Research Article

2023; Issue 6
ISSN 2591–6769
eISSN 2953–2663

East African Journal of Applied Health Monitoring and Evaluation

USING MULTIPLE IMPUTATION AND INVERSE PROBABILITY WEIGHTING TO ADJUST FOR MISSING DATA IN HIV PREVALENCE ESTIMATES: A CROSS-SECTIONAL STUDY IN MWANZA, NORTH WESTERN TANZANIA

TINASHE MHIKE [1]; JIM TODD [3 4]; MARK URASSA[3]; & NEEMA R. MOSHA[1 2]

1. Division of Epidemiology and Biostatistics, Faculty of Medicine and Health Sciences, Stellenbosch University, P.O. Box 241, Francie van Zijl Drive, 7505 Tygerberg, Cape Town-South Africa
2. Mwanza Intervention Trials Unit, P.O. Box 11936 Isamilo Road, Mwanza-Tanzania
3. National Institute for Medical Research, Mwanza Centre P.O. Box 1462 Isamilo Road, Mwanza-Tanzania
4. London School of Hygiene and Tropical Medicine, Keppel St, Bloomsbury, London WC1E 7HT, United Kingdom.

Correspondence to: Tinashe Mhike, Division of Epidemiology and Biostatistics, Faculty of Medicine and Health Sciences, Stellenbosch University, P.O. Box 241, Francie van Zijl Drive, 7505 Tygerberg, Cape Town-South Africa, E-mail: tinashemike86@gmail.com

Received: August 25, 2022 Accepted October 18, 2023 Published October 22, 2023

Introduction
Population surveys and demographic studies are the gold standard for estimating HIV prevalence. However, non-response in these surveys is of major concern, especially if it is not random and complete case analysis becomes an inappropriate data analysis method. Therefore, a comprehensive analysis that will account for the missing data must be used to obtain unbiased HIV prevalence estimates.

Methods
Serological samples were collected from participants who were residents of a Demographic Surveillance System (DSS) in Kisesa, Tanzania. HIV prevalence was estimated using three methods. Firstly, using the Complete case analysis (CCA), assuming data were Missing Completely at Random (MCAR). The other two methods, multiple imputations (MI) and inverse probability weighting (IPW) assumed that non-response was missing at random (MAR). For MI, a logistic regression model adjusting for age, sex, residence, and marital status was used to impute 20 datasets to re-estimate the HIV prevalence. The propensity for participating in the sero-survey and being tested for HIV given age, sex, residence, and marital status were generated using logistic regression models. Using the propensity scores, inverse probability weights were derived for participants who were tested for HIV.

Results
The overall CCA HIV prevalence estimate was 6.6% (95% CI: 6.0-7.2), with 5.4% (95% CI: 4.6-6.3) in males and 7.3% (95% CI: 6.6-8.1) in females. Using MI, the overall HIV prevalence was 6.8% (95% CI: 6.2-7.5), 6.2% (95% CI: 5.1-7.3) in males, and 7.4% (95% CI: 6.6-8.2) in females. Using IPW the overall HIV prevalence was 6.7% (95% CI: 6.1-7.4), with 5.5% (95% CI: 4.7-6.5) in males and 7.7% (95% CI: 7.0 - 8.6) in females. HIV prevalence differed significantly between age groups ($p<0.001$), with the highest estimate in males aged 35-39 and females aged 40-44, and the lowest in both males and females aged 15-19 years.

Conclusion
Complete case analysis underestimates HIV prevalence compared to methods that adjust for missing data. After comparing CCA, MI, and IPW, we found out that the best method to adjust for missing data in population surveys is through the use of multiple imputations.

Keywords: *Complete case analysis, Multiple imputation, Nonresponse, Propensity score, Inverse Probability Weighting*

Research Article

2023; Issue 6
ISSN 2591–6769
eISSN 2953–2663

East African Journal of Applied Health Monitoring and Evaluation

INTRODUCTION

Prevalence measures the burden of disease in a population in a given location and at a particular time, representing the proportion of people affected by the disease (1). Estimates of HIV prevalence are frequently used to monitor and study the determinants of the HIV epidemic, identify groups at high risk of HIV infection, and assess the need for HIV prevention and treatment (2).

Population surveys and demographic studies have become the gold standard for estimating national HIV prevalence (3). However, non-response in these surveys is of major concern (4). Individuals may not participate because the interviewers could not contact them for an interview or they refuse to give consent to an HIV test (4). Non-response can bias population-based estimates of HIV prevalence if non-response is associated with HIV status in any way. This could occur for two reasons namely refusal to participate in HIV testing because the individual knows his/her status or an individual is involved in high sexual risk behavior (5).

