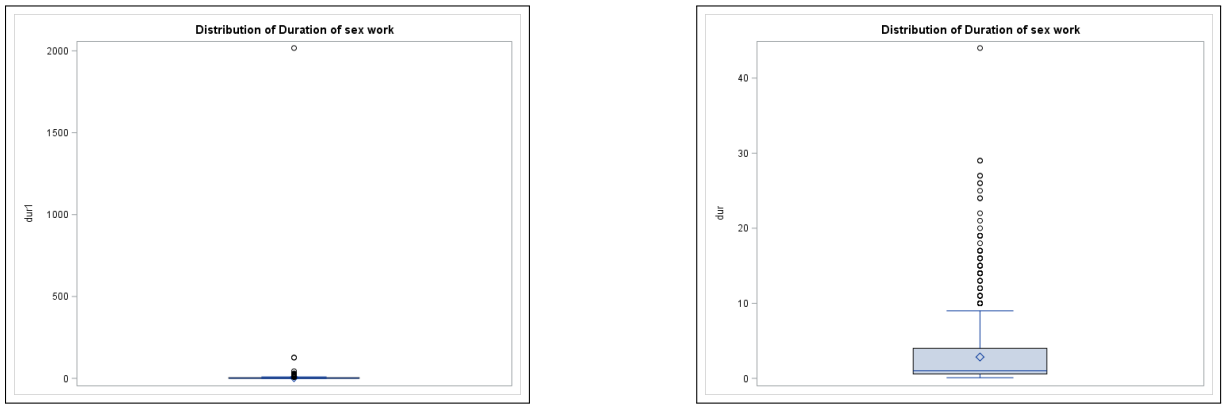
2. Distributional Properties - Diagnostic checks were done and displayed graphically using kernel density plots for continuous variables and bar plots for categorical variables.

3. Convergence of the algorithm - trace plots were displayed for continuous variables.

4. Relative efficiency of the MI methods

Outliers

Outliers were set to missing and imputed using multiple imputation. The duration of sex work contained some outliers as shown in Figure 3:



(a) with outliers **(b) without outliers**

Figure 3. Outliers in Duration of sex work

Missing Data

Missing data was imputed using the multiple imputation technique.

4 RESULTS

4.1 Introduction

This chapter gives the results obtained after testing and examining the specific objectives of the study.

4.2 Overview of the Data

One of the options of HIV status in the questionnaire was "I do not want to share". We had only three such responses in the original data, which caused quasi-complete separation of data in the analysis, and questionable convergence of the model. To get around this we set these three responses to missing and imputed stochastically as either Positive or Negative.

In Table 7, we see that 8.4% reported to be HIV positive while 10.3% did not reveal their HIV status. Majority (33.2%) of the FSWs completed secondary school while only 0.7% never attended school. More than half of the female sex workers (56.7%) used condoms consistently. We also found that, on average, a FSW engaged in 15 sex acts in a week. 77.5% reported practising vaginal douching.

Table 7. Baseline Characteristics of Female Sex Workers

Parameter	Category/Statistic	Female Sex Workers (N=1734) n (%)
HIV Status: n (%)	Positive	146 (8.4)
	Negative	1409 (81.3)
	Missing	179 (10.3)
Education Level: n (%)	Completed primary	353 (20.4)
	Did not complete primary	189 (10.9)
	Completed secondary	576 (33.2)
	Did not complete secondary	359 (20.7)
	Completed tertiary level	183 (10.6)
	Did not complete tertiary level	28 (1.6)
	Never attended school	12 (0.7)
	Other	27 (1.6)
	Missing	7 (0.4)
Use of condoms consistently: n (%)	Yes	983 (56.7)
	No	709 (40.9)
	Missing	42 (2.4)
Number of sex acts per week	n	1592
	Mean (SD)	15.1 (12.41)
	Median (Min, Max)	12.0 (1, 127)
Douching: n (%)	Yes	1343 (77.5)
	No	333 (19.2)
	Missing	58 (3.3)
Other source of income: n (%)	Yes	464 (26.8)
	No	1194 (68.9)
	Missing	76 (4.4)
Engage in sex under influence of alcohol: n (%)	Never	543 (31.3)
	Sometimes	969 (55.9)
	Most times	110 (6.3)
	Always	75 (4.3)
	Missing	37 (2.1)
Use of drugs: n (%)	Yes	611 (35.2)
	No	1007 (58.1)
	Missing	116 (6.7)
Number of casual clients per week	n	1695
	Mean (SD)	10.1 (9.71)
	Median (Min, Max)	7.0 (0, 127)
Number of regular clients per week	n	1674
	Mean (SD)	4.9 (5.25)
	Median (Min, Max)	4.0 (0, 50)
Duration of sex work (years)	n	1510
	Mean (SD)	2.8 (3.98)
	Median (Min, Max)	1.0 (0, 44)

