Change in Knowledge and Attitudes about HIV/AIDS in Sub-Saharan Africa: An Analysis of National Survey Data

Xiaochen Dai

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Simon Iain Hay, Chair

Haidong Wang

Jared Baeten

Kwun Chuen (Gary) Chan

Geoffrey Garnett

Program Authorized to Offer Degree:

Global Health

University of Washington

Abstract

Change in Knowledge and Attitudes about HIV/AIDS in Sub-Saharan Africa: An Analysis of National Survey Data

Xiaochen Dai

Chair of the supervisory committee:

Simon Iain Hay

Department of Global Health

This dissertation explores the changes in knowledge and attitudes about HIV/AIDS in sub-Saharan Africa (SSA) over time. Specifically, this work estimates the trends of 16 key indicators of knowledge and attitudes about HIV/AIDS in 47 SSA countries from 1988 to 2017 using existing national survey data (*Aim 1*), evaluates the performance of different multiple imputation methods in imputing country-level proportions of key indicators that are missing in the time-series-cross-sectional (TSCS) surveys (*Aim 2*), and calculates the composite scores of HIV/AIDS knowledge and attitudes in the 47 countries from 1998 to 2017 after imputing the country-level proportions of key indicators that are missing in the surveys using *Amelia* (*Aim 3*).

In *Aim 1*, we found that most key indicators of knowledge and attitudes about HIV/AIDS had improved over time across board although the level of improvements were heterogeneous across countries. However, two indicators, namely, people's knowledge of HIV transmission from mother to child during pregnancy, and people's attitudes toward disclosing family members' HIV/AIDS status, had deteriorated over time in many countries. We also found that men in SSA

generally had better knowledge and attitudes about HIV/AIDS than women did, except for the indicators of mother-to-child transmission (MTCT), of which women had surpassed men and had better knowledge in recent years.

In *Aim 2*, we imputed the country-level proportions of key indicators that were missing in surveys due to questions not asked in the survey. We found that *Amelia* and Multiple Imputation by Chain Equation for two-level panel data (*mice.2l.pan*) performed best among seven methods being evaluated for both methods converged fast, produced reasonable and stable imputations and had small out-of-sample root mean squared error (RMSE) less than ±5% for proportions imputed and 95% coverage rate (CR_{95}) very close to 95%. In addition, we found that including incomplete auxiliary variables that were correlated with targeted incomplete variables improved the imputation performance regardless of the missing rate of the auxiliary variables.

In *Aim 3*, we imputed the country-level proportions of key knowledge and attitudes indicators that were missing in the surveys using multiple imputation (*Amelia*) and then calculated the composite scores of knowledge and attitudes about HIV/AIDS in 47 SSA countries from 1998 to 2017. We found that the composite score of knowledge about HIV/AIDS had significantly increased in all 47 SSA countries over the past two decades. However, the composite score of attitudes about HIV/AIDS had only increased moderately in most countries. In a few countries, the attitudes score had even been declining in the recent years. In addition, we found that men generally had better scores of knowledge and attitudes about HIV/AIDS than women did but the difference between the men and women had been narrowing significantly over time.

The goal of the dissertation is to produce solid evidence on knowledge and attitudes about HIV/AIDS in SSA to inform decision and policy making for national HIV/AIDS prevention programs in SSA countries and to aid future research on HIV/AIDS in general in SSA.

Table of Contents

List of Figures

List of Tables

Acknowledgements

My PhD journey would not to be possible without the support of many.

First of all, I would like to extend my heartfelt gratitude to my dissertation committee members, Prof. Simon Hay, Haidong Wang, Jared Baeten, Gary Chan, Geoffrey Garnett and Jane Simoni, of all whom had provided tremendous support for my dissertation along the way. As the chair of my dissertation committee, Simon helped me out during the most difficult time of the process and I am beyond grateful for his support and kindness. Haidong, my academic supervisor at IHME, provided me with this research opportunity, funded and mentored me through the process, and lifted me up when I was down in life, for all of which I am extremely grateful. I also want to thank Jared, Gary, Geoff and Jane for their valuable inputs in this dissertation and for their generous supports as well.

In addition to my committee members, I would also like to thank the faculty and staff in the Department of Global Health. Especially, I would like to thank Prof. Steve Gloyd, Prof. Kenneth Sherr and Celine Abell for their enormous support along the journey. Not only did they help me academically, they also lifted me up spiritually when life knocked me down.

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

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

Chapter 1 : Introduction

Introduction

Since its discovery in the 1980s, Human Immunodeficiency Virus (HIV), the cause of Acquired Immune Deficiency Syndrome (AIDS), has claimed more than 35 million lives. The World Health Organization (WHO) estimated that globally there were 36.9 million people living with HIV (PLWH) in 2017, with 1.8 million people becoming newly infected.[1] After almost four decades, HIV/AIDS continues to be a leading infectious disease and a major global public health issue. Around the world, sub-Saharan Africa (SSA) shares a disproportionate burden of HIV/AIDS, with 25.7 million PLWH, 1.17 million new HIV infections, and 0.66 million AIDS-related deaths in 2017; accounting for 70% of global PLWH, 65% of global new infections, and more than 70% of global AIDS-related deaths respectively.[2] Most SSA countries have a generalized HIV/AIDS epidemic, meaning that the nation HIV prevalence rate is greater than 1%.[3]

Given the heavy burden of HIV/AIDS in SSA, tremendous research efforts and financial support from around the world have been mobilized to fight the epidemic in this region.[4] Despite the enormous resources mobilized for HIV/AIDS prevention and treatment, it is well recognized that successful prevention and treatment of HIV/AIDS relies not only on access to effective interventions, but also on individual behaviors,[5–12] which are in turn influenced by an individual's knowledge and attitudes about HIV/AIDS and its prevention and treatment.[13–22]

Therefore, knowing the level and trends of people's knowledge and attitudes about HIV/AIDS are crucial to the success of HIV/AIDS interventions and to understanding the trends of the HIV/AIDS epidemic in general. Since the late 1980s, donor-funded national surveys such as the Demographic Health Survey (DHS), Multiple Indicator Cluster Survey (MICS), and AIDS Indicators Survey (AIS) have been collecting data on people's knowledge and attitudes about

HIV/AIDS. In addition, some SSA countries also conducted their own national surveys, such as the Zambia Sexual Behavior Survey and Nigeria National HIV/AIDS and Reproductive Health Survey.

Although these surveys provide useful information on national levels of HIV/AIDS knowledge and attitudes, these data have rarely been compiled to establish the trend of people's knowledge and attitudes over time and across the SSA region. In addition, the national surveys are usually conducted every few years and there are more surveys for some SSA countries (e.g. Senegal) than for others (e.g. Swaziland). Therefore, the data are very scarce and unbalanced should one want to estimate the trends of HIV/AIDS knowledge and attitudes over time for all SSA countries.

To fill in this data gap, in *chapter two* of the dissertation, we estimate national trends of the key indicators of knowledge and attitudes about HIV/AIDS from 1988 to 2017 across 47 SSA countries using state-of-the-art techniques to project information over time and across space. By systematically searching for and utilizing all data available, this study seeks to produce crucial evidence on key indicators of people's knowledge and attitudes about HIV/AIDS in SSA countries, which will make comparison between countries possible and facilitate future research and treatment and prevention implementation in this region.

The health survey data used in this study, namely, the DHS, MICS and other country-specific national survey data, are all time-series-cross-sectional (TSCS) data. Although these TSCS data provide numerous measurements that open doors to many research opportunities, when used together, they often face serious missing data problem simply because some questions are not asked in some surveys (a.k.a. missing variables).[23–26] Even for DHS or MICS which are designed to be as consistent as possible over time and across countries, the questionnaires used in different

countries or in different rounds can be significantly different from each other due to local adaptation and changes in the health priorities over time.[27] When surveys from different sources are used, missing variables usually become more prevalent.

Multiple imputation (MI), a Bayesian model-based approach first introduced by Rubin[28], has become a major principled method for estimating missing data across different research fields.[24,29] The major benefits of MI are that it results in unbiased estimates, increase statistical power by using all available data and account for the uncertainty due to missing data.[28–30] Nowadays, there are two major families of MI approaches, namely, MI using joint modeling (MIJM) and MI using chain equations (MICE).[29,31,32] When first introduced by Rubin, MI was mainly used for imputing missingness in single-level cross-sectional data[28,33]. However, nowadays MI has been developed to properly impute missing data for TSCS data.[34]

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

Having estimates for each of the 16 key indicators of HIV/AIDS knowledge and attitudes is useful. However, it is also useful to have one composite score for knowledge and for attitudes about HIV/AIDS respectively. Therefore, in *chapter four* of the dissertation, we generate the composite scores of knowledge and of attitudes about HIV/AIDS and estimate their respective

trends over time in the 47 SSA countries. In addition, we use different visualization methods to show the changes of knowledge and attitudes scores over time in the SSA countries.

In summary, the dissertation generates important evidence on changes in people's knowledge and attitudes about HIV/AIDS over time in SSA countries. The results of the studies can inform decision and policy making for national HIV/AIDS prevention programs in SSA countries and to aid future research on HIV/AIDS in general in SSA.

Chapter 2 : Estimating Trends of the Key Indicators of Knowledge and Attitudes about HIV/AIDS in 47 Sub-Saharan African Countries from 1988 to 2017

Abstract

Background

HIV/AIDS has been a leading cause of death in sub-Saharan Africa (SSA) for decades. People's knowledge and attitudes about HIV/AIDS are potentially important determinants of their behaviors, which in turn are major contributors to both the spread of HIV/AIDS as well as the diffusion of effective treatment and prevention interventions. In this study, we estimate trends of key indicators of HIV/AIDS knowledge and attitudes in 47 SSA countries from 1988 to 2017.

Methods

We systematically searched for nationally representative data on HIV/AIDS knowledge and attitudes for the 47 SSA countries in the Global Health Data Exchange (GHDx), a comprehensive health data catalog established by the Institute for Health Metrics and Evaluation (IHME) at the University of Washington. All survey data were extracted in a systematic and consistent way so that national estimates of key HIV/AIDS knowledge and attitudes indicators are comparable across surveys. Lastly, point estimates of each key HIV/AIDS knowledge and attitudes indicator from different surveys were synthesized into trend estimates using spatial-temporal Gaussian process regression (ST-GPR), an innovative technique that gains strength over time and space to produce complete time series of an indicator as well as a 95% confidence interval.

Results

Among 3,682 surveys of the 47 SSA countries in GHDx, we identified 248 surveys that have at least one of 16 key HIV/AIDS knowledge and attitudes indicators. Although the trends vary greatly from country to country, overall the indicators demonstrate an increasing (improving) patterns across the board except for two indicators, namely, people's knowledge of HIV transmission from mother to child during pregnancy, and people's attitudes toward disclosing family members' HIV/AIDS status, which show a significant decreasing (deteriorating) trend for many countries. Among the 47 SSA countries, those in Somalia, Sudan, and South Sudan demonstrate lower levels of HIV/AIDS knowledge and accepting attitudes compared with people in other countries. Men in SSA generally have better knowledge and attitudes about HIV than women do, except for MTCT indicators. Lastly, older people in SSA tend to have better knowledge and attitudes about HIV/AIDS than younger people.

Conclusion

Although there is great heterogeneity regarding levels and rates of change across countries, in sub-Saharan Africa, people's knowledge of HIV/AIDS and their attitudes toward people living with HIV/AIDS have, in general, improved over the past 30 years. However, people's knowledge of mother-to-child transmission (MTCT) during pregnancy, and people's attitudes on keeping family members' HIV status secret have demonstrated deteriorating trends over time, which is concerning.

1. Introduction

Since its discovery in the 1980s, Human Immunodeficiency Virus (HIV), the cause of Acquired Immune Deficiency Syndrome (AIDS), has claimed more than 35 million lives. The World Health Organization (WHO) estimated that globally there were 36.9 million people living with HIV (PLWH) in 2017, with 1.8 million people becoming newly infected.[1] After almost four decades, HIV/AIDS continues to be a leading infectious disease and a major global public health issue. Around the world, sub-Saharan Africa (SSA) shares a disproportionate burden of HIV/AIDS, with 25.7 million PLWH, 1.17 million new HIV infections, and 0.66 million AIDS-related deaths in 2017; accounting for 70% of global PLWH, 65% of global new infections, and more than 70% of global AIDS-related deaths respectively.[2] Most SSA countries have a generalized HIV/AIDS epidemic, meaning that the nation HIV prevalence rate is greater than 1%.[3]

Given the heavy burden of HIV/AIDS in SSA, tremendous research efforts and financial support from around the world have been mobilized to fight the epidemic in this region.[4] Despite the enormous resources mobilized for HIV/AIDS prevention and treatment, it is well recognized that successful prevention and treatment of HIV/AIDS relies not only on access to effective interventions, but also on individual behaviors,[5–12] which are in turn influenced by an individual's knowledge and attitudes about HIV/AIDS and its prevention and treatment.[13–22]

Therefore, knowing the level and trends of people's knowledge and attitudes about HIV/AIDS are crucial to the success of HIV/AIDS interventions and to understanding the trends of the HIV/AIDS epidemic in general. Since the late 1980s, donor-funded national surveys such as the Demographic Health Survey (DHS), Multiple Indicator Cluster Survey (MICS), and AIDS Indicators Survey (AIS) have been collecting data on people's knowledge and attitudes about

HIV/AIDS. In addition, some SSA countries also conducted their own national surveys, such as the Zambia Sexual Behavior Survey and Nigeria National HIV/AIDS and Reproductive Health Survey.

Although these surveys provide useful information on national levels of HIV/AIDS knowledge and attitudes, these data have rarely been compiled to establish the trend of people's knowledge and attitudes across time and across the SSA region. In addition, the national surveys are usually conducted every few years and there are more surveys for some SSA countries (e.g. Senegal) than for others (e.g. Swaziland). Therefore, the data are very scarce and unbalanced should one want to estimate the trends of HIV/AIDS knowledge and attitudes over time for all SSA countries.

To fill in this data gap, the current study aims to estimate national trends of knowledge and attitudes about HIV/AIDS from 1988 to 2017 across 47 SSA countries using state-of-the-art techniques to project information over time and across space. By systematically searching for and utilizing all data available, this study sought to produce crucial evidence on people's knowledge and attitudes about HIV/AIDS in SSA countries, which will make comparison between countries possible and facilitate future research and treatment and prevention implementation in this region.

2. Methods

2.1 Countries of interest

In this study, we selected 47 SSA countries as countries of interest and estimated the national trends of key indicators of people's knowledge and attitudes about HIV/AIDS in these 47 countries. Based on the definitions of subregions in the Global Burden of Disease 2017 (GBD

2017), the 47 countries are grouped into five subregions. ***Figure 2-1*** shows a map of these countries and **Table 2-1** lists the 47 SSA countries.