Missing data in research can be classified into three types: one, data missing completely at random (MCAR), which means that missingness is independent of the outcomes and any other observed or unobserved characteristics; two, data missing at random (MAR), that is missingness can be dependent on observed covariates but is independent of the unobserved data and thirdly, data missing not at random (MNAR), that is data are neither missing completely at random nor missing at random. When missing data depends on both the observed and unobserved data, they are considered MNAR (6).

In the population based HIV studies, data can be assumed to be MCAR if the patient gave a blood sample, but the sample was destroyed before it was tested such that the missingness is not associated with their HIV status or any other observed covariate (7). If, however, a patient misses a test, because he had a long way to walk, then data would be MAR, because although missingness is not directly related to their HIV status, it may be related to their residence or other observed covariates, which may, in turn, be associated with the HIV status (8). And finally, MNAR is when an eligible study participant does not come or consent for testing because they already know their HIV status or they have a high probability of being HIV positive or belong to high-risk groups. Here, the missingness depends on the missing HIV status, in which case the MAR assumption is violated. Such mechanism data are considered missing not at random (MNAR) or non-ignorable (9).

When observations are missing completely at random, the missing observations are a random subset of all observations; the missing and observed values will have similar distributions and produce unbiased estimates. However, if observations are MAR there might be systematic differences between the missingness and observed values, but these can be entirely explained by other observed variables. For example, if HIV status is missing at random, conditional on age, sex, residence, and marital status, then the distributions of the missing and observed HIV status will be similar among people of the same age, sex, residence, and marital status (10). However, if observations are MNAR even after conditioning

on the observed covariates, the distributions will differ and any estimates maybe biased (11).

Most researchers use conventional methods such as the complete case or available case analysis where the assumption is data are MCAR. The use of these methods in the presence of missing data that are not MCAR results in loss of information and biased estimates of HIV prevalence (12). There has been development of statistical methods that can be applied to adjust for missing data when the missingness is not completely at random. Methods such as inverse probability weighting (IPW), maximum likelihood estimation, multiple imputations, and double robust methods can produce less biased estimates.

The IPW methods rely on the intuitive idea of creating a pseudo-population of weighted copies of the complete cases to remove selection bias introduced by the missing data. However, different weighting approaches are required depending on the missing data pattern and mechanism (13). Maximum likelihood estimation and multiple imputations (MI) are the other methods used to adjust for missing data (14). In MI, missing data are replaced by data drawn from an imputation model. This is done M times, generating M complete datasets. Each generated data is analyzed and an estimate of the model parameters is calculated (15). The overall estimate is simply the average of the M estimates and the standard errors of the estimates are obtained using Rubin's rules (8).

However, in surveys for HIV prevalence, the application of these statistical methods is rare due to their complexity, the extra time needed for the analysis and the availability of software. Depending on the pattern and mechanism of the missingness, some techniques are superior than others.

The objective of this study was to determine the effect of missing data on the estimates of HIV prevalence from a population survey in Tanzania, using complete case analysis, multiple imputation (MI) and inverse probability weighting (IPW).

METHODS

Data Source

Data were obtained from Kisesa observation HIV cohort study in Magu District, Mwanza Region, Northwestern Tanzania. This cohort is located within a Health and Demographic Surveillance System (HDSS) which had the baseline census in 1994 and then regular household visits to record all births, deaths and migration. Currently there are 34 completed rounds of HDSS (16). HIV and other infectious

diseases are monitored in the cohort using a series of epidemiological serological surveys to measure the HIV status of residents at three-year intervals from 1994 to 2016, and currently there are 8 completed serological surveys.

This study used data from HDSS round 30 (2015) and sero-survey round 8 (sero8) implemented during 2015/2016. All residents (aged 15 years and above) from Kisesa HDSS round 30 were eligible to take part in sero8. Participants were invited through invitation slips, informing them about the location of the temporary clinic and their date of participation. At the clinic, all participants were requested for their written consent to participate in the survey and testing for HIV. Consents for the minors (under the age of 18 years) were obtained at home from parents or guardians and assent provided by the minor at the clinic. During the sero8 operations, participants were interviewed using a structured questionnaire to report on their socio-demographic characteristics. Blood samples were collected through finger prick and tested for HIV antibodies using Alere Determine™ HIV-1/2 rapid test for screening and Trinity Biotech Uni-Gold™ HIV rapid test for confirmation.