4.3 Proportion and Pattern of Missing Data

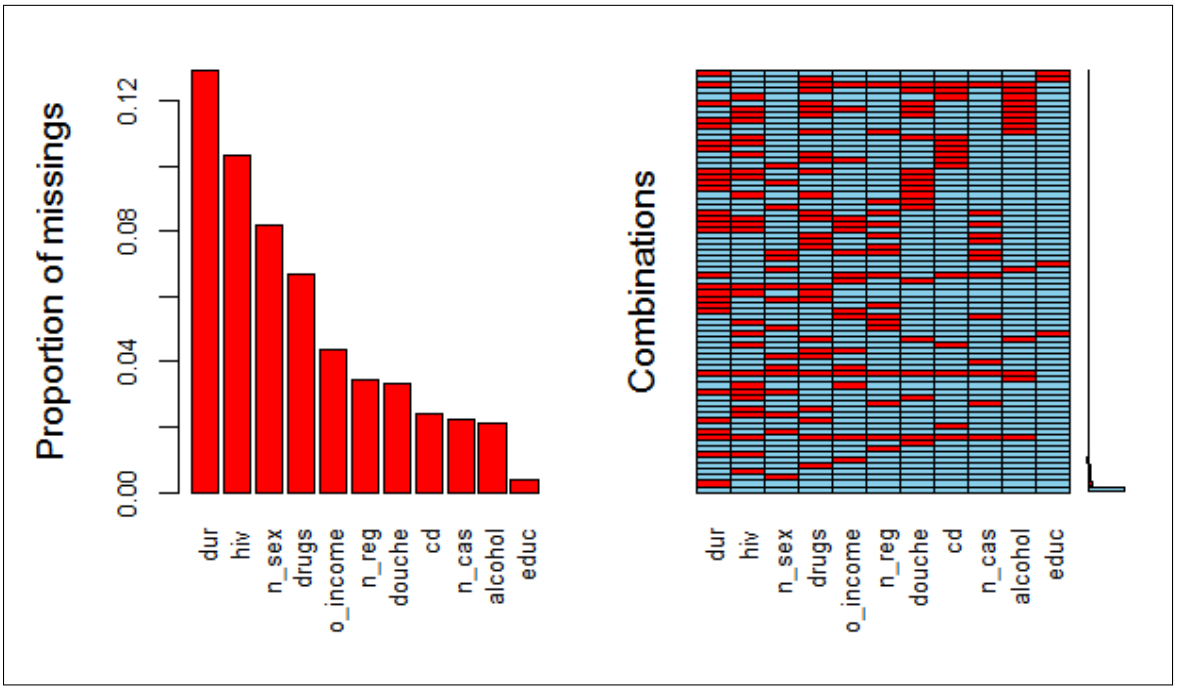


Figure 4. Missing Data Aggregate plot: Blue - observed data, Red - Missing data

On the left side of Figure 4 we see the proportions of missing data in all the variables, in decreasing order. Generally, there were low proportions of nonresponse. All analysis variables contained some level of missingness. The duration of sex work, had the highest proportion (12.9%) of missingness while education level had the lowest (0.4%). On the right side, we have the possible combinations of missing (red) and observed (blue) data, which give us the missing data patterns. There are a few cases with univariate missingness pattern, and arbitrary missingness in most parts and therefore we conclude that the missingness pattern is arbitrary.

4.4 Missing Data Mechanism

As shown in Table 8, the significant p-value of 0.0011 from Little's MCAR test suggested that the MCAR assumption was not met.

Table 8. LittleMCAR results

Chi square	706.8584
Degrees of freedom	596
p-value	0.001144109
Missing patterns	73

We therefore assumed the MAR assumption for this data.

4.5 Impact of missing data

The parameter estimates, their standard errors and p-values are presented in Table 9. In all three methods, consistent use of condoms (cd) was the only predictor found to be significantly, though negatively, associated with HIV positivity, adjusting for other variables. We also see that missingness had no impact on statistical significance of the association of the covariates with HIV positivity. However, standard errors of the parameter estimates are higher in CCA as compared to MI. Level of education has the highest standard errors in CCA. High standard errors implying low precision.

Of all considered confounders, duration of sex work was found to be significant.