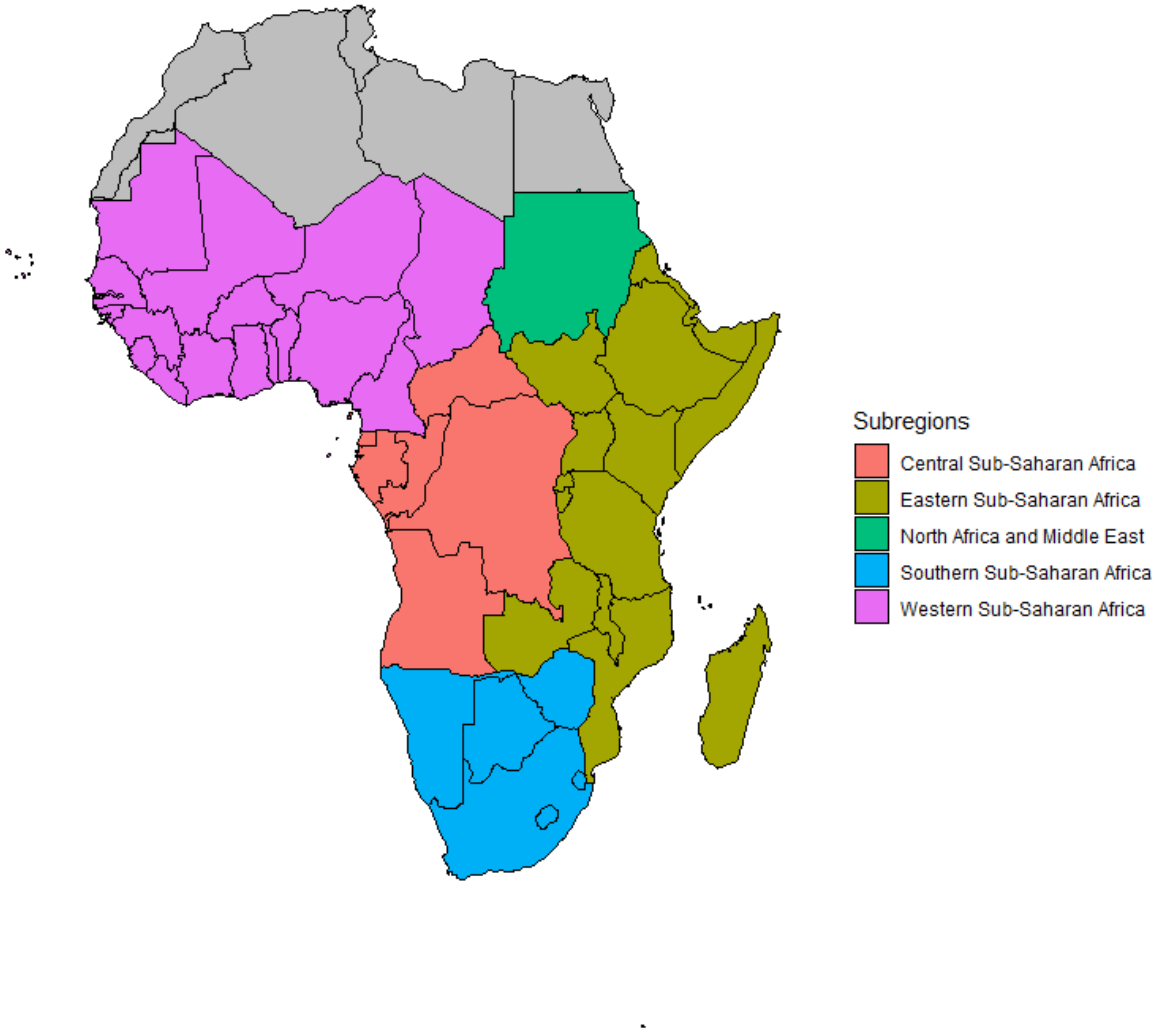


Figure 2-1 A Map of sub-Saharan African Countries by Subregion

Table 2-1 The 47 sub-Saharan African countries of interest

Subregions of SSA	Countries of Interest
Central SSA	Angola (AGO), Central African Republic (CAF), Congo (COG), Democratic Republic of the Congo (COD), Equatorial Guinea (GNQ), and Gabon (GAB)
Eastern SSA	Burundi (BDI), Comoros (COM), Djibouti (DJI), Eritrea (ERI), Ethiopia (ETH), Kenya (KEN), Madagascar (MDG), Malawi (MWI), Mozambique

	(MOZ), Rwanda (RWA), Somalia (SOM), South Sudan (SSD), Tanzania (TZA), Uganda (UGA), and Zambia (ZMB)
Western SSA	Benin (BEN), Burkina Faso (BFA), Cameroon (CMR), Cape Verde (CPV), Chad (TCD), Cote d'Ivoire (CIV), The Gambia (GMB), Ghana (GHA), Guinea (GIN), Guinea-Bissau (GNB), Liberia (LBR), Mali (MLI), Mauritania (MRT), Niger (NER), Nigeria (NGA), Sao Tome and Principe (STP), Senegal (SEN), Sierra Leone (SLE), and Togo (TGO)
Southern SSA	Botswana (BWA), Lesotho (LSO), Namibia (NAM), South Africa (ZAF), Swaziland (SWZ), and Zimbabwe (ZWE)
Northern SSA	Sudan (SDN)

2.2 Data searching, data eligibility, and data extraction

To find all survey data available on HIV/AIDS knowledge and attitudes in the 47 SSA countries, we conducted a systematic data search using the Global Health Data Exchange (GHDx), the world's most comprehensive catalog for health-related data.[35] The data search was restricted to all surveys conducted in the 47 SSA countries from 1980 to 2017. The eligible surveys have nationally representative data on at least one of the key indicators of HIV/AIDS knowledge and attitudes.

For eligible surveys with microdata, we systematically extracted relevant variables from each survey and mapped these variables into standardized indicators that measure the same thing, have the same variable name, and are comparable across all surveys. When extracting the original data and standardizing the indicators, we carefully reviewed the questionnaire of each survey to ensure that the indicators are comparable across surveys and that the skip pattern of each survey is appropriately accounted for. The codebook mapping the original indicators to standardized indicators is included in Supplementary Materials. The expected results of data extraction are a set of datasets, each of which contains individual level data from a survey on standardized indicators of HIV/AIDS knowledge and attitudes as well as demographic

information including age, sex, location, year, and sampling weight. These datasets contain the original data from the surveys but only in a standardized way.

For surveys with report data only, we manually extracted the available indicators from the report. The extracted data includes sample size and the country-level estimates of the indicators by sex and by age group if available. The sampling variance is calculated using the binomial variance equation shown below.

$$SE_{i,s} = \sqrt{\frac{DEFT_{i,s} * p_{i,s}(1 - p_{i,s})}{n_{i,s}}}$$

where $SE_{i,s}$ is the standard error of indicator i for gender s, $p_{i,s}$ is the country-level estimate and is a proportion, $n_{i,s}$ is the sample size, and $DEFT_{i,s}$ is the design effect of indicator i in survey s due to multistage cluster sampling. STATA 13 and R 3.5.0 are used for data cleaning and management. A full list of extracted indicators is shown in **Table 2-2**.

2.3 Definitions of key indicators and the estimation period

To measure people's knowledge and attitudes toward HIV/AIDS, DHS and other surveys collect data on multiple indicators on people's knowledge of HIV transmission (including MTCT), prevention, testing, and misconceptions of people's attitudes toward PLWH.[36] From all the indicators collected by these surveys, we selected 16 indicators, including 12 knowledge indicators and four attitudes indicators. The reason we selected these 16 indicators is that they are core indicators of HIV/AIDS knowledge and attitudes in most surveys and have been consistently collected in most surveys for a long time. **Table 2-2** provides detailed information of the 16 indicators, including their definitions, variable name, and meaning of indicator values.

Note that the values of the indicators have been recoded so that "1" always suggests a positive answer, i.e. knowing ways of HIV/AIDS transmission and prevention, rejecting misconceptions about HIV/AIDS, and showing positive attitudes towards PLWHs; "0" always suggests the opposite. "Don't know (DK)" is always coded as a negative response, which is particularly important for attitudes questions since people who answer "don't know" usually imply a negative attitudes but may not feel comfortable to verbalize it.

Table 2-2 Standardized indicators of HIV/AIDS knowledge and attitudes

Category	Definition of indicator	Variable name	Values
Gateway question	Ever heard of HIV/AIDS	heard_aids	1: Yes 0: No/DK
Knowledge on HIV/AIDS prevention	Knowing that one can reduce chance of getting AIDS by having just one uninfected partner who has no other sex partners	kw_pv_one_partner	1: Yes 0: No/DK
	Knowing that one can reduce change of getting AIDS by using a condom every time they have sex	kw_pv_condom	1: Yes 0: No/DK
Knowledge on HIV/AIDS mother-to-child transmission	Knowing that HIV can be transmitted from mtc during pregnancy	kw_mtct_preg	1: Yes 0: No/DK
	Knowing that HIV can be transmitted from mtc during delivery	kw_mtct_delivery	1: Yes 0: No/DK
	Knowing that HIV can be transmitted from mtc through breasfeeding	kw_mtct_bf	1: Yes 0: No/DK
	Knowing that there is a drug to prevent mtct	kw_mtct_drug	1: Yes 0: No/DK
Knowledge on HIV/AIDS misconceptions	Believing that HIV can be transmitted by mosquito bites	mis_mosquito	1: No 0: Yes/DK
	Believing that HIV can be transmitted by sharing food with PLWHs	mis_food	1: No 0: Yes/DK
	Believing that one can get AIDS from witchcraft	mis_witchcraft	1: No 0: Yes/DK
Knowledge on HIV/AIDS testing	Knowing a place where people can get tested for HIV	kw_where_test	1: Yes 0: No/DK
Other knowledge on HIV/AIDS	Knowing that healthy-looking person can have HIV	kw_looking	1: Yes 0: No/DK

Attitudes towards people living with HIV/AIDS	Would buy fresh vegetable from a HIV-infected vendor	att_vegetable	1: Yes 0: No/DK
	Would want to remain a secret if a family member got infected with HIV	att_secret	1: No 0: Yes/DK
	Willing to care for an infected family member in his/her own household	att_willing_care	1: Yes 0: No/DK
	Believing that a female teacher with HIV should be allowed to continue teaching in the school	att_f_teacher	1: Yes 0: No/DK

It is worth noting that *heard_aids* is the gateway indicator for all of the other indicators. For knowledge indicators, if the person interviewed had not heard of HIV/AIDS by the time of the interview, i.e. *heard_aids* = 0, all the knowledge indicators for the person are set to 0. However, for attitudes indicators, if *heard_aids* = 0, all the attitudes indicators for the person are set to NA because it is meaningless to ask a person's attitudes towards HIV/AIDS if the person is not aware of the disease. In many surveys, there are also gateway indicators for knowledge of HIV/AIDS prevention, misconceptions, and mother-to-child transmission. These gateway indicators are carefully dealt with in each of the relevant surveys in a case-by-case manner. For example, *kw_mtct_preg*, *kw_mtct_delivery*, and *kw_mtct_bf* are the gateway indicators for *kw_mtct_drug* in some surveys. If the person knew none of the three ways of mother-to-child HIV transmission, we assume the person did not know the existence of the drug that prevents MTCT of HIV. Detailed data cleaning and management strategies are included in the Appendix.

Since the data on HIV/AIDS knowledge and attitudes were first collected by DHS in 1988, we estimated the trends of the HIV/AIDS knowledge and attitudes key indicators from 1988 to 2017.

2.4 Country-level estimates of the key indicators

After extracting the individual-level data, we obtained the country-level estimates of the key indicators by calculating the weighted mean of individual responses. The weights used are the sampling weights of the survey. The weighted mean of each indicator is calculated by gender and age group, including 15-49, 15-24, and 25-49. Since the individual data are all binary, the weighted means are proportions between 0 and 1. We used the *survey* package in R to obtain the weighted means and their corresponding standard errors while accounting for survey design elements such as stratification and clustering.

2.5 Country-level covariates

To estimate the trends of HIV/AIDS knowledge and attitudes indicators, important covariates are used for bias adjustment and in the data synthesis process to improve the estimate. The country-level covariates used in this study are listed in **Table 2-3** below.

Table 2-3 List of country-level covariates

Covariate	Description	5-year-age-group specific?	Gender-specific?
education	Mean years of education per capita	Yes	Yes
GDP	GDP per capita base, 2010 international dollars	No	No
ASFR	Age-specific fertility rate	Yes	No
contra_prev	Modern contraception prevalence in women by age groups	Yes	No
HAQI	Healthcare access and quality index [37]	No	No
prop_urban	Proportion of population living in urban area	No	No
Muslim	Binary indicator: value 1 if country is greater than 50% Muslim	No	No
HSA	Health system access: a composite score of immunization, measles	No	No

	immunization, hospital beds, in-facility delivery, and skilled birth attendant		
ANC4	Proportion of pregnant women receiving 4 or more antenatal care visits from a skilled provider	No	No

For covariates that are originally age-group specific, levels of the covariate within the 15-49 and 15-24 age groups are the weighted mean of the covariate within relevant age groups using the population of each age group as the weights. Male and female mean years of education are used separately in the analysis for males and females respectively. The level of all covariates are estimated using the IHME GBD study 2017. Data of the covariates are extracted from the IHME database for the period 1988 to 2017.

2.6 Data bias adjustment

Since the country-level estimates may differ systematically between different surveys, we incorporated data bias adjustment before synthesizing the data to produce the trend. The approach is adapted from Wang *et al.*'s study.[38] Specifically, we modeled the logit-transformed country-level data for each indicator using a linear mixed effect model, which includes a fixed effect for data source type across all locations and a random effect for data source type nested within each country (country-source random effect), controlling for important covariates such as year of data, mean years of education per capita, and log-transformed GDP per capita (base 2010 international dollar).

$$\text{logit}(I_{cys}) = \beta_1 * education_{cy} + \beta_2 * GDP_{cy} + \gamma_c + \gamma_{cs} + \alpha_s + \varepsilon_{cys} \quad (1)$$

where I is indicator, c is country, y is year, s is source type, γ_c is country random effect, γ_{cs} is country-source random effect, α_s is source type fixed effect across countries, and ε_{cys} is the residual term. There are three source types total, i.e. DHS (including AIS), MICS, and other

surveys, among which DHS is believed to be the least biased and thus is used as the reference source.

Based on equation (1), each data source has an associated random effect (γ_{cs}) and a source type fixed effect (α_s). The values of these random and fixed effects for the reference source (DHS) are deemed to be the true deviation from the unbiased estimate. Therefore, we adjusted the non-reference source types by replacing the estimated random and fixed effect values for these non-reference source types with the values for the reference type, as shown below.

$$\text{logit}\left(I_{adj,cys}\right) = \beta_1 * education_{cy} + \beta_2 * GDP_{cy} + \gamma_c + \hat{\gamma}_{c,ref} + \hat{\alpha}_{ref} + \varepsilon_{cys} \quad (2)$$

where $\hat{\gamma}_{c,ref}$ and $\hat{\alpha}_{ref}$ are the random and fixed effects estimated for the DHS survey, respectively. Therefore, national estimates from DHS surveys are unchanged but the estimates from other sources are adjusted accordingly. Using equation (2), we corrected the bias in the data due to source type for males and females separately.