Statistical methods

The outcome of interest was HIV status (positive/negative) with HIV prevalence estimated using three methods: Complete case analysis on the sero8 survey data alone assuming HIV status through non-attendance at the survey, to be missing completely at random (MCAR); Multiple imputation (MI) and inverse probability weighting (IPW) methods, which assumed data to be missing at random (MAR), with attendance at the survey dependent on age, gender, residence and marital status.

In the complete case analysis, all participants with missing HIV status or missing any of the covariates were excluded from the analysis. Participants who had missing HIV status were treated as a random subset of the complete sample of subjects, and, the set of participants with no missing HIV status were also treated as a random sample from the source population (7). This approach can only result in unbiased estimates when it is demonstrable that missing data are not associated with HIV status in any way (17).

Multiple imputations (MI) involved imputing values for the missing HIV status, for those who did not attend the sero8 survey, based on age, sex, residence and marital status (12). We imputed 20 datasets (M=20) using the Markov Chain Monte Carlo (MCMC) algorithm with a binomial distribution replacing each missing HIV value with values consistent with that person's age, sex, residence and marital status. After imputation, each dataset was used to estimate the HIV prevalence using logistic regression. The 20 estimates of HIV prevalence were averaged to come up with a pooled estimate. The Rubin's rules were used to combine the average standard error and obtain the 95% confidence interval for the pooled estimate (18).

For IPW, we first used a logistic regression model to estimate the propensity scores for participating in the sero-survey and being tested for HIV given age, sex, residence and marital status as the covariates. Propensity scores (PS) obtained from the models balanced the distribution of observed baseline covariates for those tested for HIV and those not tested. Using the propensity scores, p(x), we derived inverse probability weights (IPW) for participants who were tested for HIV. The inverse probability weights were normalized to reflect the age, sex, residence and marital status of the HDSS population, and the HIV prevalence was estimated using the normalized inverse probability weights.

Ethics approval and consent to participate

At the clinic, all participants were requested for their written consent to participate in the survey and testing for HIV. Consents for the minors (under the age of 18 years) were obtained at home from parents or guardians and assent provided by the minor at the clinic.

RESULTS

Description of the study participants

Figure 1 shows that a total of 21857 participants aged 15 years or older were resident in the cohort, 19985 (91%) were seen in the HDSS survey, 7490 (34%) enrolled in the sero8 survey with 5618 (26%) seen in both HDSS and sero8. The 1872 (9%) participants not in HDSS were new residents, had moved into the area after the HDSS survey. More than 70% of the eligible participants did not attend the corresponding sero-survey, hence missing the HIV status (Figure 1). A flow diagram below shows the enrollment of the study participants.

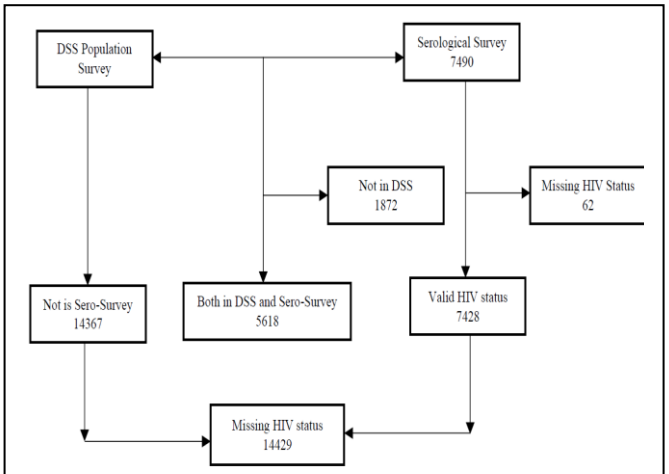

Figure 1: Flow chart of participants aged 15 years and above enrolled in the study

Study characteristics of the participants

In this population aged 15 years and above, there were 10,150 (46%) males and 11,706 (54%) females, with a 10,755(49%) married participants compared to 7,543 (36%) who were single and 2,829 (13%) who were separated or widowed. For areas of residence, overall, there were 11,274 (52%) from rural areas and 10,578 (48%) from urban areas. A larger percentage of the participants, 4,752 (22%) in this study were in the 15-19 age group, with the lowest number of participants, 779 (4%) in the 55-59 age category. There were differences in the proportions in these categories between those who attended sero8 and those who were seen in the HDSS (Table 1).