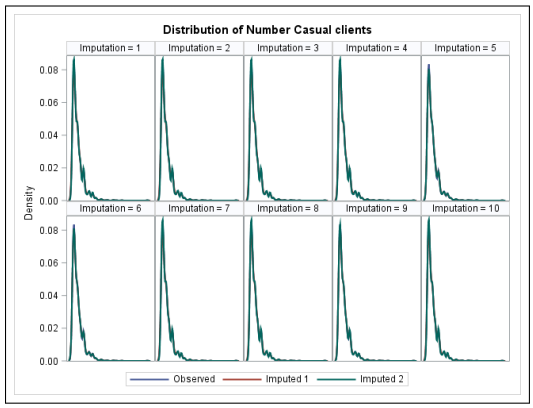
Table 9. Model Parameter Estimates

		CCA (N=1128)			FCS-Model based (N=1734)			FCS-PMM (N=1734)		
Parameter	Class levels	β	SE	p-value	β	SE	p-value	β	SE	p-value
Intercept		−4.371	117.400	0.970	−2.493	0.287	<.0001	−2.661	0.280	<.0001
educ	Never attended school	0			0			0		
	Completed primary	1.891	117.400	0.987	0.191	0.251	0.448	0.324	0.250	0.195
	Completed secondary	1.210	117.400	0.992	−0.399	0.257	0.122	−0.259	0.249	0.299
	Completed tertiary level	1.128	117.400	0.992	−0.495	0.380	0.194	−0.399	0.344	0.247
	Did not complete primary	2.045	117.400	0.986	0.121	0.300	0.687	0.266	0.285	0.351
	Did not complete secondary	1.819	117.400	0.988	−0.011	0.261	0.966	0.117	0.257	0.649
	Did not complete tertiary level	1.283	117.400	0.991	−0.772	0.920	0.402	−0.534	0.903	0.555
	Other	−12.346	821.500	0.988	−0.567	0.960	0.556	−0.726	0.933	0.437
cd	No	0			0			0		
	Yes	0.412	0.127	0.001	0.353	0.101	0.001	0.363	0.102	0.000
n_sex		−0.004	0.011	0.747	0.004	0.009	0.609	0.008	0.010	0.431
douche	No	0			0			0		
	Yes	0.062	0.146	0.671	0.029	0.121	0.809	0.037	0.117	0.754
o_income	No	0			0			0		
	Yes	0.044	0.127	0.731	0.035	0.107	0.742	0.051	0.104	0.627
alcohol	Never	0			0			0		
	Always	0.515	0.376	0.171	0.500	0.301	0.099	0.468	0.292	0.109
	Most times	−0.830	0.470	0.077	−0.325	0.315	0.304	−0.348	0.333	0.299
	Sometimes	−0.076	0.226	0.738	−0.220	0.165	0.183	−0.168	0.168	0.318
drugs	No	0			0			0		
	Yes	0.025	0.131	0.847	−0.046	0.112	0.684	−0.054	0.103	0.599
n_cas		0.003	0.014	0.838	−0.003	0.012	0.770	−0.005	0.012	0.690
n_reg		−0.018	0.026	0.496	−0.028	0.021	0.195	−0.026	0.021	0.217
dur		0.075	0.021	0.000	0.108	0.018	<.0001	0.100	0.018	<.0001

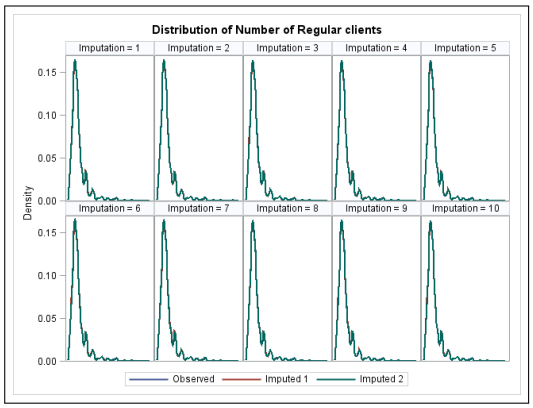
4.6 Performance of MI

4.6.1 Distributional Properties

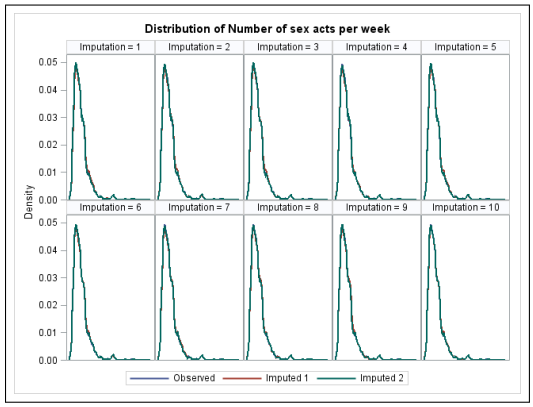
In Figure 5, we see that in all continuous variables of our analysis model, the kernel densities for imputation 1 (model-based FCS) and imputation 2 (PMM) are identical to that of the observed.



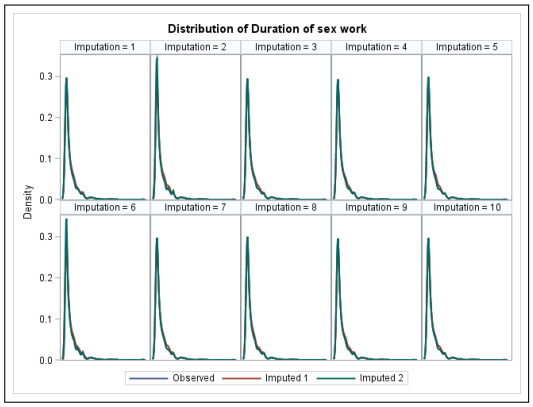
(a) n_cas



(b) n_reg



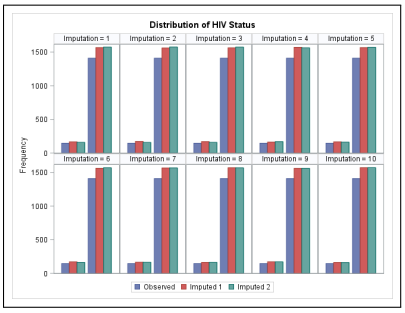
(c) n_sex



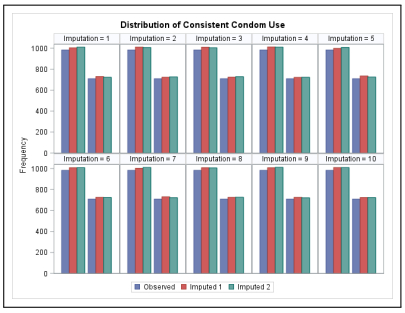
(d) dur

Figure 5. Kernel Densities of Imputed Continuous Variables

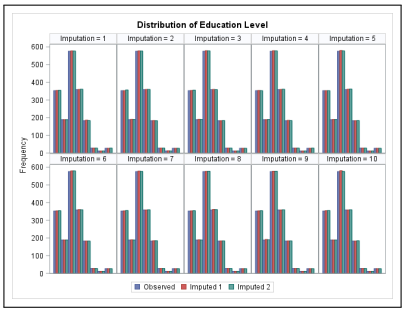
Frequency distributions by the two imputation strategies were also very similar to those of the observed, as illustrated in Figure 6. We therefore conclude that the MI techniques preserved the distributions in the data.