2.7 Crosswalking

Among the 249 surveys, all surveys have data for women but only 184 (73.9%) have data for men. Given the scarcity of data in many countries, the imbalance of data between men and women is concerning because the estimated trends for men and for women can be very different simply because the data are missing for men but not for women in certain years. Moreover, it is reasonable to assume that an indicator for women is informative of the same indicator for men and vice versa. To address the issue, we used a procedure called "crosswalking" which is widely used in the GBD studies.[6,38–41]

We use the crosswalking from women to men as an example. Briefly, if a survey only has data for women, we impute the data for men using fitted values from the following mixed effect linear model fitted using all surveys with data for both men and women:

$$\text{logit}(I_{adj,m,k}) = \beta_0 + \beta_1 \text{logit}(I_{adj,f,k}) + \beta_2 \text{logit}(I_{adj,f,k}) * year + \gamma_c + \varepsilon_k \quad (3),$$

where $I_{adj,m,k}$ and $I_{adj,f,k}$ are the respective bias-adjusted country-level estimates of an indicator for male and for female in survey k, $\text{logit}(I_{adj,f,k}) * year$ is the interaction between logit-transformed data and $year$ (centered to 2000). γ_c, is the country random effect and ε_k is the error term.

There are three other candidate models that are considered for the crosswalking, namely, a simple linear model, a linear model with country random intercepts, and a linear model with both country random intercepts and country random slopes on $I_{f,k}$. We choose model (3) over the other three models based on the smallest root mean squared error (RMSE) and the relative simplicity of the model. We include the interaction between data and year because we find that the relationship between data of men and of women has changed significantly over time for many indicators, such as *heard_aids*, *kw_looking*, *kw_mtct_preg*, and *kw_where_test*. After including the interaction term, the model fit for the indicators has improved.

The prediction takes into account both the fixed and the random effects in the model. The prediction variance, $var(\hat{I}_{m,k})$, is estimated using the bootstrapping method, which utilizes simulation with repeated sampling. Specifically, we simulate 1,000 predictions for each $loigt(I_{m,k})$ with each prediction using a random draw from the joint distribution of the model parameters. We then transform the 1,000 draws back into proportion and calculate the variance

of the 1,000 simulated proportions to obtain the prediction variance.[42,43] The estimation is conducted using the R package lme4[43].

The same method is used for crosswalking between different age groups. We first crosswalked the data between genders and then between age groups. Since all surveys have data for women and for age group 15-49, we used data for women and data for age group 15-49 as the reference group for the crosswalking. After the crosswalking, every indicator is expected to have the same number of data points for men and women and for all age groups.

2.8 Data synthesis using ST-GPR

Once we obtained the bias-adjusted and crosswalked country-level estimates of the key indicators, we used spatiotemporal Gaussian process regression (ST-GPR) to impute missing location-year combinations and to estimate trends and the uncertainty intervals of the indicators in all 47 countries. The estimation is done by gender and age group. Full details on the ST-GPR method have been published previously.[38–41,44] However, there have been some important updates and adjustments to the ST-GPR previously used.

First-stage: generalized additive mixed model with random slop and random intercept

Briefly, ST-GPR is a three-stage estimation process. In the first stage, we fit the bias-adjusted and crosswalked data obtained from the steps above to a generalized additive mixed model (GAMM) with random slope and random intercept and then predict the time series of an indicator with the fitted model. The model is as follows:

$$\begin{aligned} logit(I_{cy}) = {} & (\beta_0 + \gamma_{0,c}) + \gamma_{1,c} year + \boldsymbol{NS(year)} + \beta_2 education_{cy} + \beta_3 \, log(GDP_{cy}) \\ & + \beta_4 \, logit(contra\_prev_{cy}) + \beta_5 ASFR_{cy} + \beta_6 HAQI_{cy} + \beta_7 Muslim_c \\ & + \beta_8 logit(prop\_urban_{cy}) + \beta_9 HSA_{cy} + \beta_{10} logit(ANC4_{cy}) + \varepsilon_{cy} \quad (6) \end{aligned}$$

where c is country, y is year, I_{cy} is indicator level of country c in year y, $\gamma_{0,c}$ is country random intercept, $\gamma_{1,c}$ is country random slop of $year$, $\boldsymbol{NS}(\boldsymbol{year})$ is the natural cubic spline of $year$ given vector of knots $\boldsymbol{k}$ and vector of spline coefficients $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$, and ε_{cy} is the error term. The natural cubic spline can be described by the following equations:

$$\begin{cases} \sum_{i=0}^{3} \lambda_i X^i + \sum_{j=1}^{J} \theta_j \left(X - k_j\right)_+^3 \\ \lambda_2 = \lambda_3 = 0 \\ \sum_{j=1}^{J} \theta_j = 0 \\ \sum_{j=1}^{J} \theta_j k_j = 0 \end{cases}$$

where i is the spline degree (cubic spline has a degree of 3), J is the total number of knots and X is the indicator of interest. The spline basis

$$\left(X - k_j\right)_+ = \begin{cases} 0, & if\ X \le k_j \\ X - k_j, & if\ X > k_j \end{cases}$$

The first equation is a general cubic spline and the constraint equations 2-4 ensure that the second derivatives of the spline at the two boundary knots is 0, which makes the natural spline function linear beyond the boundary knots.

The estimates of cubic spline are often erratic near the boundaries, making the extrapolation dangerous. By constraining the spline to be linear beyond the boundaries, natural spline produces more reasonable estimates at and beyond the boundaries in many cases.[45,46] This feature of natural cubic spline is particularly useful for the current study since many of the HIV/AIDS

knowledge and attitudes indicators demonstrate a linear trend in recent years. To make the model flexible enough to fit different trends of the indicators in different countries but not too flexible to overfit the scarce data in each country, we chose the minimum of 3 knots for the spline, including two boundary knots and one interior knot. The locations of the knots are selected based on model fit statistics such as AIC/BIC and visual inspection of the plausibility of the estimated trends. Eubank, in his book *Nonparametric Regression and Spline Smoothing*, recommended visual inspection as the simplest method for selecting spline knots and stated that selecting spline knots through visual inspection "often tends to work quite well in terms of giving a visual pleasing fit to the data and has a definite computational advantage over other methods".[47] Using visual inspection to select spline knots is also common in research.[48,49] In our study, selecting knots through visual inspection is even more important and useful because it helps provide plausible extrapolations in the early years when there are few or no data. Following Eubank's recommendation, we place the three knots to where the trend seems to change most significantly.[47] Regarding the plausibility of the trends, based on the early history of HIV/AIDS (Appendix **Box 1.**), we believe that people's knowledge and attitudes about HIV/AIDS should be generally low in 1988, unless the data suggest otherwise. Detailed rules of the knot selection and the final locations of the knots selected for each indicator are included in the Appendix.

The covariates in the model are selected based on a literature review[50–60] and expert opinions. Logit and log transformation are used for certain covariates to improve the model fit, and the restricted maximum likelihood (REML) approach is used to fit the model. With this model, we obtain the first-stage prediction of the time series of the indicators.

Second-stage: spatiotemporal smoothing

Although model (6) is carefully specified and accounts for variations across countries, the assumptions of the model are still too strict to fit the trends of the indicators in different countries. To relax these assumptions and improve the estimation, in the second stage we smooth the residuals between the first-stage predictions and the bias-adjusted raw data across time, age group, and countries by applying a combination of smoothing functions to these residuals. For each country-year-age, we weighted all observed residuals based on their proximity to this country-year-age in space, time, and age group. The time weight w_t is specified as follows:

$$w_t = e^{-\lambda|year_i - year_j|} \qquad (7)$$

where i is the country-year to be predicted, j is the observed data point, and λ is the chosen parameter controlling how much strength is borrowed across years. The strength borrowed over time decreases exponentially given equation (7).

The space weight w_s of each residual is specified as follows:

$$w_s = \begin{cases} \zeta^0 = 1 & for\ residuals\ within\ country\ c \\ \zeta^1 & for\ residuals\ within\ region\ r\ but\ not\ country\ c \\ \zeta^2 & for\ residuals\ outside\ region\ r \end{cases} \qquad (8)$$

where ζ is a scaler determining how much down-weight is placed on region and super region residuals. For example, when we estimate smoothed residuals within Kenya, the observed residuals within Kenya always have weight 1 whatever ζ is. However, the observed residuals in other Eastern SSA countries such as Tanzania or Uganda have weight of ζ (e.g. 0.01) and the observed residuals within countries outside Eastern SSA, such as Ghana or South Africa, have a weight of ζ^2 (e.g. 0.0001). With this space weight, we can better estimate the trends in a country by borrowing information from its neighboring countries while still valuing the information within the country the most. This is particularly useful for countries with no or few data points.

The age weight w_a is specified as follows:

$$w_{a,ij} = e^{-\omega * |\, agegroup_i - agegroup_j \,|} \qquad (9)$$

where ω controls how quickly the age weights diminish. The larger the ω is, the less strength is drawn across age groups.

Using the time, space, and age weights, we obtain the smoothed residuals for each country-year-age by calculating the weighted average of all the observed residuals

$$SR_{cya} = \sum w_{t,ijk|cya} w_{s,ijk|cya} w_{a,ijk|cya} * OR_{ijk} \qquad (10)$$

where SR_{cya} is the smoothed residual for each country-year-age combination, OR_{ijk} is the observed residual for country i and age group k in year j, $w_{t,ijk|cya}$, $w_{s,ijk|cya}$ and $w_{a,ijk|cya}$ are time, space and age weight for an observed residual OR_{ij} with respect to a specific country-year-age combination.

We then obtain the second-stage predictions of the time series by adding the smoothed residuals to the first-stage prediction.

Third stage: Gaussian process regression

In the third stage, GPR is used to produce the final time series of the indicator, as well as the uncertainty bounds. The model of GPR is shown below

$$f(t) \sim GP(M, C) \qquad (11)$$

where M and C are the mean function and the covariance matrix respectively. Full details of GPR have been well described in other studies.[38,44] In a nutshell, we use the second-stage predictions as the prior of the mean function, use a Matérn covariance function to describe the

covariance prior[38] and specify a likelihood function which describes the probability of observing the data given a particular set of parameters. The likelihood function is described below

$$logit(I_t) \sim Normal(f(t), V_t) \qquad (11)$$

where $f(t)$ is the mean and V_t is the data variance including both sampling and non-sampling variance (NSV). NSV refers to variability of data due to random errors in the data collection process and is largely unknown. We approximate the NSV between data using the method proposed by Foreman et.al. [44] After specifying the prior of the mean function, the prior of the covariance and the likelihood function, we use Markov Chain Monte Carlo (MCMC)[61] to approximate the posterior distribution of $f(t)$ which incorporates information from both second-stage predictions and the observed country-level indicators. In total, 5000 MCMC samples are produced. The first 3,000 samples are burned and the remaining 2,000 samples are thinned by a factor of 2 leaving 1,000 draws in the end. The third-stage prediction or the final estimates and the uncertainty bounds are generated form the median and the 2.5th and 97.5th percentiles of the 1,000 draws, respectively.

Parameter selection

There are a total of four parameters that need to be specified to run ST-GPR, namely, λ, ζ, ω and scale. As mentioned above, λ and ω control how quickly the time and age weights diminish and ζ determines how much to down-weight the region and super region residuals. Scale is a parameter of the GPR and determines how correlated the GPR estimates are over time. A scale of 15 simply means that estimates within 15 years apart are correlated. Given the limited number of data points, we follow the newest guideline used in the GBD 2017 and choose $\lambda = 0.25$, $\zeta = 0.01$, $\omega = 1$ and $scale = 15$ for the estimation of all the indicators. These parameters borrow

enough strength across time and space while still ensuring that the GPR follows data within the country rather than data outside the country when estimating a time series for a location.[39]

In addition to the three parameters noted above, amplitude is another important parameter affecting the variance around the mean estimates in the GPR stage. A large amplitude gives the GPR more flexibility to move away from the second-stage estimates while a small amplitude forces the estimates to be around the second-stage estimates. To properly account for the limited data available and the uncertainty of the data, we set the amplitude to be the median of all squared data variances multiplied by 1.96.

2.9 Identification and removal of outliers

To ensure the quality of the estimates, we carefully reviewed the raw data and the estimates to identify and remove outliers. Firstly, we remove data sources with quality concern such as Nigeria DHS 1999 and Malawi Global Fund Household Health Coverage Survey 2007-2008. Secondly, we removed certain indicator(s) from a survey due to quality concerns such as having too many missing values for no reason or applying wrong skip logic for some indicators. Lastly, we utilized statistical method and expert opinion to identify extreme or unusual county-level estimates. Specifically, once we have obtained the residuals between the first-stage predictions and the bias-adjusted country-level estimates, we calculated the MAD of residuals within each country and identified data points with residual three MAD away from the median.[62] We then consulted experts to determine whether these extreme data points are actual outliers.

In short, outliers were removed with great caution and they were carefully documented for reference. The detailed descriptions of problematic surveys, indicators, and outliers are included in the appendix.

2.10 Estimation of mean trends by gender, by region, and by age groups

After identifying the country-sex-age-group specific time series of each indicator, we synthesized the time series of each indicator to obtain the mean trends of the indicators by sex, by region, and by age group using the GAM smooth function with the formula

$$\bar{I}_{cys} = cs(year) + \varepsilon_{cys}$$

where c, y, s, ε are country, year, sex, and the error term. The $cs(year)$ is the shrinkage version of penalized cubic splines of year with knots spread evenly through the covariate values. [63–65]

3. Results

3.1 The data sources

By April 1st, 2019, there were a total of 3,682 surveys in the 47 SSA countries from 1980 to 2017 in GHDx, among which 248 are accepted. Among the 248 accepted surveys, 231 have microdata (individual level data) and 17 have report data only. A complete list of all sources is available in the Supplementary Materials.