Table 1: Distribution of the characteristics of the participants included in the study (N=21857)

Characteristics	DSS and Sero-survey n (%)	Serological survey only n (%)	DSS only n (%)	Total
Overall	5618 (26)	1872 (9)	14367 (66)	21857
Sex				
Female	3483 (30)	1203 (10)	7020 (60)	11706
Male	2135 (21)	669 (7)	7346 (72)	10150
Age group				
15 - 19	1264 (27)	488 (10)	3000 (63)	4752
20 - 24	665 (19)	312 (9)	2549 (72)	3526
25 - 29	525 (22)	225 (9)	1688 (69)	2438
30 - 34	427 (19)	230 (10)	1590 (71)	2247
35 - 39	475 (24)	176 (9)	1337 (67)	1988
40 - 44	441 (24)	130 (7)	1287 (69)	1858
45 - 49	344 (28.2)	79 (6.5)	796 (65.3)	1219
50 - 54	384 (35)	66 (6)	643 (59)	1093
55 - 59	239 (31)	42 (5)	498 (64)	779
60 and above	854 (44)	124 (6)	979 (50)	1957
Marital status				
Single	1691 (22)	669 (9)	5183 (69)	7543
Married	3009 (28)	889 (8)	6857 (64)	10755
Separated/Widowed	917 (32)	298 (11)	1614 (57)	2829
Missing	1 (0.1)	16 (2.1)	713 (97.8)	730
Residence				
Urban	2214 (21)	879 (8)	7485 (71)	10578
Rural	3404 (30.2)	990 (8.8)	6880 (61)	11274

HIV prevalence – A complete case analysis

Figure 2 shows the HIV prevalence and 95% CI estimates by sex and age groups for those who attended the sero8 survey. In all age groups, except for the 35-39 age group, females had a higher HIV prevalence than males.

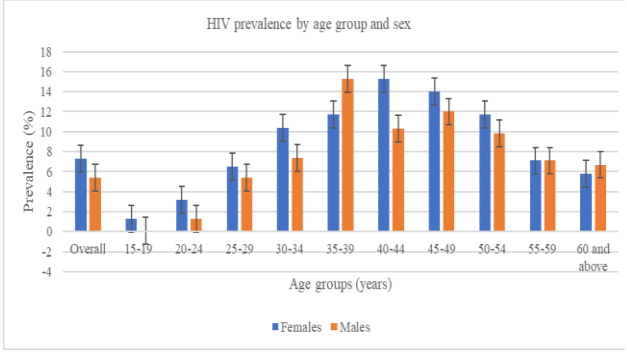

Figure 2: Prevalence of HIV estimates and their 95% confidence intervals by age group and sex among study participants (N=7428)

Using the complete case analysis, the overall HIV prevalence was 6.6% (95% CI: 6.0-7.2), with higher estimate in females (7.3%, 95% CI: 6.6-8.1) than males (5.4%, 95% CI: 4.6-6.3). HIV prevalence differed by age from 0.7% (95% CI: 0.4-1.3) in the 15-19 age group to 13.4% (95% CI: 10.9 – 16.5) in those aged 40-44 years (p<0.001). The HIV prevalence was similar for rural and urban residents (p=0.38), but there was a higher HIV prevalence among the separated/widowed group (12.4%, 95% CI: 10.6 – 14.4) and a lower prevalence among those who were single (2.0%, 95% CI: 1.5 – 2.6) (Table 2).

Table 2: Prevalence of HIV among participants who attended sero8 survey (complete case analysis)

Characteristics	N	HIV positive	95% CI	P-value
Overall	7428	488(6.6)	6.0 - 7.2	
Sex				
Female	4652	339(7.3)	6.6 - 8.1	0.001
Male	2776	149(5.4)	4.6 - 6.3	
Age group				
15 - 19	1750	13(0.7)	0.4 - 1.3	< 0.001
20 - 24	973	24(2.5)	1.7 - 3.7	
25 - 29	746	46(6.2)	4.6 - 8.1	
30 - 34	654	63(9.6)	7.6 - 12.1	
35 - 39	645	83(12.9)	10.5 - 15.7	
40 - 44	566	76(13.4)	10.9 - 16.5	
45 - 49	421	56(13.3)	10.4 - 16.9	
50 - 54	447	49(11)	8.4 - 14.2	
55 - 59	280	20(7.1)	4.7 - 10.8	
60 and above	946	58(6.1)	4.8 - 7.9	
Marital status				
Single	2355	47(2.0)	1.5 - 2.6	< 0.001
Married	3874	293(7.6)	6.8 - 8.4	
Separated/Widowed	1189	147(12.4)	10.6 - 14.4	
Residence				
Urban	3055	210(6.9)	6.0 - 7.8	0.381
Rural	4370	278(6.4)	5.7 - 7.1	

Missing data description

In this study, 14429 (66%) participants did not have valid HIV tests. Also, 730 (3%) participants had missing marital

status (Table 3). One person had no sex recorded whilst 5 had missing residence records.