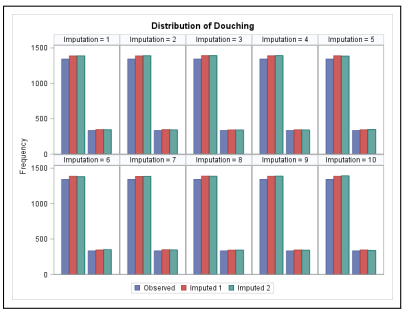
(a) hiv



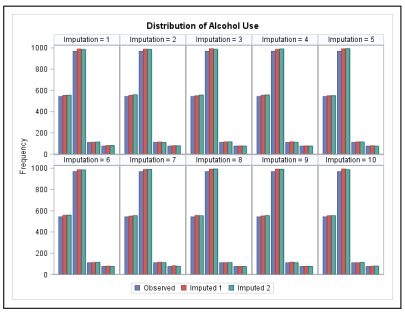
(b) cd



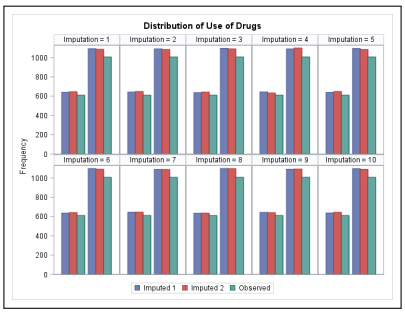
(c) educ



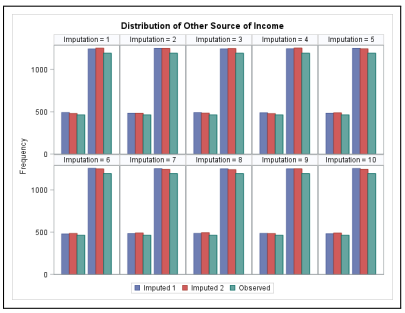
(d) douche



(e) alcohol



(f) drugs

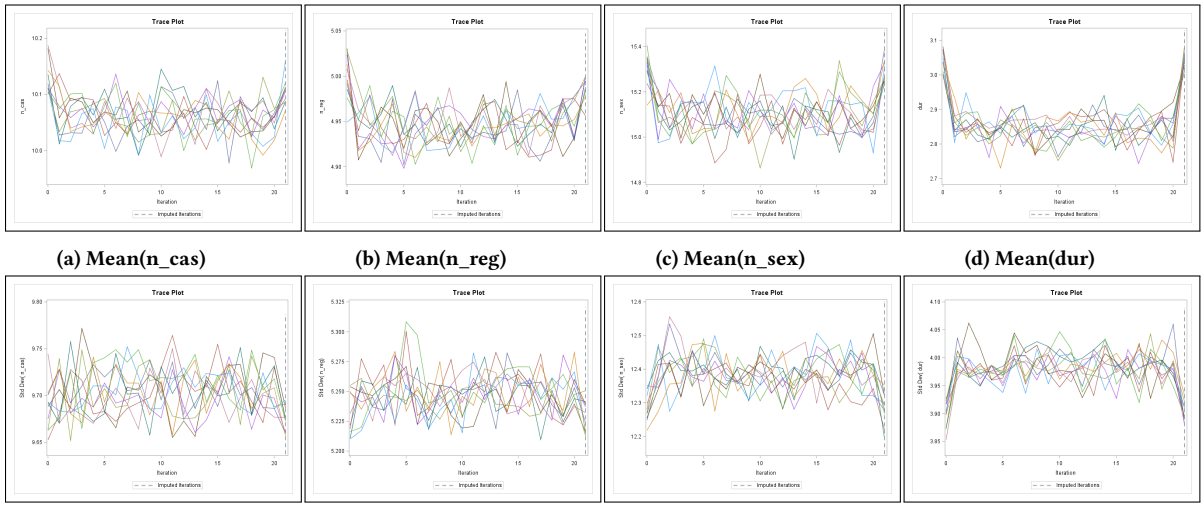


(g) o_income

Figure 6. Frequency Distributions of Imputed Categorical Variables

4.6.2 Convergence of Algorithm

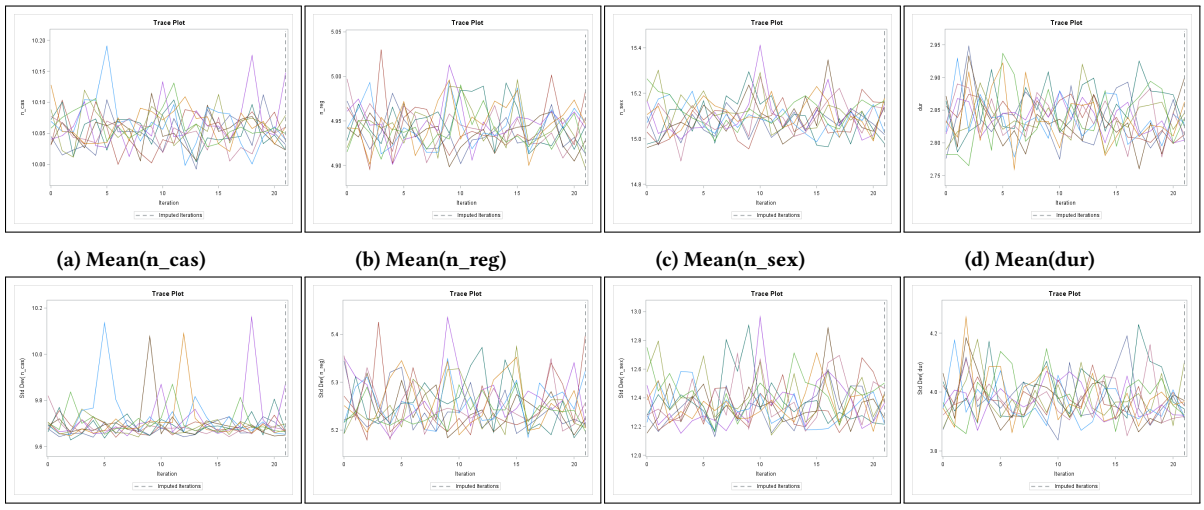
Long-term trends in a trace plot can indicate high correlation of successive iterations and that the series of iterations has not converged. All trace plots in Figure 7 and almost all in Figure 8, show no trends which implies healthy convergence in both MI methods. The standard deviation trace plot for number of casual clients (n_cas) shows unhealthy convergence in some iterations.



(a) Mean(n_cas) (b) Mean(n_reg) (c) Mean(n_sex) (d) Mean(dur)

(e) SD(n_cas) (f) SD(n_reg) (g) SD(n_sex) (h) SD(dur)

Figure 7. Trace plots - model based multiple imputation



(a) Mean(n_cas) (b) Mean(n_reg) (c) Mean(n_sex) (d) Mean(dur)

(e) SD(n_cas) (f) SD(n_reg) (g) SD(n_sex) (h) SD(dur)

Figure 8. Trace plots - PMM

4.6.3 Adjusted Odds Ratios and Interval Width

Table 10 shows that even when not statistically significant, conclusions drawn across the three methods are largely consistent.

In all three methods, adjusting for other variables, all levels of education had lower odds of HIV positivity compared to those who never attended school.

An interesting find in consistent condom use was that in all the three methods, adjusting for other variables, the odds of HIV positivity for female sex workers who consistently used condoms were twice as high as those who didn't: [2.278(1.385,3.747)], [2.025(1.365,3.004)] and [2.069(1.388,3.085)] for CCA, model-based FCS and PMM respectively. These results were found to be statistically significant.

In CCA we see that, adjusting for other variables, an increase in the average number of sex acts per week decreased the odds of HIV positivity by 0.4% [0.996(0.974,1.019)], whereas in MI we see that an increase in the average number of sex acts per week increased the odds of HIV positivity by 0.4% [1.004(0.988,1.022)] and 0.8% [1.008(0.988,1.028)] for model-based FCS and PMM respectively.

Those who practised douching had slightly higher odds of HIV positivity compared to those who didn't, adjusting for other variables, in all methods: [1.132(0.638,2.01)], [1.060(0.659,1.707)] and [1.076(0.679,1.706)] for CCA, model-based FCS and PMM respectively.