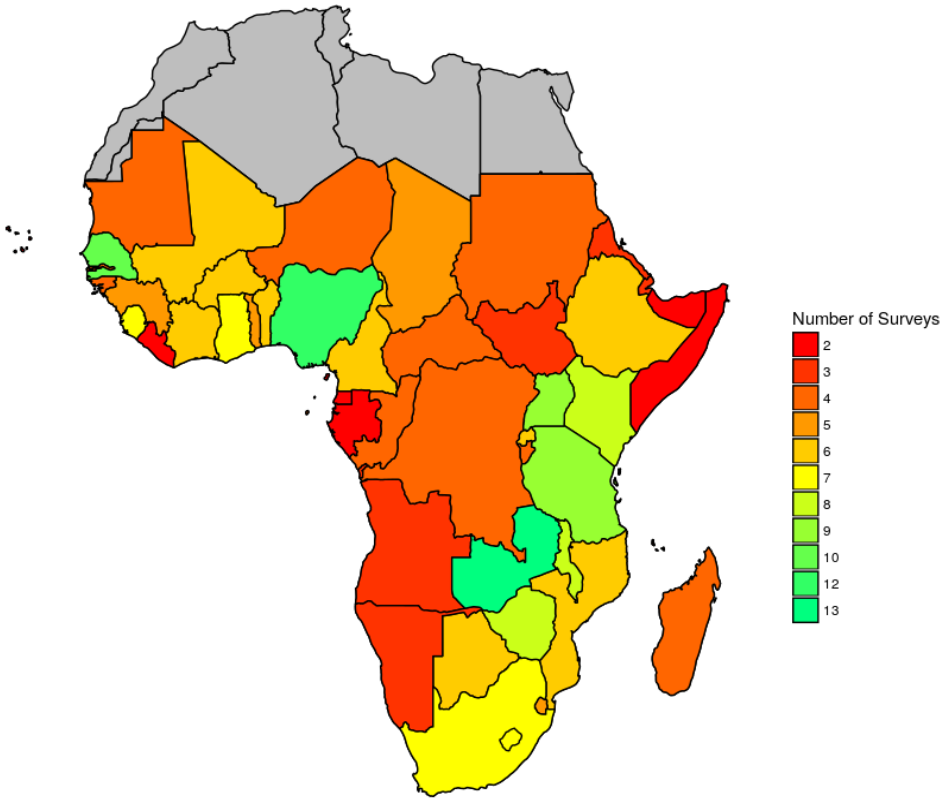


Figure 2-2 A Map of number of surveys by country

As shown in ***Figure 2-2***, data are in general scarce in SSA countries, especially in central SSA countries where no country has more than five data sources. Eastern and western SSA have more

data sources compared with the other regions but the number of data sources varies greatly across countries. Country-specifically, Zambia, Nigeria and Senegal have the most data sources (>10) followed by Tanzania, Uganda, Kenya, Malawi, Zimbabwe, Ghana, Sierra Leone, South Africa, and Lesotho (7-9). On the other end of the spectrum, Liberia, Gabon, Equatorial Guinea, Angola, Namibia, South Sudan, Somalia, Djibouti, and Eritrea have the scarcest data sources (2-3) on HIV/AIDS knowledge and attitudes.

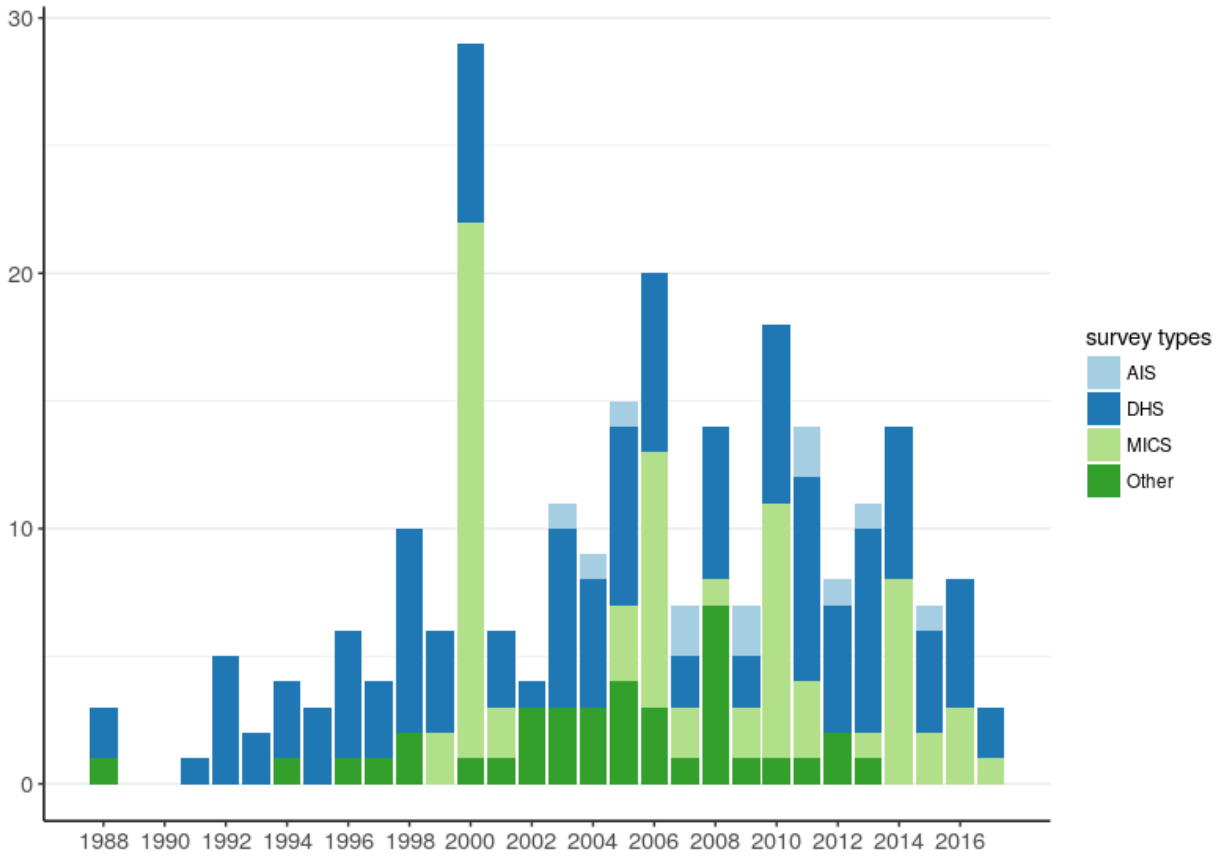


Figure 2-3 Number of surveys by survey type and year

Among the 248 surveys used in the study, there are 128 DHS, 71 MICS, 12 AIS, and 37 other surveys, accounting for 51.6%, 28.6%, 4.8%, and 14.9% of the surveys respectively. The DHS and the MICS are the majority sources, jointly accounting for 80% of all sources. ***Figure 2-3*** visualizes the distribution of the surveys by survey type and by year. There is an increasing trend

in the number of surveys from 1988 to 2006 except for the big jump in 2000 due to the MICS. After 2006, there seems to be a decreasing trend in the number of available surveys, which is partly due to the fact that surveys conducted in recent years have not yet become available to the public.

To estimate the trends of HIV/AIDS knowledge and attitudes for men and for women, we extracted data for both sexes from the surveys. As described in the methods section, we deal with the imbalance of data between genders through crosswalking. However, there are 16 key indicators and most surveys have some but not all of the indicators. ***Figure 2-4*** shows the indicator coverage by year. In ***Figure 2-4***, the y-axis is ordered by the total coverage rate of the indicators and the two x-axes show the coverage rate and the number of surveys respectively. We can see that the indicators *heard_aids* and *kw_looking* have the highest coverage rates of over 90% followed by *kw_pv_condom*, *kw_pv_one_partner*, *kw_pv_one_partner*, and *mis_mosquito* which have coverage rates of over 80%. In fact, all indicators are available in more than half of the surveys except for *kw_mtct_drug*, which are available in only 105 surveys or a coverage rate of 42%. From the temporal perspective, the great majority of data for all indicators are between the year 2000 and 2014. Before 2000, there are very limited data for most indicators; all indicators are available after 2000. Since 2005, the 16 key indicators have been consistently collected in most surveys, especially in the DHS, MICS, and AIS.

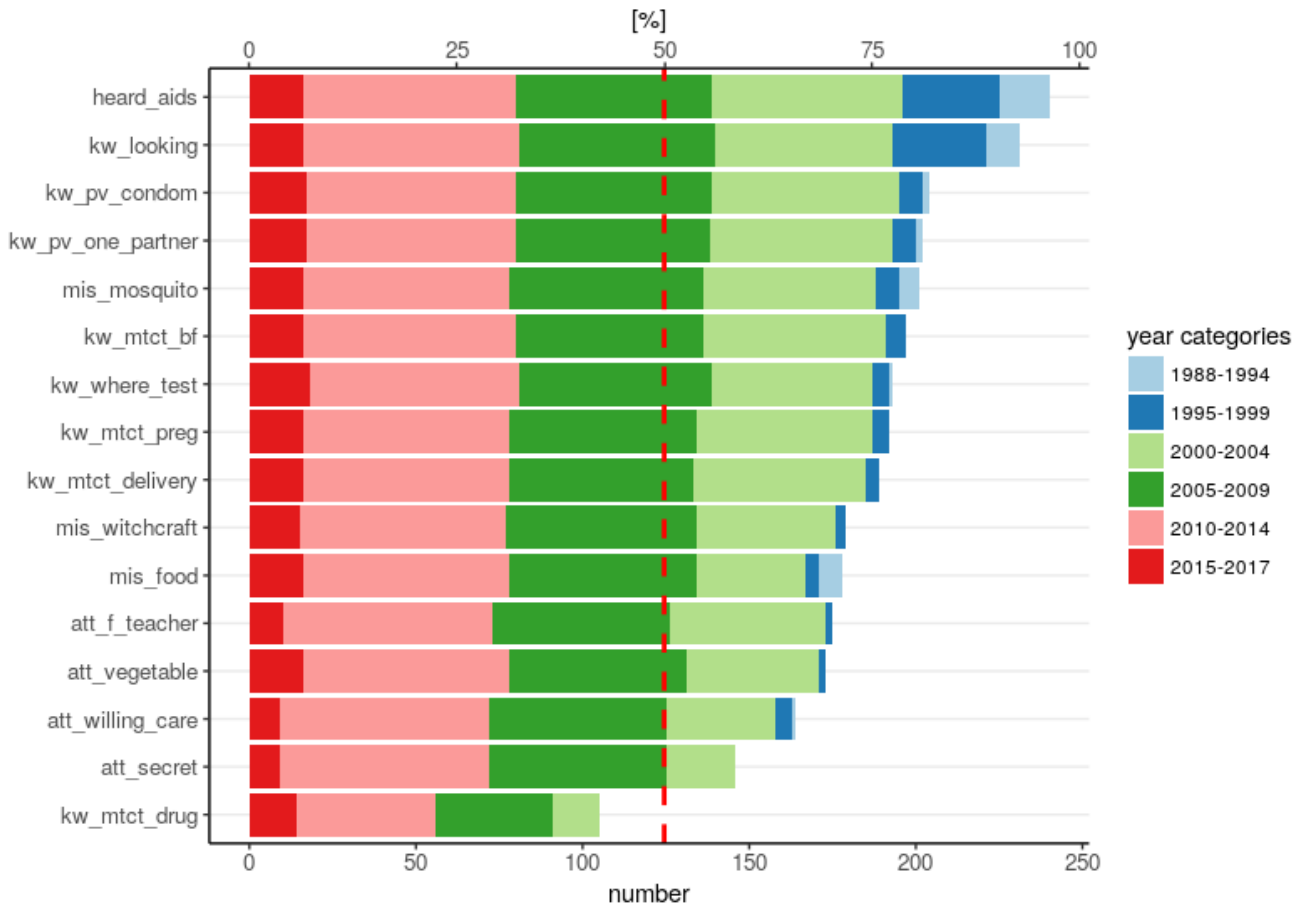


Figure 2-4 Indicator coverage plot by year

Figure 2-5 shows the number of indicator data by country as a heat map. The countries on the y-axis are ordered by the country's total number of surveys and the indicators on the x-axis are ordered by the indicator's coverage rate among the 249 surveys. In general, the color shifts from green to red from the left to the right and from the top to the bottom, indicating decreasing number of data on the indicators. If a country has more surveys, it usually has more data for the indicators as well. However, this is not always the case, as exemplified by South Africa, which has seven surveys but has a limited number of data on the indicators because most surveys of South Africa only have data on a few indicators. These unusual cases can be easily identified in the heat map.

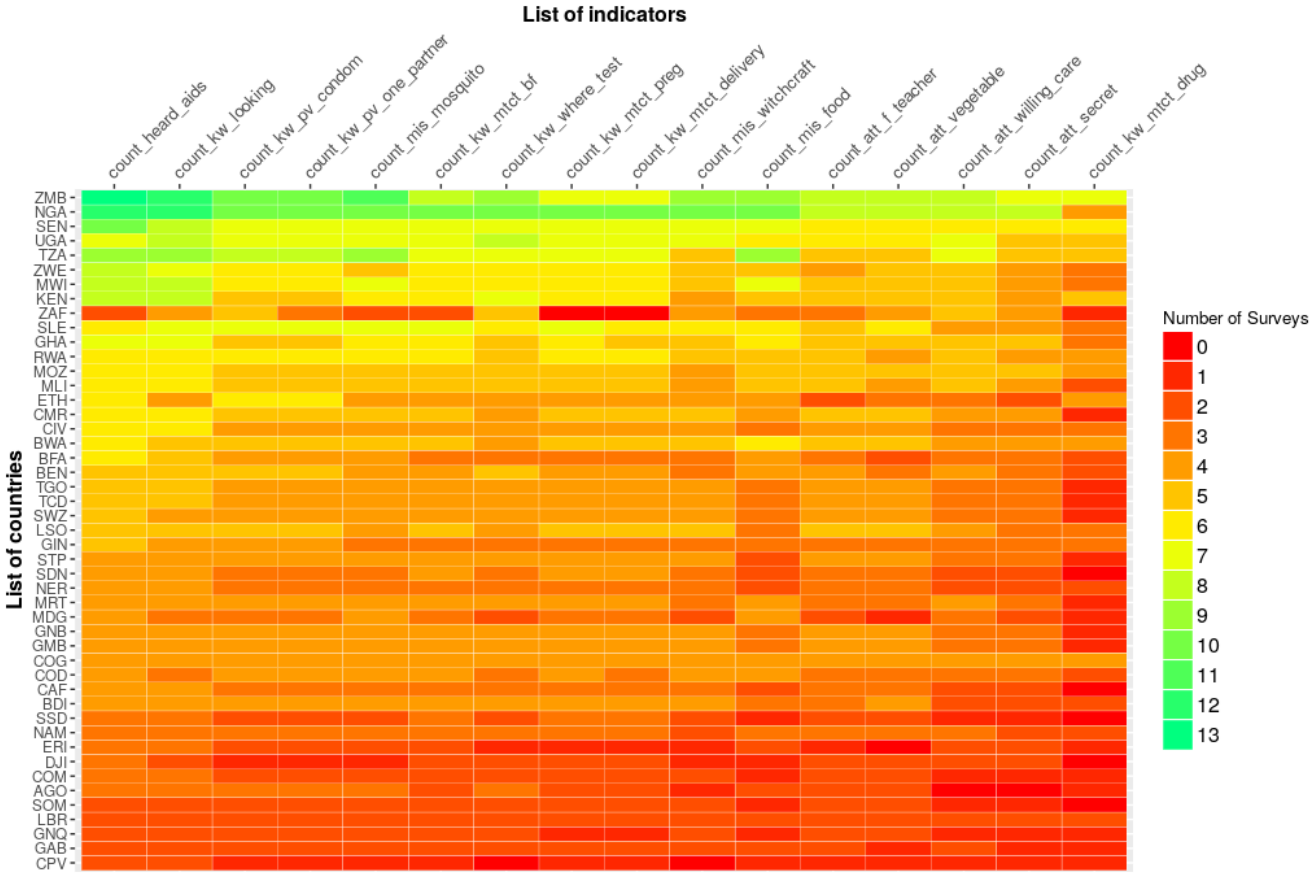


Figure 2-5 Indicator coverage by country

3.2 Crosswalking

As described above, we use crosswalking to deal with the data imbalance between men and women and between different age groups. ***Figure 2-6*** and ***Figure 2-7*** show the scatter plots of the logit-transformed data of four selected indicators for men and women and for age group 15-24 and 25-49 respectively. We can see that the relationships of logit-transformed data between men and women and between age group 15-24 and 25-49 are very likely to be linear, substantiating the mixed effect linear model used for the crosswalking. **Table 4** lists the predicted root mean squared error (RMSE) of all the crosswalking models for all indicators. All the models have RMSE < 0.05, suggesting that the crosswalking models fit the data very well and the predicted values of the model are very close to the true values on average.