Table 3: Frequencies and percentage of missing data (N=21857)

Characteristics	Frequency of missing values	Percentage of missing values
HIV status	14429	66
Sex	1	0
Age	0	0
Marital status	730	3
Residence village	5	0

Multiple imputations (MI) and inverse probability weighting (IPW) HIV prevalence estimates

All individuals (N=736) with any missing covariates were dropped from the analysis. The overall HIV prevalence estimate was 6.8% (95% CI: 6.2-7.5) using MI and 6.7% (95% CI: 6.1-7.4) using IPW. HIV prevalence was estimated separately (stratified) for each sex. For males the overall HIV prevalence was 5.4% (95% CI: 4.6 – 6.3) under CCA, 6.2%, (95% CI: 5.1-7.3) under MI, and 5.5%, (95% CI: 4.7-6.5) under IPW (Table 4).

Table 4: HIV prevalence estimates using complete case analysis, multiple imputations, and inverse probability weighting for males (N=9797)

Characteristics	Complete case	Multiple imputations	IPW
		Prevalence (95% CI)	
Overall	5.4 (4.6; 6.3)	6.2 (5.1; 7.3)	5.5 (4.7, 6.5)
Age group			
15 - 19	0.1 (0.02; 0.9)	0.5 (0.1; 1.1)	0.1 (0.02; 1.0)
20 - 24	1.3 (0.6; 3.2)	2.3 (0.6; 4.0)	1.3 (0.5; 3.2)
25 - 29	5.4 (3.1; 9.2)	6.1 (3.4; 8.7)	5.0 (2.8; 8.8)
30 - 34	7.5 (4.4; 12.5)	8.5 (5.3; 11.8)	8.1 (4.6; 13.9)
35 - 39	15.4 (11.0;21.2)	13.1 (9.4; 16.9)	14.6 (10.4; 20.3)
40 - 44	10.3 (6.9;15.1)	11.4 (7.6; 15.3)	10.7 (7.0; 15.9)
45 - 49	12(7.7; 18.3)	10.9 (6.7; 15.1)	11.2 (7.1; 17.2)
50 - 54	9.8 (6.1; 15.4)	9.6 (5.1; 14.0)	10.4 (6.4; 16.5)
55 - 59	7.1 (3.4; 14.3)	6.1 (3.0; 9.3)	7.0 (3.4; 14.1)
60 and above	6.7 (4.5; 9.8)	5.7 (3.7; 7.8)	6.7 (4.5; 10.0)
Marital Status			
Single	1.1 (0.7; 1.9)	2.5 (1.5; 3.6)	1.8 (1.1; 3.1)
Married	8.5 (7.1; 10.1)	8.5 (6.8; 10.3)	8.7 (7.3; 10.4)
Separated/Widowed	12.1 (7.7;18.4)	12.8 (8.3; 17.3)	13.3 (8.4; 20.3)
Residence			
Urban	5.2 (4.2; 6.3)	5.9(4.9; 6.8)	5.2 (4.3; 6.4)
Rural	5.7 (4.4; 7.3)	6.5 (4.9; 8.1)	5.9 (4.5; 7.6)

The overall HIV prevalence for females was 7.3% (95% CI: 6.6 – 8.1) under CCA, 7.4%, (95% CI: 6.6-8.2) under MI, and 7.7%, (95% CI: 7.0-8.6) under IPW (Table 5).

Table 5: HIV prevalence estimates using complete case analysis, multiple imputations, and inverse probability weighting for females (N=11,324)

Characteristics	Complete case	Multiple imputations	Propensity scores
		Prevalence (95% CI)	
Overall	7.3 (6.6; 8.1)	7.4 (6.6; 8.2)	7.7 (7.0; 8.6)
Age group			
15 - 19	1.3 (0.7; 2.3)	1.1 (0.4; 1.7)	1.2 (0.7; 2.0)
20 - 24	3.2 (2.0; 4.9)	2.9 (1.1; 4.6)	3.2 (2.0; 5.1)
25 - 29	6.5 (4.7; 9.0)	6.6 (4.6; 8.6)	6.8 (4.8; 9.4)
30 - 34	10.4 (8.0;13.5)	10.5 (7.9;13.2)	10.8 (8.2; 14.1)
35 - 39	11.8 (9.1;15.2)	12.9 (9.4; 16.3)	11.9 (9.2; 15.4)
40 - 44	15.3 (11.9;19.5)	14.7 (10.6; 18.8)	15.4 (11.9; 19.6)
45 - 49	13.7 (10.1; 18.4)	13.4 (9.4; 17.5)	13.9 (10.2; 18.7)
50 - 54	11.7 (8.4; 16.0)	12.4 (8.4; 16.4)	11.9 (8.6; 16.3)
55 - 59	7.1 (4.2; 11.9)	7.6 (3.6; 11.6)	7.4 (4.3; 12.4)
60 and above	5.8 (4.2; 8.0)	6.7 (4.6; 8.8)	5.8 (4.2; 8.0)
Marital Status			
Single	3 (2.1; 4.2)	2.9 (1.8; 4.0)	4.3 (3.0; 6.1)
Married	7 (6.1; 8.1)	7.7 (6.7; 8.6)	7.2 (6.2; 8.3)
Separated/Widowed	12.4 (10.5; 14.6)	13.3 (11.3; 15.3)	13.5 (11.4; 15.8)
Residence			
Urban	7.2 (6.2; 8.2)	7.4 (6.2; 8.6)	7.7 (6.7; 8.8)
Rural	7.4 (6.4; 8.6)	7.4 (6.4; 8.5)	7.8 (6.7; 9.1)