Interval width in CCA is larger than MI in most parts, particularly categorical variables. For continuous predictors, the interval width is not far off from the MI methods. PMM also shows larger interval widths in Education level (educ) compared to model based MI. In other variables, we see fairly equal widths to those of the model based MI.

Table 10. Adjusted Odds Ratios (ORs) and Confidence Interval (CI) Width

		CCA (N=1128)		FCS-Model based (N=1734)		FCS-PMM (N=1734)	
Parameter	Class Levels	OR (CI)	Width	OR (CI)	Width	OR (CI)	Width
educ	Never attended school	-		-		-	
	Completed primary	0.34 (0.024, 4.736)	4.712	0.175 (0.033, 0.916)	0.882	0.412 (0.089, 1.915)	1.826
	Completed secondary	0.172 (0.012, 2.415)	2.403	0.097 (0.018, 0.524)	0.506	0.230 (0.049, 1.090)	1.042
	Completed tertiary level	0.159 (0.01, 2.432)	2.422	0.088 (0.015, 0.533)	0.518	0.200 (0.040, 1.002)	0.962
	Did not complete primary	0.397 (0.028, 5.605)	5.577	0.163 (0.029, 0.930)	0.902	0.389 (0.083, 1.825)	1.742
	Did not complete secondary	0.317 (0.023, 4.443)	4.421	0.143 (0.026, 0.778)	0.752	0.335 (0.070, 1.599)	1.529
	Did not complete tertiary level	0.185 (0.007, 5.083)	5.076	0.067 (0.006, 0.786)	0.781	0.175 (0.013, 2.292)	2.279
	Other	<0.001 (<0.001, >999.999)		0.082 (0.005, 1.331)	1.326	0.144 (0.011, 1.977)	1.967
cd	No	-		-		-	
	Yes	2.278 (1.385, 3.747)	2.362	2.025 (1.365, 3.004)	1.639	2.069 (1.388, 3.085)	1.697
n_sex		0.996 (0.974, 1.019)	0.044	1.004 (0.988, 1.022)	0.034	1.008 (0.988, 1.028)	0.039
douche	No	-		-		-	
	Yes	1.132 (0.638, 2.01)	1.372	1.060 (0.659, 1.707)	1.049	1.076 (0.679, 1.706)	1.026
o_income	No	-		-		-	
	Yes	1.091 (0.664, 1.792)	1.128	1.073 (0.706, 1.631)	0.925	1.106 (0.736, 1.663)	0.927
alcohol	Never	-		-		-	
	Always	1.131 (0.431, 2.972)	2.542	1.576 (0.713, 3.485)	2.772	1.524 (0.720, 3.225)	2.505
	Most times	0.295 (0.084, 1.028)	0.944	0.691 (0.299, 1.594)	1.294	0.674 (0.280, 1.622)	1.341
	Sometimes	0.627 (0.379, 1.037)	0.658	0.767 (0.509, 1.157)	0.647	0.807 (0.535 , 1.216)	0.681
drugs	No	-		-		-	
	Yes	1.052 (0.631 , 1.754)	1.123	0.913 (0.589, 1.415)	0.827	0.897 (0.599, 1.344)	0.744
n_cas		1.003 (0.976, 1.03)	0.054	0.997 (0.974 , 1.020)	0.045	0.995 (0.973, 1.018)	0.045
n_reg		0.983 (0.934, 1.033)	0.099	0.973 (0.933 , 1.014)	0.081	0.974 (0.934, .016)	0.082
dur		1.078 (1.034, 1.124)	0.090	1.114 (1.075, 1.155)	0.080	1.106 (1.067, 1.145)	0.078