It is worth mentioning that the regression coefficient of the interaction term $logit(I_{f,k}) * year$ in the crosswalking models is informative of the change in the relationship between data of different genders or age groups. If this coefficient is significant, i.e. p-value< 0.05, it is likely that there is significant change in the relationship between data of different genders or age groups over time. **Table 2-4** shows the p-values of the interaction for all indicators. For crosswalking from women and men, this coefficient is significant for *heard_aids, kw_looking, kw_where_test,* and *kw_mtct_preg*. For crosswalking from ages 15-24 to 25-49, this coefficient is significant for all indicators except *kw_pv_one_partner*, *kw_mtct_bf,* and *att_willing_care*. The scatter plots of the data for all indicators and the detailed results of the crosswalking models are included in the Appendix.

Table 2-4 The RMSE and p-value of the interaction term of crosswalking models from women to men and from 15-24 to 25-49 for all indicators

Indicators	**From women to men**		**From 15-24 to 25-49**	
	Predicted RMSE	**p-value of the interaction**	**Predicted RMSE**	**p-value of the interaction**
heard_aids	0.0121	**<0.001**	0.0258	**<0.001**
kw_pv_condom	0.0321	0.140	0.0305	**<0.001**
kw_pv_one_partner	0.0319	0.374	0.0375	0.074
kw_looking	0.0299	**0.017**	0.0343	**<0.001**
kw_where_test	0.0451	**0.025**	0.0465	**<0.001**
kw_mtct_preg	0.0441	**<0.001**	0.0329	**0.041**
kw_mtct_delivery	0.0398	0.076	0.0295	**<0.001**
kw_mtct_bf	0.0450	0.492	0.0258	0.220
kw_mtct_drug	0.0441	0.564	0.0306	**<0.001**
mis_mosquito	0.0307	0.810	0.0331	**0.036**
mis_food	0.0267	0.939	0.0325	**<0.001**
mis_witchcraft	0.0303	0.743	0.0310	**<0.001**
att_vegetable	0.0282	0.539	0.0334	**<0.001**
att_secret	0.0424	0.104	0.0274	**<0.001**

att_willing_care	0.0344	0.100	0.0219	0.646
att_f_teacher	0.0319	0.153	0.0338	**<0.001**

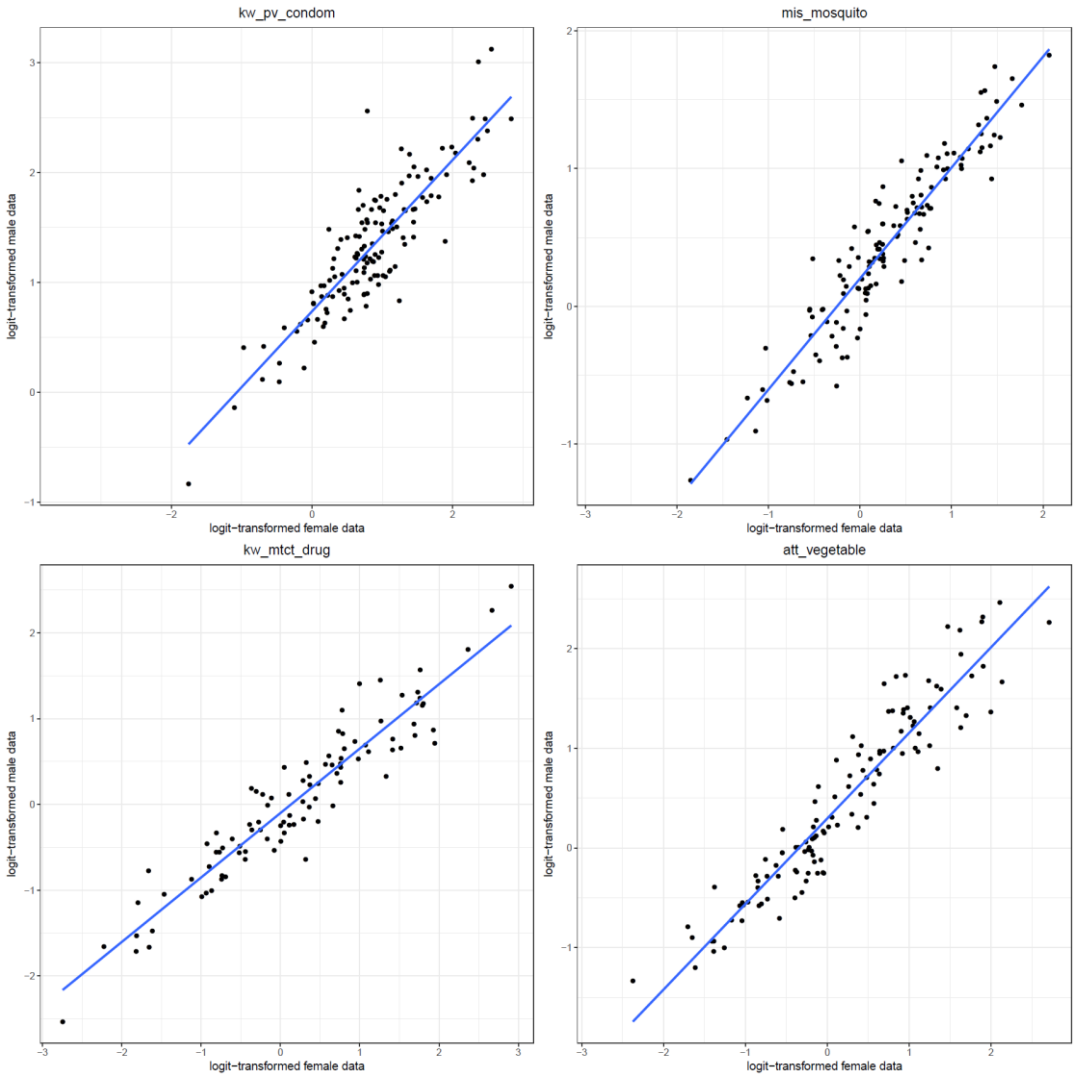


Figure 2-6 Scatter plots of data for men and women for four indicators

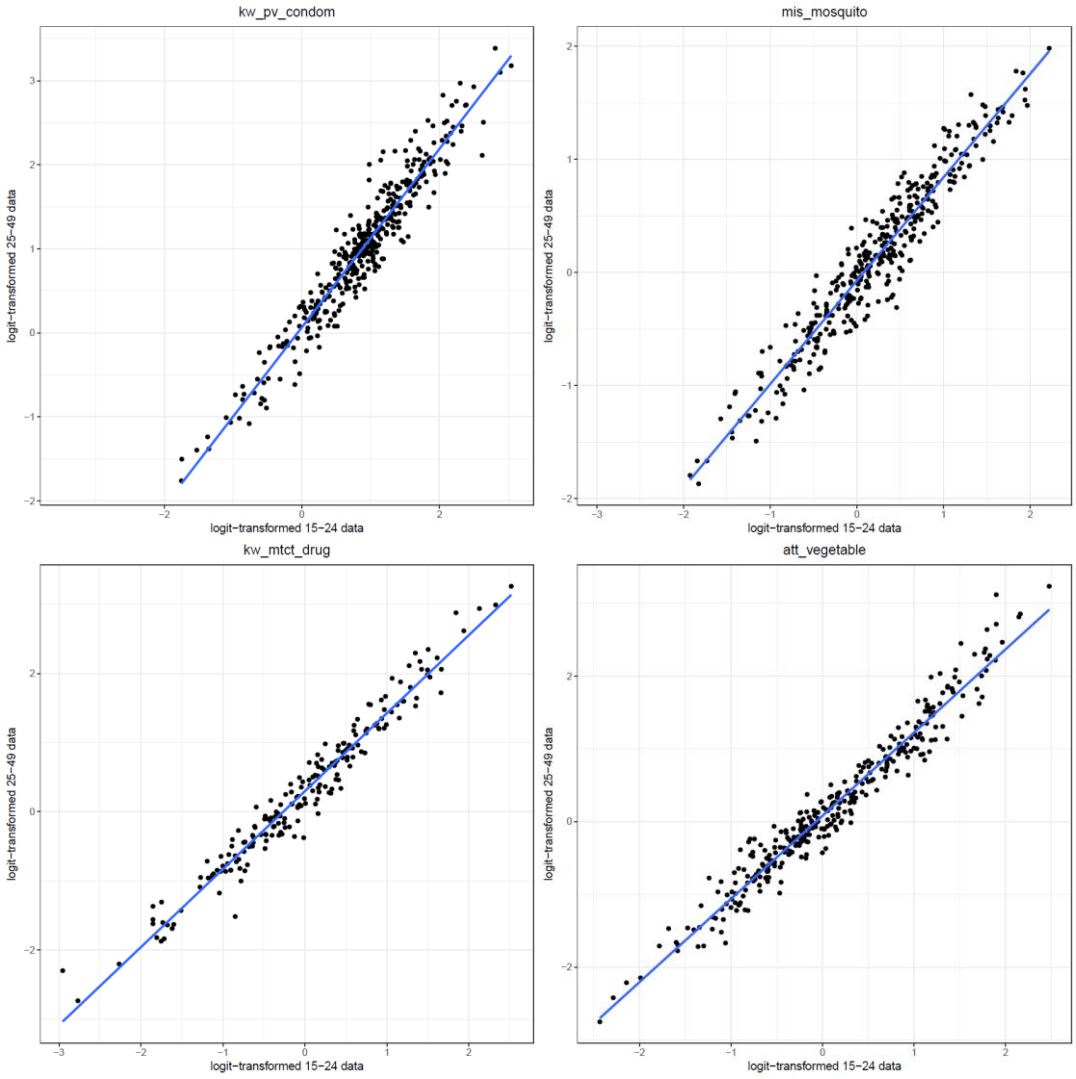


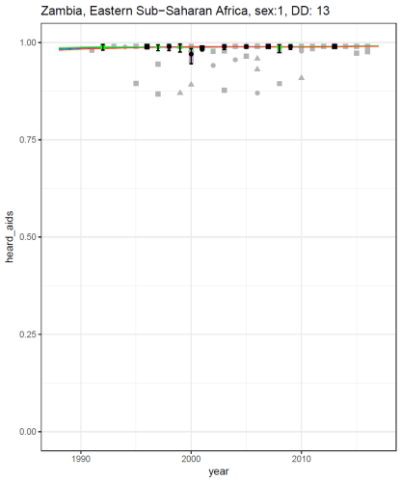
Figure 2-7 Scatter plots of data for age groups 15-24 and 25-49 for four indicators

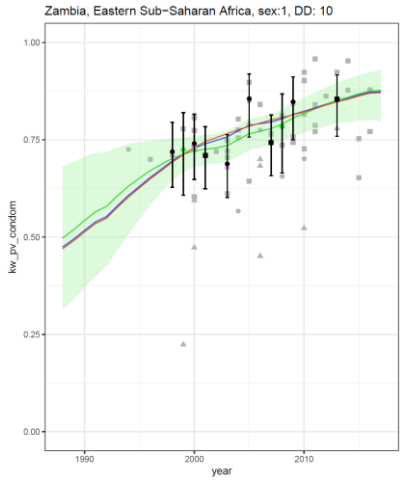
3.3 Country specific trends of the indicators for age group 15-49

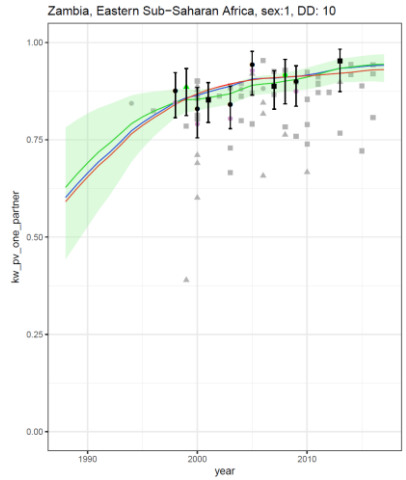
We first ran ST-GPR on the bias-adjusted and gender-crosswalked data for age group 15-49. The results are a country- and sex-specific time series of each indicator from 1988 to 2017 for all 47 SSA countries. Therefore, there are a total of 1,504 country-sex-indicator specific time series for age group 15-49. Full results of the 1,504 time series including graphs and tables are in the Appendix. ***Figure 2-8*** shows the sex-indicator specific trends plots for Zambia, which has the most data sources among all countries. The black, grey, green, and red points are country data,

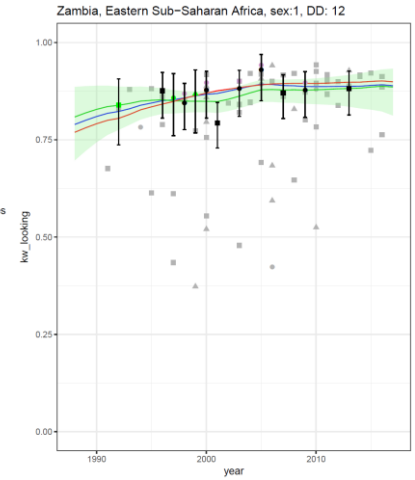
regional data, crosswalked data, and removed outliers, respectively. Different shapes of the points represent different survey types, with square, triangle, and circle representing DHS/AIS, MICS, and other surveys, respectively. The red, blue, and green lines are 1st, 2nd (ST), and final ST-GPR estimates, respectively, with the credible interval of the final estimates shown by light green shade.

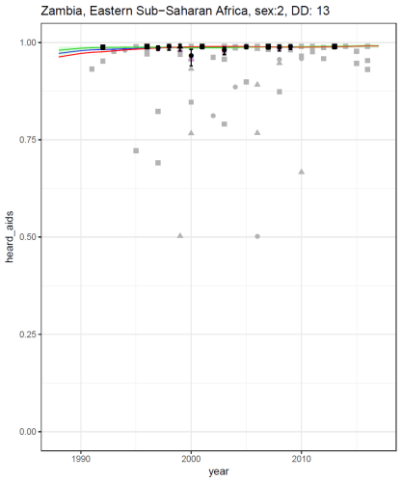
From ***Figure 2-8***, we can see that the trends of all indicators have been improving over time in Zambia, except for *kw_mtct_preg* and *att_secret*, which have been significantly deteriorating since 2000 when data became available. The unusual deteriorating trend of *kw_mtct_preg* can also be found in many other countries including Kenya, Tanzania, Rwanda, Ghana and Nigeria, where PMTCT programs have been successful. The decreasing trend of the attitudes indicator *att_secret* is more prevalent. The indicator has been deteriorating in almost all countries since 2000 when data became available. ***Figure 2-9*** and ***Figure 2-10*** show some country examples of these two indicators respectively. Full results of country-specific trends are in the appendix.