Overall, in males the estimated HIV prevalence under IPW (5.5%, 95% CI: 4.7, 6.5) was similar to CCA (5.4%, 95% CI: 4.6, 6.3), while the MI estimate was higher (6.2%, 95% CI: 5.1, 7.3). In younger males (15-19 years and 20-24 years), the estimated HIV prevalence was similar using IPW and CCA methods but was greater under MI (Table 4). IPW and CCA estimates were similar in older males, but the MI estimates were lower in younger males, aged 15-19, compared to older age groups. The increased HIV prevalence estimate was greater in single males under MI, and in separated/widowed males under IPW.

The overall HIV prevalence among women under the CCA was 7.3, (95% CI: 6.6-8.1), 7.4 (95% CI: 6.6-8.2) under MI

and 7.7% (95% CI: 7.0-8.6) under IPW method. The 15-19 age group had the lowest HIV prevalence under the three approaches: 1.1% (95% CI: 0.4-1.7) under the MI method, 1.3% (95% CI: 0.7-2.3) for the complete case analysis and 1.2% (95% CI: 0.7-2.0) under the IPW method. The separated/widowed had the highest HIV prevalence under the three approaches, 12.4% (95% CI: 10.5-14.6) for the CCA, 13.3% (95% CI: 11.3-15.3) under the MI approach and 13.5% (95% CI: 11.4-15.8) for the IPW approach. Single participants had the least HIV prevalence under the three approaches, 3.0% (95% CI: 2.1-4.2) for the complete case, 2.9% (95% CI: 1.8- 4.0) under the MI approach and 4.3% (95% CI: 3.0-6.1) for the IPW approaches. Females residing in the rural areas had the least HIV prevalence under the complete case analysis 7.2%, (95% CI: 6.2-8.2) compared to 7.4% (95% CI: 6.4-8.6) for their counterparts in the urban areas. However, under the MI approach, females residing in the rural areas had similar HIV prevalence as urban ones, 7.4% (95% CI: 6.2-8.6) and 7.4% (95% CI: 6.4-8.5) for the urban counterparts. Using the IPW approach, HIV prevalence for rural and urban residents was very close i.e. 7.7% (95% CI: 6.7-8.8) in the rural areas compared to 7.8% (95% CI: 6.7-9.1) in the urban areas (Table 5).

Generally, tables 4 and 5 showed that HIV prevalence increased with an increase in age, from the minimum age group to 35-39 for males and 40-44 for females when it started to decrease. Those who were separated or widowed had the highest HIV prevalence with the lowest HIV prevalence amongst the single never married participants. Estimating HIV prevalence by residence had similar estimates for all the three methods.

There was an increase in HIV prevalence estimates after adjusting for missing data using multiple imputations and inverse probability weighting methods. The estimates obtained using multiple imputations were slightly larger than those obtained using inverse probability weighting and the 95% confidence intervals for MI were narrower than those obtained using IPW and CCA for both sexes. The age and sex pattern for HIV prevalence was similar for MI and inverse probability weighting methods. The separated/widowed participants had the highest HIV prevalence. Urban residence had a higher HIV prevalence than rural residents but the difference was not statistically significant using the three approaches.

DISCUSSION

This study compared three methods of analysis that adjust for missing data in HIV surveys. The overall HIV prevalence using the complete case analysis was 6.6 (95% CI: 6.0-7.2), 6.8 (95% CI: 6.2-7.5) using MI and 6.7 (95% CI: 6.1-7.4) using the inverse probability weighting (IPW) method.