4.6.4 Relative Efficiency

Table 11 shows that PMM performed better with higher efficiency (closer to 1) in most variables and variable levels. PMM had more occurrences of 0.99 and a few 0.98 while the model based FCS had mostly 0.98 and 0.97.

Table 11. Relative Efficiency of MI methods

Parameter	Class levels	Model-based FCS	PMM
intercept		0.9860	0.9935
educ	Never attended school	-	-
	Completed primary	0.9886	0.9945
	Completed secondary	0.9847	0.9955
	Completed tertiary level	0.9737	0.9948
	Did not complete primary	0.9799	0.9930
	Did not complete secondary	0.9863	0.9939
	Did not complete tertiary level	0.9773	0.9760
	Other	0.9671	0.9865
cd	No	-	-
	Yes	0.9883	0.9877
n_sex		0.9837	0.9595
douche	No	-	-
	Yes	0.9841	0.9911
o_income	No	-	-
	Yes	0.9802	0.9850
alcohol	Never	-	-
	Always	0.9783	0.9880
	Most times	0.9805	0.9767
	Sometimes	0.9906	0.9902
drugs	No	-	-
	Yes	0.9801	0.9946
n_cas		0.9830	0.9849
n_reg		0.9853	0.9823
dur		0.9872	0.9852

5 DISCUSSION

This study was motivated by the missing data problem in HIV research, particularly in key populations like sex workers. Missing data is almost inevitable in every research, naïve application of CCA can often lead to biased estimates under the MAR assumption. When the MAR assumption is met, multiple imputation produces unbiased results. We applied MI using the fully conditional specification method to handle missing data in the SWOP female sex worker data. The broad aim of this study was to assess whether missing data affected association of HIV risk factors to HIV status; and also compare the performance of two MI methods: FCS using model based imputation (fitting an appropriate model for each variable type: FCS Logistic regression for binary and ordinal variables, FCS Discriminant function for multinomial variables and FCS Regression for continuous variables) and FCS PMM.

The FCS method of MI was used to handle the missing data problem in HIV research of female sex workers to minimise bias, maximize the available information and obtain valid estimates whilst preserving uncertainty of missing values. The SWOP program of University of Manitoba is a valuable source of sex worker data. For this study, we used the bio-data of female sex workers in the SWOP database. The nature of the data was ideal to test the objectives. Efficient handling of missing data is important for realising unbiased answers to research questions. Auxiliary variables were used to predict HIV status: training on harm reduction on drug and alcohol use as described by Muraguri (2010); Costa (2007), and training on condom negotiation (Muraguri, 2010; Israel et al., 2008).

There was a response rate of 94.9% in the entire data matrix, only considering the analysis variables, and 12.9% was the highest proportion of missingness in the variables. The fairly low levels of missingness, gave us confidence that MI estimates are unbiased. High proportions of missingness lead to biased estimates, (De Silva et al., 2019; Welch, 2015). The arbitrary missing pattern also allowed us to use the flexible FCS method.

We established that the missingness did not affect conclusions drawn on the associations between the covariates of HIV positivity among female sex workers and their HIV status. For this reason, we conclude plausibility of MAR assumption, (Carpenter and Kenward, 2007). Little's MCAR test was significant, ruling out the MCAR assumption. Also, because there were no marked changes between MI and CCA, we rule out the possibility of perfect prediction by the MI methods used, given the number of categorical variables imputed.

The interpretation of Tables 9 and 10 is consistent pertaining the magnitude of standard errors and the interval width of the odds ratios. CCA had larger standard errors compared to MI in almost all variables. The results agree with most literature on MI, given the significant drop in standard errors of parameter estimates and consequently narrower width of confidence intervals. The differences in standard errors could be due to the drawbacks of CCA: loss of information/data and biased estimates if not MCAR, which is similar to previously done studies (De Silva et al., 2019; Chinomona and Mwambi, 2015; Welch, 2015). The significant loss of information in CCA is seen in the huge drop of number of observations (CCA=1128, MI=1734). Performance of the model improved with increased sample size, which is consistent with Rombach et al. (2018), both MI methods performed better than CCA.