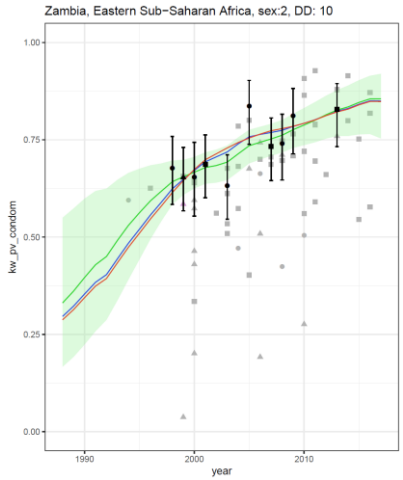

Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 13
heard_aids
year
survey types
DHS/AIS
MICS
Other

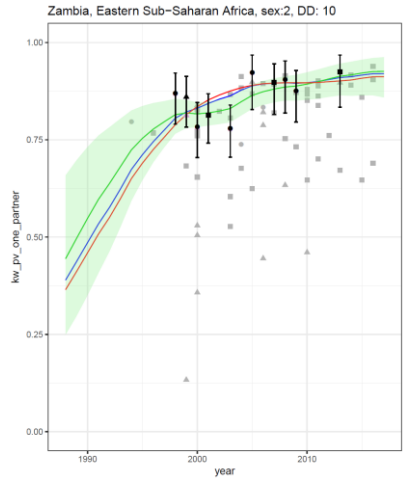

Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 10
kw_pv_condom
year
survey types
DHS/AIS
MICS
Other

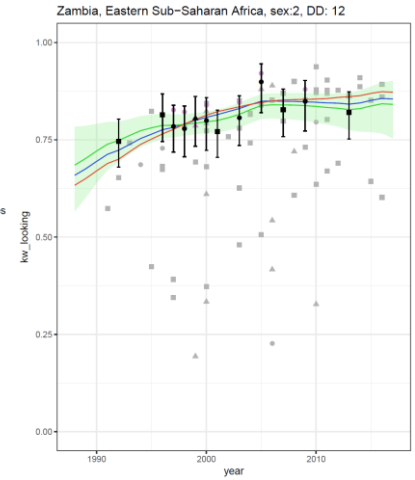

Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 10
kw_pv_one_partner
year
survey types
DHS/AIS
MICS
Other

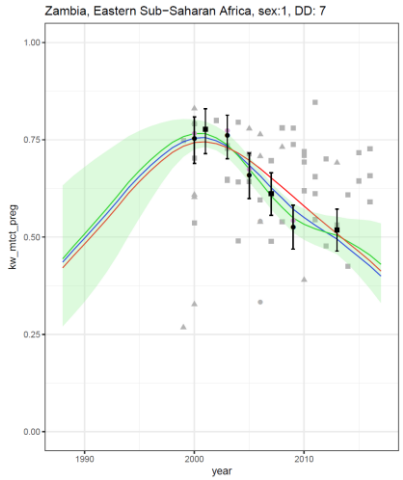

Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 12
kw_looking
year
survey types
DHS/AIS
MICS
Other

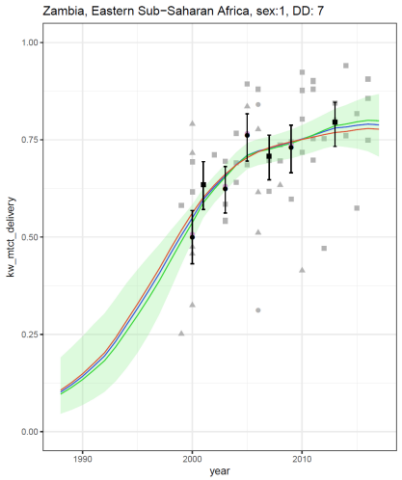

Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 13
heard_aids
year
survey types
DHS/AIS
MICS
Other

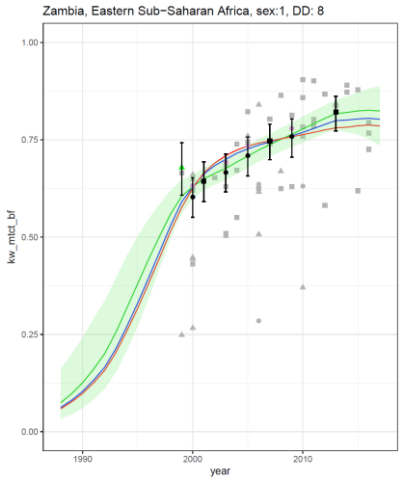

Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 10
kw_pv_condom
year
survey types
DHS/AIS
MICS
Other

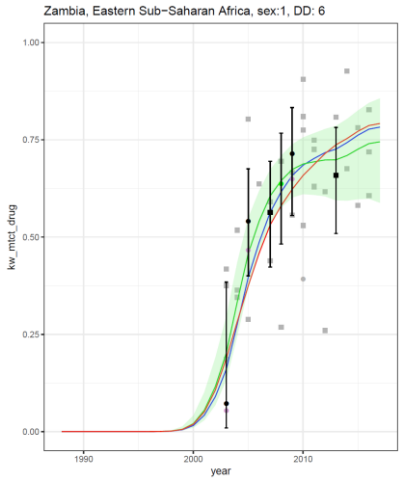

Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 10
kw_pv_one_partner
year
survey types
DHS/AIS
MICS
Other

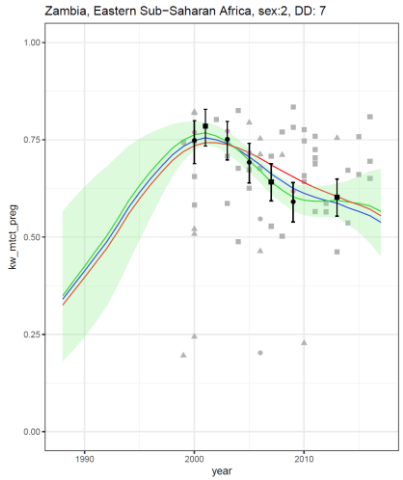

Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 12
kw_looking
year
survey types
DHS/AIS
MICS
Other

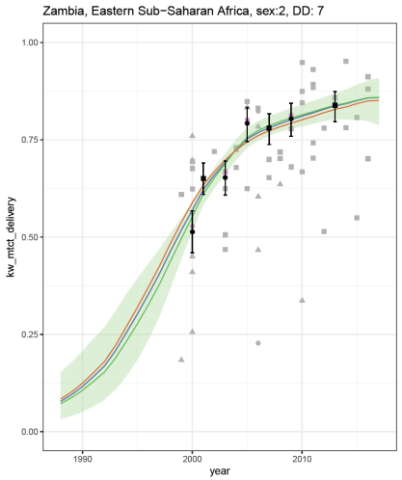

Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 7
kw_mtct_preg
year
survey types
DHS/AIS
MICS
Other

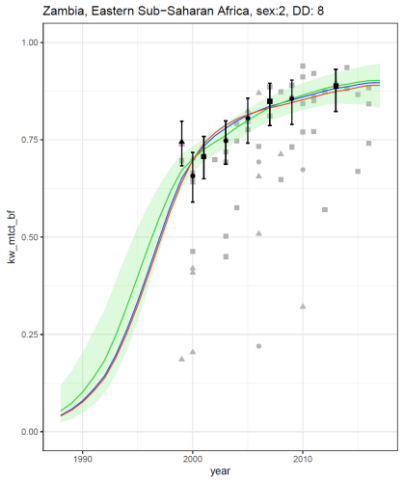

Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 7
kw_mtct_delivery
year
survey types
DHS/AIS
MICS
Other

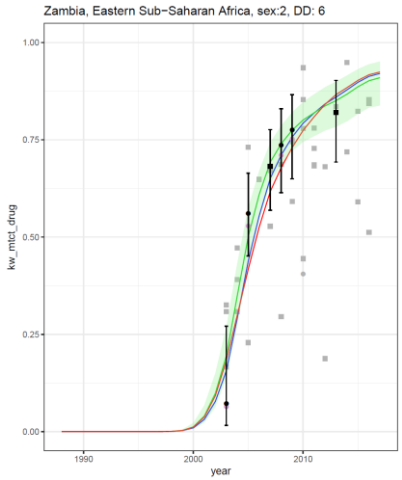

Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 8
kw_mtct_bf
year
survey types
DHS/AIS
MICS
Other

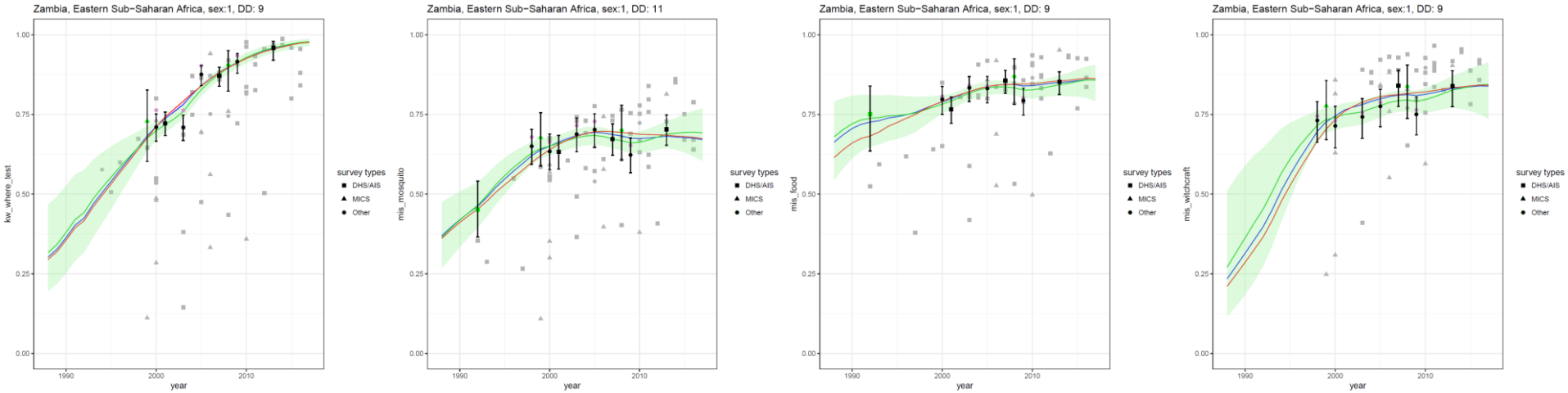

Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 6
kw_mtct_drug
year
survey types
DHS/AIS
Other

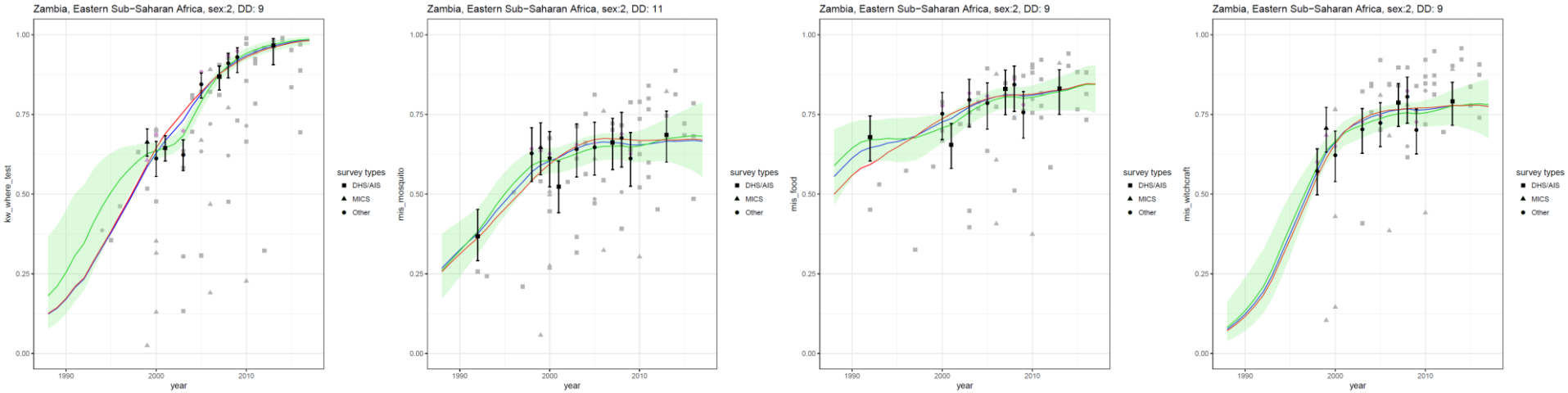

Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 7
kw_mtct_preg
year
survey types
DHS/AIS
MICS
Other


Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 7
kw_mtct_delivery
year
survey types
DHS/AIS
MICS
Other


Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 8
kw_mtct_bf
year
survey types
DHS/AIS
MICS
Other


Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 6
kw_mtct_drug
year
survey types
DHS/AIS
Other


Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 9
kw_where_test
year
survey types
DHS/AIS
MICS
Other
Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 11
mis_mosquito
Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 9
mis_food
Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 9
mis_witchcraft


Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 9
kw_where_test
year
survey types
DHS/AIS
MICS
Other
Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 11
mis_mosquito
Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 9
mis_food
Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 9
mis_witchcraft