In this study, females had a higher HIV prevalence than males using the three approaches, that is, more females were HIV positive than males, with the lowest estimates among participants aged 15-19 years which maybe because most of these participants were of school going age, not yet married, and may not have had sexual debut (20). Participants between 25-59 years had high HIV prevalence as most of them are sexually active and have multiple partners. The lower HIV prevalence among those aged 60 years and above was a result of potentially lower sexual activities in the group (21).

The separated or widowed participants had the highest HIV prevalence, as some may have had partners infected with HIV who have died or divorced (20). Single participants had the lowest HIV prevalence under CCA, MI and IPW methods, as many were young and not involved in sexual relationships. Variations in HIV prevalence were also a result of place of residence. Urban residents had high HIV prevalence than the rural residents but the difference was not statistically significant (p=0.38). The insignificant difference between the HIV prevalence between rural and urban residents could be explained by the fact that the entire area of Kisesa is becoming more urbanized and access to rural areas has increased a lot in the recent times.

We found that there were minor differences in HIV prevalence estimates obtained using each of the methods i.e. complete case analysis, multiple imputation and IPW. However, in some specific groups MI and IPW produced narrower confidence interval estimates. The complete case analysis method ignores the missing data hence can underestimate the HIV prevalence. A systematic review which looked at the analytical methods used in estimating the prevalence of HIV/AIDS from demographic and cross-sectional surveys with missing data recommended the use of advanced methods to adjust for missing data in the analysis of HIV survey data to reduce bias in the estimates. Failure to adjust for missing data may result in biased estimates of parameters of interest (22).

The HIV prevalence estimated using the methods that assumed the missingness was MAR were 2-3% higher than the complete case analysis which assumed MCAR. Thus, the assumption of MCAR gave a biased estimate of the HIV prevalence, which concurs with the conclusions of a systematic review of missing data in HIV prevalence estimation (22). Our results were consistent with Mwambi and Chinomona who found that the prevalence of HIV was underestimated by complete case analysis, with the conclusion that multiple imputation provided a more accurate estimation of the HIV prevalence in the presence of missing data (20). In another analysis using multiple imputations, complete case analysis provided inefficient though valid results when missing data are MCAR, but biased results when data were MAR. Multiple imputation approach led to unbiased results with correct standard errors, in situations where data were MCAR or MAR (7). A simulation study indicated that it's not advisable to use complete case analysis especially if the proportion of missing values is high (23). With IPW, assuming no model misspecification, the prevalence estimates are corrected from the bias introduced by CCA analysis irrespective of the sample size as the standard errors are larger compared to IPW. (24).

Multiple imputation generally had the highest HIV prevalence estimates in most of the covariates, and the 95% confidence intervals were narrower than the complete case and the IPW methods. This reflects the effects of the extra precision the MI introduces in the estimation process (20). The 95% confidence intervals for CCA and IPW were similar because IPW and CCA are restricted to the sample who were tested for HIV, and the only difference was that using IPW we weighted the

estimates in respect of their covariates observed in calculating the prevalence. In contrast the MI method imputed data for the missing HIV status, and the extra information made the standard errors smaller resulting in narrower 95% confidence intervals which were more precise.

CONCLUSION

Estimating HIV prevalence from population and survey data is prone to bias when the assumptions about missing data are incorrect. Robust statistical methods have to be employed in order to properly account for missing data. Both multiple imputation and IPW are able to account for missing data. The results of this study showed that multiple imputation (MI) is a reliable method for estimating HIV prevalence in the presence of missing data. This method was more superior to the complete case and the IPW approaches as it did not underestimate HIV prevalence and had tighter 95% confidence intervals. Therefore, in the presence of missing data, we recommend the use of MI in estimating HIV prevalence to address the problem of varied types of missing data. Thus, based on the MI estimations, overall HIV prevalence in Kisesa was 6.8% and higher among females with 7.4% (95% CI: 6.6-8.2) than males with 6.2% (95% CI: 5.1-7.3).

STUDY LIMITATIONS

The potential limitation of this study is the use of secondary data. Researchers had no control on what was contained in the dataset and on how well the data was collected.

LIST OF ABBREVIATIONS

CCA – Complete Case Analysis; DSS - Demographic Surveillance System; HDSS - Health and Demographic Surveillance System; HIV – Human immunodeficiency Virus; IPW – Inverse Probability weighting; KOC – Kisesa Observational Cohort; MAR – Missing at Random; MCAR – Missing Completely at Random; MCMC – Monte Carlo Markov Chain; MI – Multiple imputation; MNAR – Missing Not at Random; PS – Propensity Score

ACKNOWLEDGEMENTS

I would like to thank the participants of the TAZAMA Project (NIMR Mwanza) studies who contributed the data for this project. This work was undertaken as part of the MSc in Biostatistics in the University of Stellenbosch by TM. I would also like to thank the National Institute of Health (NIH) for sponsoring my MSc in Biostatistics.