Like Campeau et al. (2018), HIV risk behavioural factors of the female sex workers were not significantly associated with HIV. An interesting find was that consistent use of condoms

was significantly negatively associated with being HIV positive compared to inconsistent use. Further investigation established that about three quarters (71.9%) of those who were HIV positive at the time of enrolment used condoms consistently while about half (56.1%) of those who were HIV negative at the time of enrolment used condoms consistently. Though not statistically significant, under MI, multiple sex partners and frequency of partner change measured by number of sex acts per week was found to increase the odds of being HIV positive, adjusting for other variables, which is consistent with Coetzee et al. (2017), unlike CCA. Duration of sex work was found to be a significant confounder of the number of sex acts per week.

As suggested by Vink (2016) distributional shapes should be checked as a standardized measure of evaluating imputations. Both MI methods gave identical distributions of densities and frequencies in all variables, indicating that the imputed data was not different from the observed.

Numeric variables were considered for assessment of convergence of their means and standard deviations. All the variables achieved a healthy convergence in both MI methods except the standard deviation of number of casual clients under PMM. This variable had 2.25% missingness and converged well in the model-based FCS. This could mean that the imputation method can affect convergence. In this effect we can conclude that the model-based FCS converged relatively better than PMM. The importance of assessing convergence is well described by Vink (2016); Carpenter and Kenward (2008). Convergence guarantees that the imputed datasets differ sufficiently.

Adjusted ddds ratios of parameter estimates, and corresponding confidence intervals were obtained. The adjusted odds ratios in CCA are obtained from corresponding Wald estimates. Odds ratios follow a log-normal distribution, and so cannot be pooled in the same way

the parameter estimates are in MI. Log transformations are necessary to normalize the estimates, then pooling can be done using Rubin's rules. The pooled estimates are then transformed back to their original log-normal scale. We obtained the interval width by subtracting the lower limit from the upper limit. CCA had the largest interval widths compared to MI, and this can be attributed to the relatively larger standard errors. PMM had a slightly wider interval width in most variables, indicating relatively lower precision compared to model-based FCS method.

Relative efficiency of the two MI methods was quite high >0.95 in all variables. PMM seemed to perform better with 0.99 0.98 in most parts and one 0.95 at number of sex acts. FCS model-based also performed well, with 0.98 0.97 in most parts. This is due to relatively lower between variance in PMM compared to model-based FCS. We conclude that PMM had better relative efficiency compared to the model-based approach. Standard errors of parameter estimates were almost similar for both PMM and the model-based FCS, with PMM showing slight gains with relatively lower standard errors in some parts.

5.1 Strengths and Limitations

Multiple imputation is a flexible method of handling missing data whilst preserving distributional properties and uncertainty of the imputation. These results are not conclusive and may need to be validated through a simulation study.

5.2 Conclusion

Results in this study were based on bio-data obtained from the University of Manitoba's SWOP program on sex workers and other key populations. Our focus was female sex workers. Missing data is a complication to analysis. Naïve approaches that disregard the

proportions, patterns and mechanisms of missingness can lead to biased results. Complete case analysis often introduces bias when missingness is not MCAR. Multiple imputation preserved the distributions in the observed data. Missingness was found to have no impact on the statistical significance of association of risk factors of HIV and HIV positivity among female sex workers. A slight difference in interpretation of results was observed in one variable, but was not statistically significant. Multiple imputation produced unbiased results and performed better than complete case analysis, with much lower standard errors. FCS model-based method showed better performance in convergence and interval width, while, PMM performed better in relative efficiency.

5.3 Future Research

Sensitivity analyses could be conducted to check plausibility of MNAR assumption in this data, as it is impossible to test MAR versus MNAR. A simulation study could also be conducted to validate these results.

References

Amornkul, P. N., Vandenhoudt, H., Nasokho, P., Odhiambo, F., Mwaengo, D., Hightower, A., ... Glynn, J. (2009). HIV prevalence and associated risk factors among individuals aged 13-34 years in Rural Western Kenya. *PloS one*, **4(7)**, e6470.

Audigier, V., White, I. R., Jolani, S., Debray, T. P., Quartagno, M., Carpenter, J., ... Resche-Rigon, M. (2018). Multiple imputation for multilevel data with continuous and binary variables. *Statistical Science*, **33(2)**, 160-183.

AVERT. (2019). HIV/AIDS in Kenya: Kenya 2018. URL https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/kenya.

Beaujean, A. A. (2012). BaylorEdPsych: R package for Baylor University educational psychology quantitative courses. R package version 0.5, URL http://CRAN. R-project. org/package= BaylorEdPsych.

Beksinska, A., Prakash, R., Isac, S., Mohan, H. L., Platt, L., Blanchard, J., ... Beattie, T. S. (2018). Violence experience by perpetrator and associations with HIV/STI risk and infection: a cross-sectional study among female sex workers in Karnataka, south India. *BMJ open*, **8(9)**, e021389.