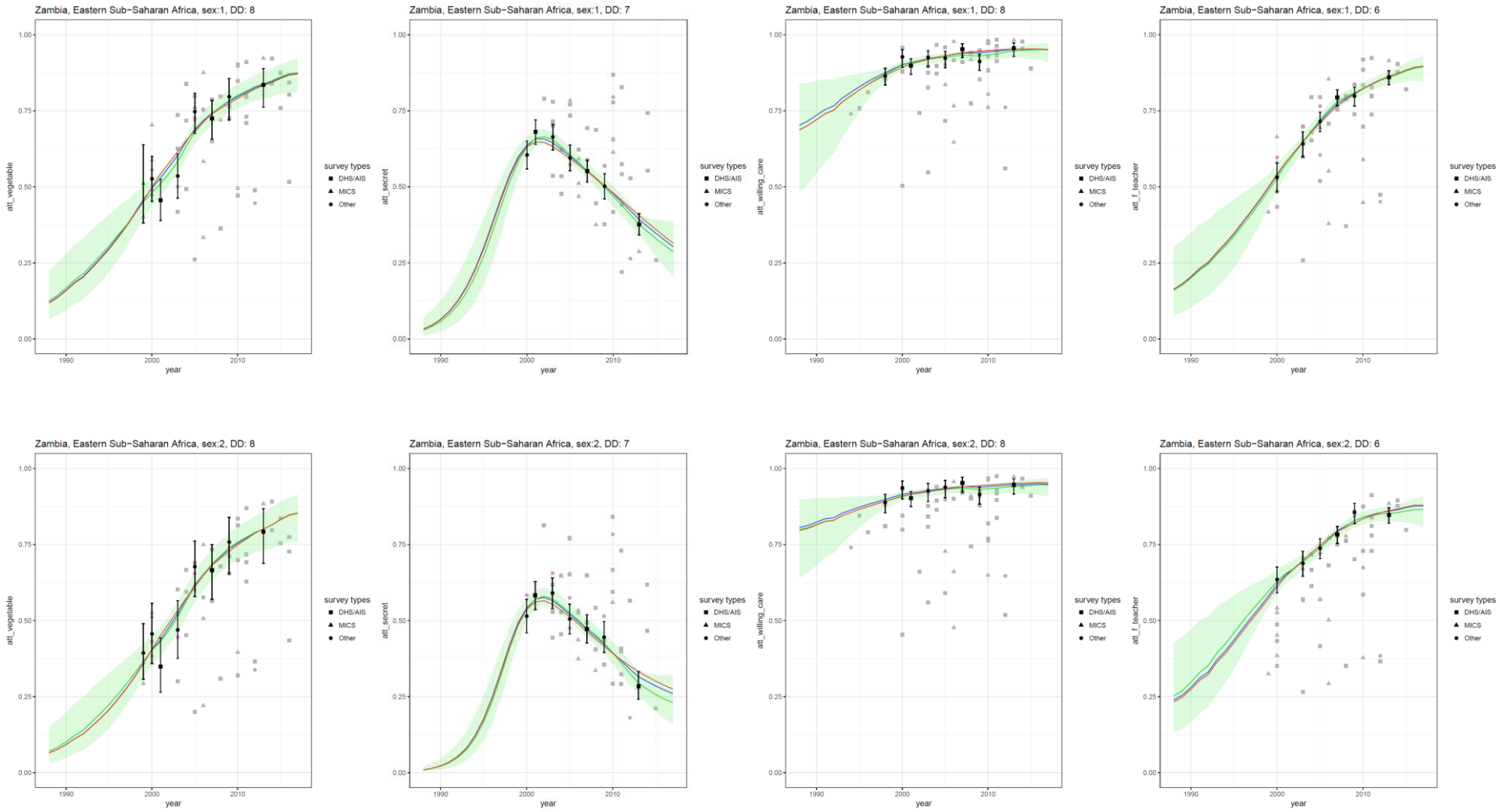


Figure 2-8 Sex-and-indicator specific trends for people aged 15-49 in Zambia

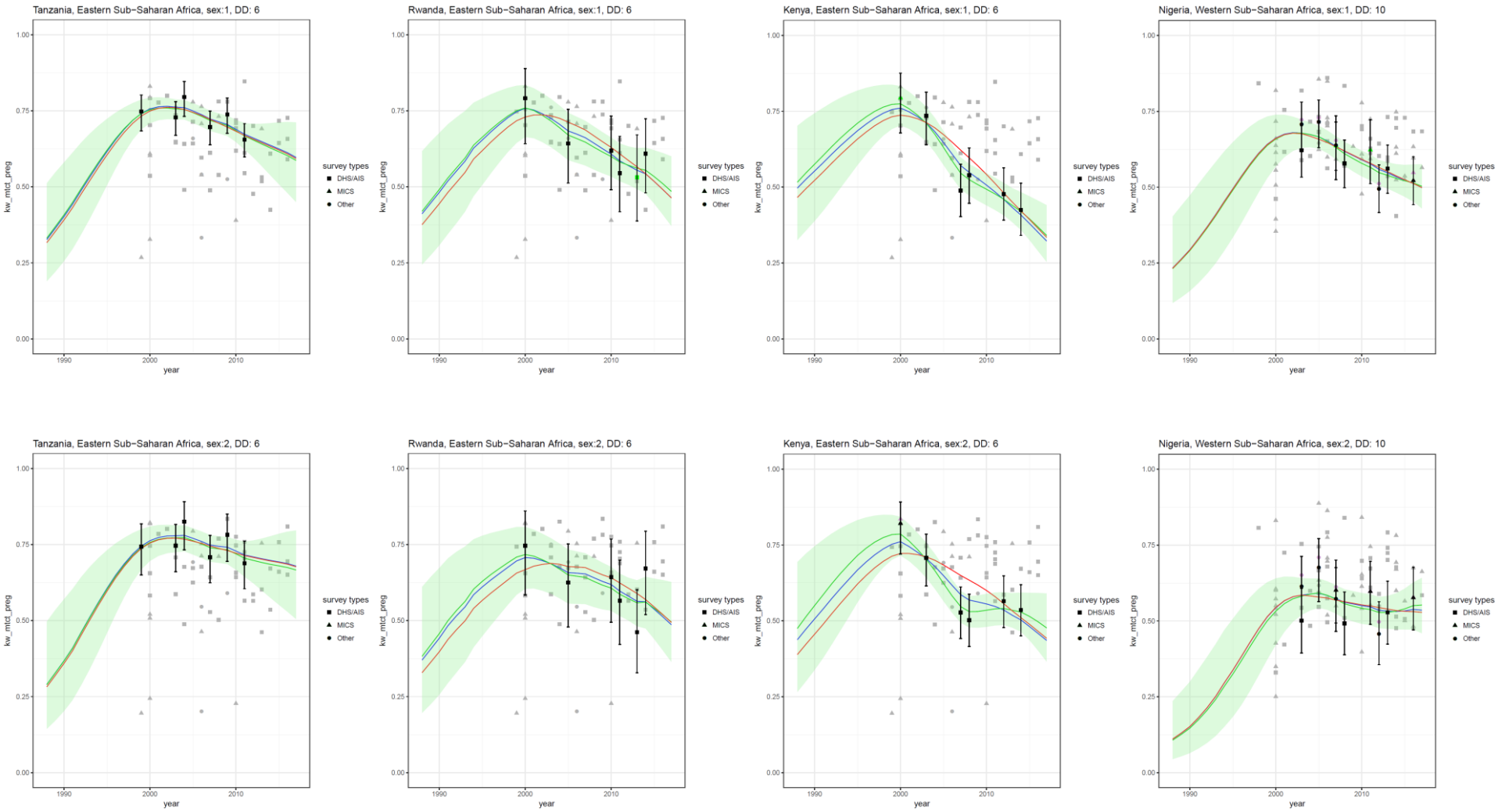


Figure 2-9 Country examples of trends of kw_mtc_preg

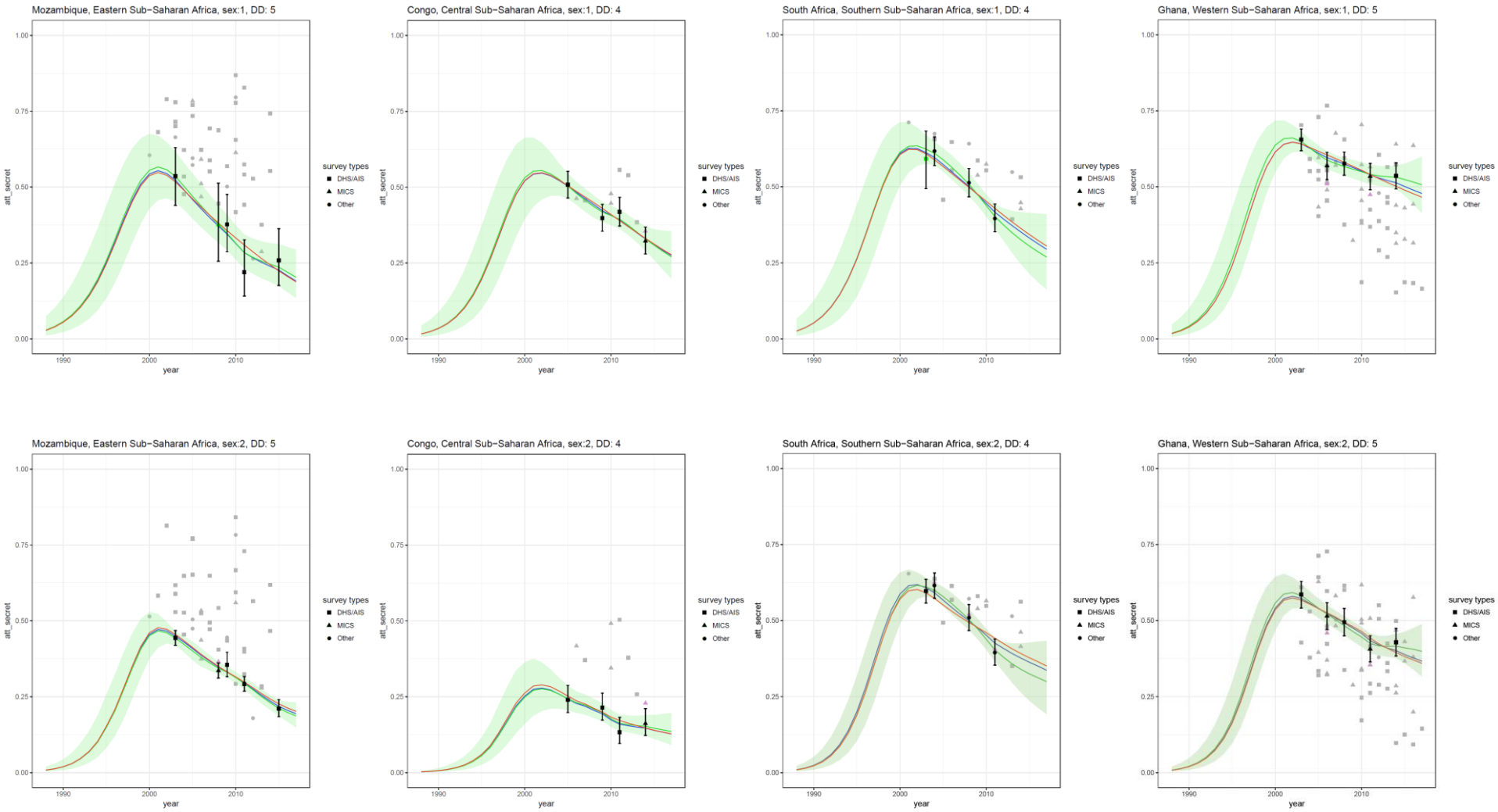


Figure 2-10 Country examples of trends of att_secret

3.4 Mean trends of the indicators for people aged 15-49 by gender and by region

After analyzing the country-sex specific trends of each indicator for people aged 15-49, we estimated the mean trends of the indicators by gender and by region using a generalized additive model. ***Figure 2-12*** shows the mean trends of the indicators by gender, by subregion, and by age group, respectively. Each transparent thin line in the back represents the final ST-GPR estimates of the indicator for a specific sex in a country. The thick solid lines in the graphs are the estimated mean trends of the indicators for men and women, for each subregion, and for each age group respectively.

From ***Figure 2-12***, we can see that the mean trends of the indicators for both men and women have been generally increasing over the past three decades except for indicators *kw_mtct_preg*, *kw_mtct_delivery*, and *att_secret*, which have been decreasing over time. For *kw_mtct_preg* and *kw_mtct_delivery*, the deteriorating trend is more prominent for men than for women, but for *att_secret*, this attitudes indicator has been significantly decreasing for both gender since 2000. Except for knowledge on MTCT, men generally do better than women on other HIV/AIDS knowledge and attitudes indicators. However, the difference between men and women on these indicators has been decreasing over time especially for *heard_aids*, *kw_looking*, *kw_one_partner*, *kw_where_test* and *mis_mosquito*, suggesting that women have been catching up with men over time. For knowledge indicators on MTCT, namely, *kw_mtct_preg*, *kw_mtct_delivery*, *kw_mtct_bf* and *kw_mtct_drug*, women's knowledge started low in early years but has caught up with and surpassed that of men's since 2000, and the gap between women and men has increased thereafter.

Regarding the mean trends of the indicators by subregion, ***Figure 2-12*** shows that the HIV/AIDS knowledge and attitudes indicators have been generally improving over the past three decades in

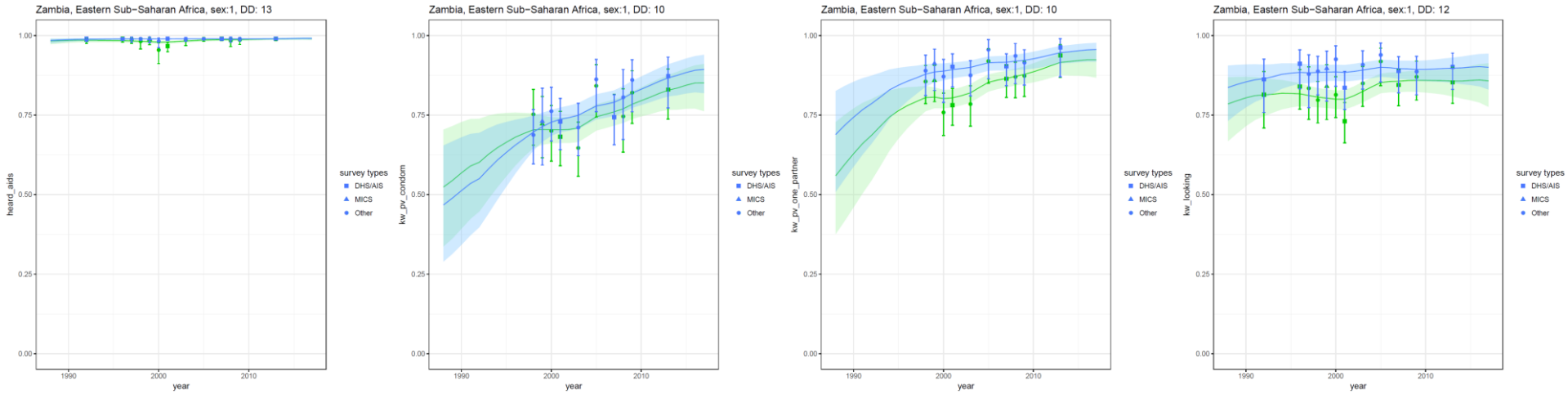
all subregions of SSA expect for *kw_mtct_preg*, *kw_mtct_delivery*, and *att_secret*, which have been deteriorating over time across all subregions. Among the four subregions (northern and eastern SSA are combined into one subregion), people in southern SSA tend to do significantly better on all indicators than people in the other subregions except for *att_secret*, of which people in southern SSA have almost the same level with people in Eastern SSA. In addition, the improving rates of indicators *kw_where_test, kw_mtct_drug, att_vegetable, and att_f_teacher* for southern SSA are much higher than those for the other regions.