AUTHORS' CONTRIBUTIONS

TM, JT and NM formulated the research question and developed the study protocol. MU prepared the data. TM, JT and NM performed the analysis and interpretation of the results. All the authors contributed to the manuscript development and approved the final version for submission to be published.

CONFLICT OF INTEREST

The authors declare that there are no competing interests.

AVAILABILITY OF DATA AND MATERIALS

The datasets used and/or analysed during the current study are available from the corresponding author on a reasonable request. Data will also be available in the London School of Hygiene and Tropical Medicine repository.

FUNDING

The MSc studies of TM were supported by the Fogarty International Centre of the National Institutes of Health under Award Number D43 TW010547. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

1. Ward MM. ADMINISTRATIVE DATA : SOME ASSEMBLY REQUIRED. 2014;40(8):1241–3.
2. Bärnighausen T, Bor J, Wandira-Kazibwe S, Canning D. Correcting HIV prevalence estimates for survey nonparticipation using heckman-type selection models. Epidemiology. 2011;22(1):27–35.
3. Boerma JT, Ghys PD, Walker N. Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard. 2003;362:1929–31.
4. McGovern ME, Marra G, Radice R, Canning D, Newell ML, Bärnighausen T. Adjusting HIV prevalence estimates for non-participation: N application to demographic surveillance. J Int AIDS Soc. 2015;18(1):1–11.
5. Marston M, Harriss K, Slaymaker E. Non-response bias in estimates of HIV prevalence due to the mobility of absentees in national population-based surveys : a study of nine national surveys. 2008;84(Suppl I):71–7.
6. Li C. Little ' s Test of Missing Completely at Random. 1988;(Little):1–15.
7. Donders ART, Heijden GJMG Van Der, Stijnen T, Moons KGM. Review : A gentle introduction to imputation of missing values. 2006;59:1087–91.
8. Schafer JL, Graham JW. Missing Data : Our View of the State of the Art. 2002;7(2):147–77.
9. Kang S, Little RJ, Kaciroti N. Missing not at random models for masked clinical trials with dropouts. 2015;(January).
10. Bhaskaran K, Smeeth L. Education corner What is the difference between missing completely at random and missing at random ? 2014;(April):1336–9.
11. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. & reporting Multiple

imputation for missing data in epidemiological and clinical research : potential and pitfalls.

12. Chinomona A, Mwambi HG. Estimating HIV Prevalence in Zimbabwe Using Population-Based Survey Data. 2015;1–17.
13. Robins JM. NIH Public Access. 2014;22(1):14–30.
14. Allison PD, Horizons S. SAS Global Forum 2012 Statistics and Data Analysis Handling Missing Data by Maximum Likelihood SAS Global Forum 2012 Statistics and Data Analysis. 2012;1–21.
15. Seaman SR, White IR, Copas AJ, Li L, Unit MRCB. Combining Multiple Imputation and Inverse-Probability Weighting Combining Multiple Imputation and Inverse-Probability Weighting. 2014;(November 2011).
16. Kishamawe C, Isingo R, Mtenga B, Zaba B, Todd J, Clark B, et al. Health & Demographic Surveillance System Profile Health & Demographic Surveillance System Profile : The Magu Health and Demographic Surveillance System (Magu HDSS). 2015;(September):1851–61.
17. Malla L, Perera-salazar R, Mcfadden E, Ogero M. Handling missing data in propensity score estimation in comparative effectiveness evaluations : a systematic review. 2018;7:271–9.
18. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation : current practice and guidelines. 2009;8:1–8.
19. Kang H. The prevention and handling of the missing data. 2014;(January).
20. Chinomona A, Mwambi H. Multiple imputation for non-response when estimating HIV prevalence using survey data Biostatistics and methods. BMC Public Health [Internet]. 2015;15(1):1–10. Available from: http://dx.doi.org/10.1186/s12889-015-2390-1
21. Wing EJ. The Aging Population with HIV Infection. 2017;128(2).
22. Mosha NR, Aluko OS, Todd J, Machekano R, Young T. Analytical methods used in estimating the prevalence of HIV / AIDS from demographic and cross-sectional surveys with missing data : a systematic review. 2020;1–10.
23. Mayer B, Puschner B. Propensity score adjustment of a treatment effect with missing data in psychiatric health services research. 2015;12(1):1–7.
24. Data R. HHS Public Access. 2019;113(521):369–79.