Bui, T. D., Pham, C. K., Pham, T. H., Hoang, L. T., Nguyen, T. V., Vu, T. Q., Detels, R. (2001). Cross-sectional study of sexual behaviour and knowledge about HIV among urban, rural, and minority residents in Viet Nam. *Bulletin of the World Health Organization*, **79**, 15-21.

Campeau, L., Blouin, K., Leclerc, P., Alary, M., Morissette, C., Blanchette, C., ... Roy, E. (2018). Impact of sex work on risk behaviours and their association with HIV positivity among people who inject drugs in Eastern Central Canada: cross-sectional results from an open cohort study. *BMJ open*, **8(1)**, e019388.

Carpenter, J., Kenward, M. (2008). Brief comments on computational issues with multiple imputation.

Carpenter, J., Kenward, M. (2007). Guidelines for handling missing data in social science research.

Carpenter, J., Kenward, M. (2012). *Multiple imputation and its application*. John Wiley Sons.

Chersich, M. F., Luchters, S. M. F., Malonza, I. M., Mwarogo, P., King'Ola, N., Temmerman, M. (2007). Heavy episodic drinking among Kenyan female sex workers is associated with unsafe sex, sexual violence and sexually transmitted infections. *International journal of STD AIDS*, **18(11)**, 764-769.

Chinomona, A., Mwambi, H. (2015). Multiple imputation for non-response when estimating HIV prevalence using survey data. *BMC public health*, **15(1)**, 1059.

Coetzee, J., Jewkes, R., Gray, G. E. (2017). Cross-sectional study of female sex workers in Soweto, South Africa: factors associated with HIV infection. *PloS one*, **12(10)**, e0184775.

Costa, A. M. (2007). Reducing the harm of drug use and dependence. Retrieved August, 15, 2016.

De Silva, A. P., Moreno-Betancur, M., De Livera, A. M., Lee, K. J., Simpson, J. A. (2019). Multiple imputation methods for handling missing values in a longitudinal categorical

variable with restrictions on transitions over time: a simulation study. *BMC medical research methodology*, **19(1)**, 14.

Harel, O., Pellowski, J., Kalichman, S. (2012). Are we missing the importance of missing values in HIV prevention randomized clinical trials? Review and recommendations. *AIDS and Behavior*, **16(6)**, 1382-1393.

Horton, N. J., and Lipsitz, S. R. (2001). "Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables." American Statistician 55:244–254.

Israel, E., Laudari, C., Simonetti, C. (2008). HIV prevention among vulnerable populations: The Pathfinder International approach.

Jakobsen, J. C., Gluud, C., Wetterslev, J., Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. *BMC medical research methodology*, **17(1)**, 162.

Johnson, D. R., Young, R. (2011). Toward best practices in analyzing datasets with missing data: Comparisons and recommendations. *Journal of Marriage and Family*, **73(5)**, 926-945.

Joint United Nations Programme on HIV/AIDS. (2019). AIDSinfo. Available at: aidsinfo. unaids. org. Accessed, 31.

Lee, K. J., Carlin, J. B. (2010). Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. American journal of epidemiology, 171(5), 624-632.

Little, R. J., Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley Sons.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., Verbeke, G. (2014). *Handbook of missing data methodology*. Chapman and Hall/CRC.

Muraguri, N. (Ed.). (2010). National guidelines for HIV/STI programs for sex workers. Ministry of Public Health and Sanitation.

Nuwasiima, Afra. (2018). Multiple Imputation and Random Survival Forests: Application to the Demographic and Health Survey Child Survival Data. 10.13140/RG.2.2.30983.85929.

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical epidemiology*, **9**, 157.

Prantner, B. (2011). Visualization of imputed values using the R-package VIM.

Raghunathan, T. (2015). *Missing data analysis in practice*. Chapman and Hall/CRC.

Raghunathan, T., Berglund, P. A., Solenberger, P. W. (2018). *Multiple Imputation in Practice: With Examples Using IVEware*. Chapman and Hall/CRC.

Rombach, I., Gray, A. M., Jenkinson, C., Murray, D. W., Rivero-Arias, O. (2018). Multiple imputation for patient reported outcome measures in randomised controlled trials: advantages and disadvantages of imputing at the item, subscale or composite score level. *BMC medical research methodology*, **18(1)**, 87.

SAS Institute. (2017). SAS/STAT® user's guide version 14.3, The MI Procedure.

Schafer, J. L., Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, **7(2)**, 147.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, **338**, b2393.

Templ, M., Alfons, A., Kowarik, A., Prantner, B. (2019). VIM: Visualization and Imputation of Missing Values, 2011a. URL http://CRAN. R-project. org/package= VIM. R package version, 3(0).

UNAIDS, G. H., Statistics, A. I. D. S. (2019). 2018 Fact Sheet, 2019.

Van Buuren, S. (2012). *Flexible imputation of missing data*. Chapman and Hall/CRC.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, **16(3)**, 219-242.

Vink, G. (2016). Towards a standardized evaluation of multiple imputation routines.

Welch, C. A. (2015). Implementation, evaluation and application of multiple imputation for missing data in longitudinal electronic health record research (Doctoral dissertation, UCL (University College London)).

World Health Organization. (2018). Global health estimates 2016: disease burden by cause, age, sex, by country and by region, 2000–2016. Geneva: World Health Organization.

Yuan, Y. C. (2010). Multiple imputation for missing data: Concepts and new development (Version 9.0). *SAS Institute Inc, Rockville, MD*, **49(1-11)**, 12.