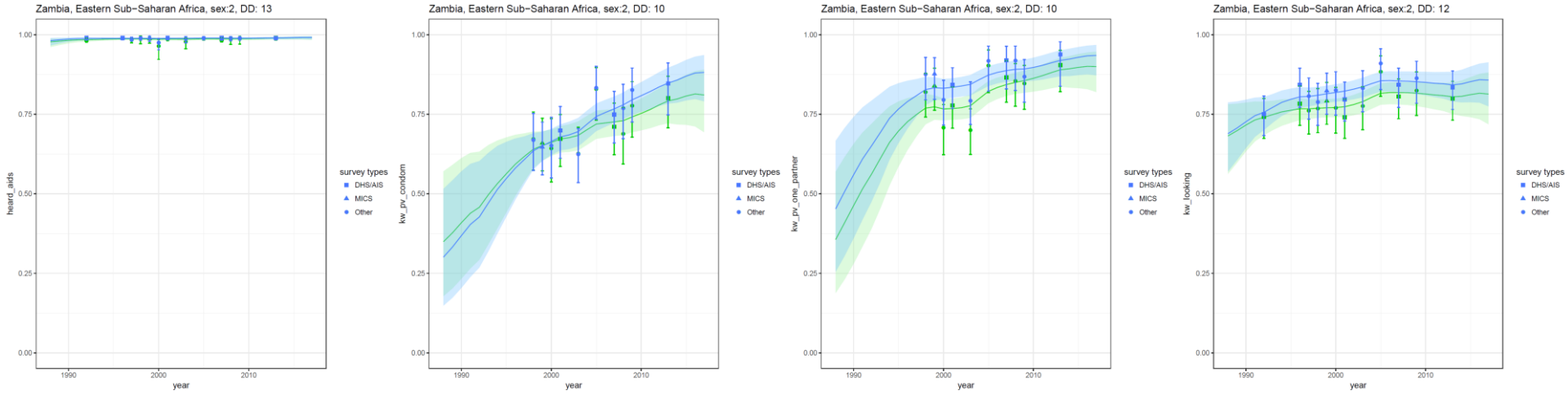
3.5 Mean trends of the indicators by age groups

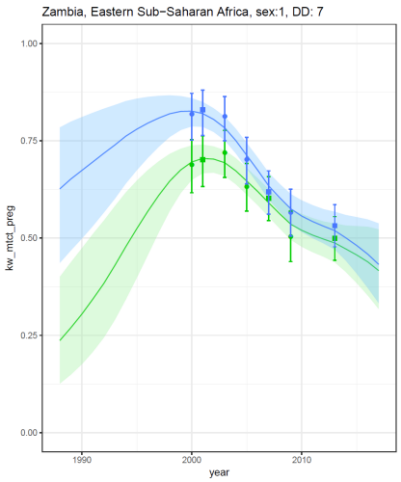
To explore the difference in HIV/AIDS knowledge and attitudes between young and older people over time, we estimated the country- and sex-specific trends of the indicators for people aged 15-24 and 25-49, respectively. The full results of the country-sex specific trends by two age groups are included in the Appendix. **Figure 11** shows the trends of the indicators by age groups in Zambia. The green lines and dots are for young people aged 15-24 and the blues ones are for people aged 25-49. We can see that in Zambia, older people tend to do better than young people do for both genders on all indicators but *mis_mosquito,* of which young people have slightly better knowledge than older people do.

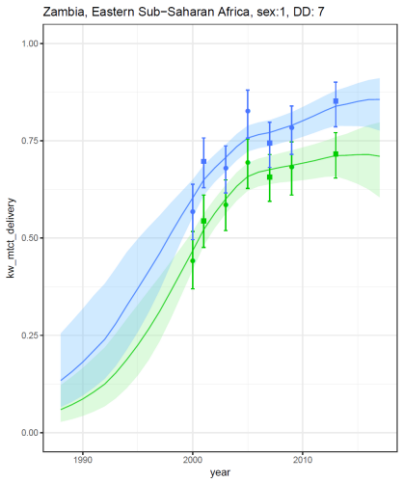
After obtaining the country-specific trends, we estimate the mean trends of the indicators in SSA by the two age groups using a generalized additive model. **Figure 12** shows the mean trends of the indicators for age group 15-24 and 25-49 respectively. Overall, the mean trends of the indicators have been increasing for both genders over time except for *kw_mtct_preg* and att_secret, which have been decreasing in recent years. People aged 25-49 generally have better knowledge and attitudes than people aged 15-24, especially for *kw_where_test, kw_pv_condom, kw_pv_one_partner, kw_looking,* and four MTCT indicators, of which older people have

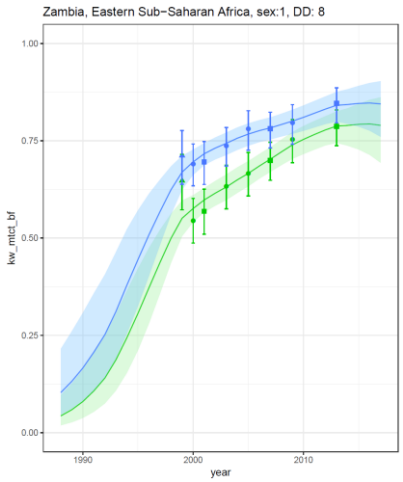
significantly better knowledge than young people do. Regarding attitudes, older people tend to have a slightly better attitudes toward PLWHs than young people do. It is also worth noting that the knowledge gap of *kw_looking*, *kw_pv_condom*, *kw_where_test*, *kw_mtct_delivery*, and *kw_mtct_drug* between older and young people are significantly widening over time. For attitudes indicators, the gap of *att_vegetable* and *att_f_teacher* between older and young people also seems to be widening slightly.

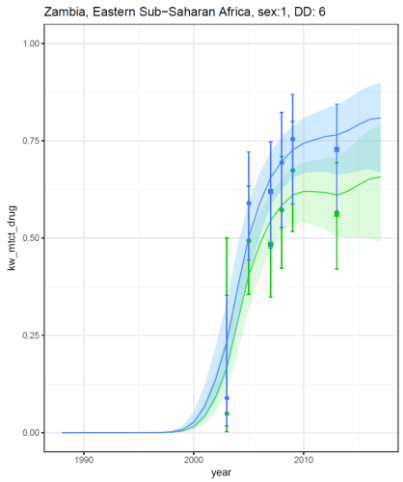

Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 13
heard_aids
year
survey types
DHS/AIS
MICS
Other
Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 10
kw_pv_condom
Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 10
kw_pv_one_partner
Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 12
kw_looking

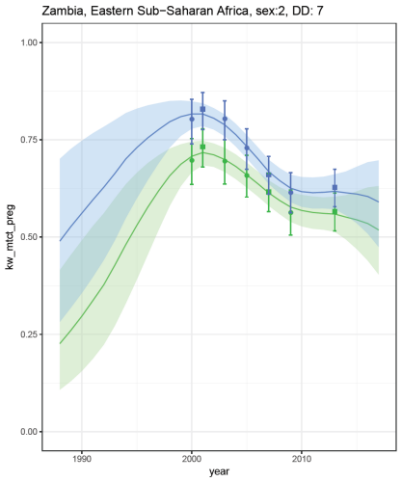

Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 13
heard_aids
year
survey types
DHS/AIS
MICS
Other
Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 10
kw_pv_condom
Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 10
kw_pv_one_partner
Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 12
kw_looking

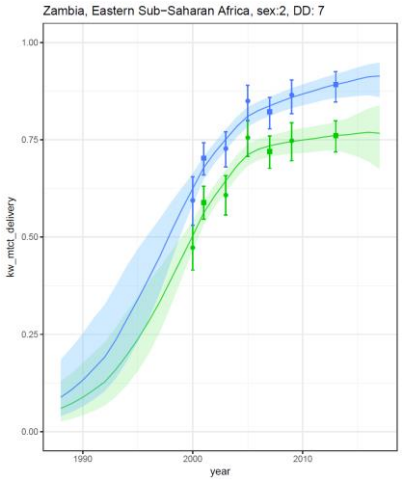

Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 7
kw_mtct_preg
year
survey types
DHS/AIS
Other

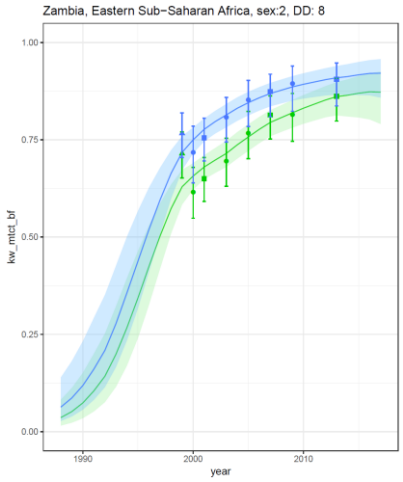

Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 7
kw_mtct_delivery
year
survey types
DHS/AIS
Other

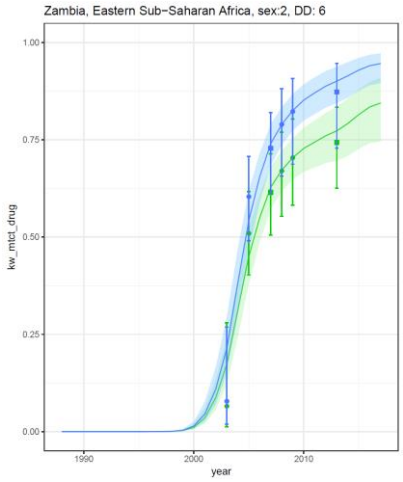

Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 8
kw_mtct_bf
year
survey types
DHS/AIS
MICS
Other

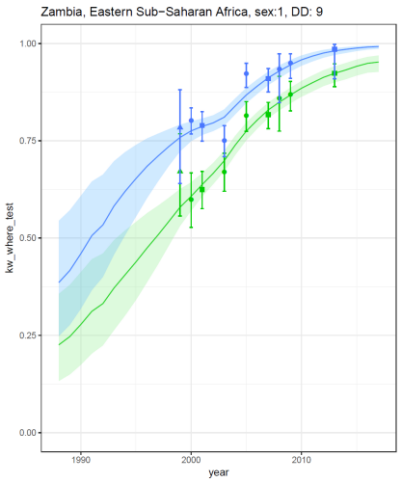

Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 6
kw_mtct_drug
year
survey types
DHS/AIS
Other

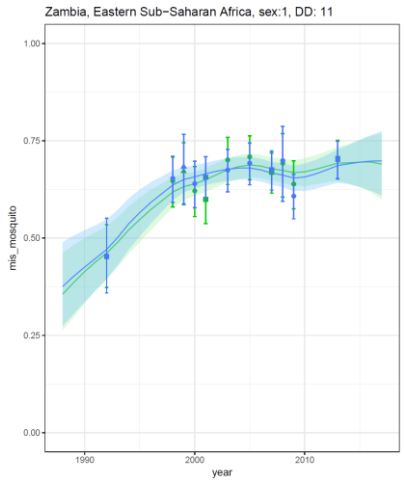

Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 7
kw_mtct_preg
year
survey types
DHS/AIS
Other

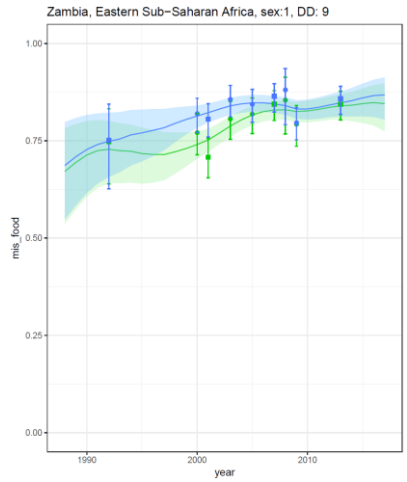

Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 7
kw_mtct_delivery
year
survey types
DHS/AIS
Other

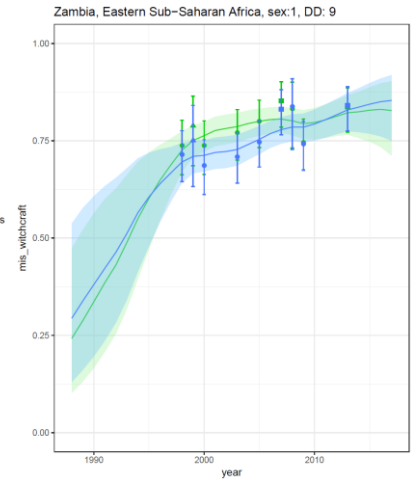

Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 8
kw_mtct_bf
year
survey types
DHS/AIS
MICS
Other

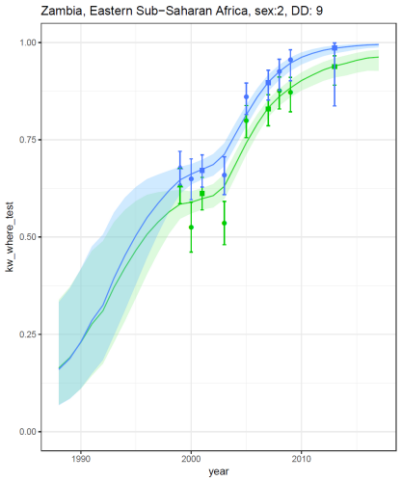

Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 6
kw_mtct_drug
year
survey types
DHS/AIS
Other

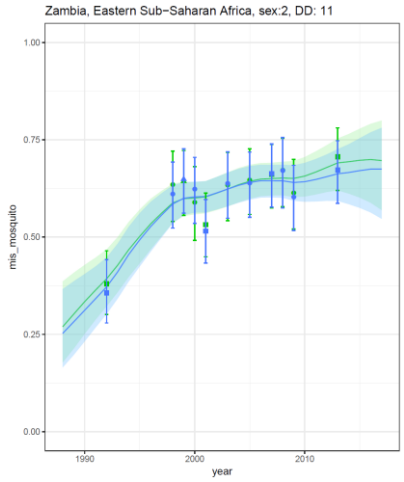

Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 9
kw_where_test
year
survey types
DHS/AIS
MICS
Other

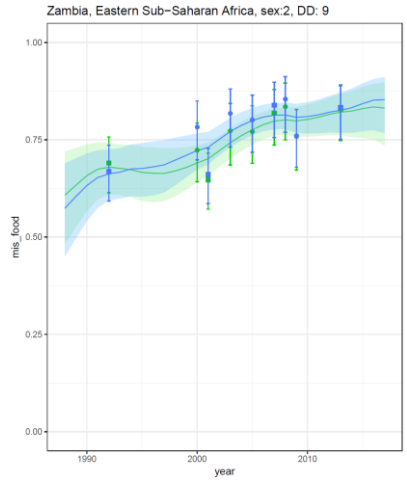

Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 11
mis_mosquito
year
survey types
DHS/AIS
MICS
Other

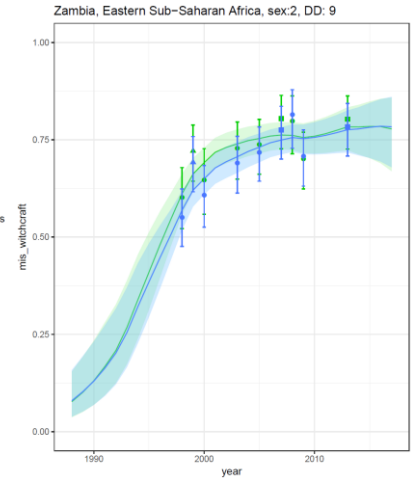

Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 9
mis_food
year
survey types
DHS/AIS
Other


Zambia, Eastern Sub−Saharan Africa, sex:1, DD: 9
mis_witchcraft
year
survey types
DHS/AIS
MICS
Other


Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 9
kw_where_test
year
survey types
DHS/AIS
MICS
Other


Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 11
mis_mosquito
year
survey types
DHS/AIS
MICS
Other


Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 9
mis_food
year
survey types
DHS/AIS
Other


Zambia, Eastern Sub−Saharan Africa, sex:2, DD: 9
mis_witchcraft
year
survey types
DHS/AIS
MICS
Other

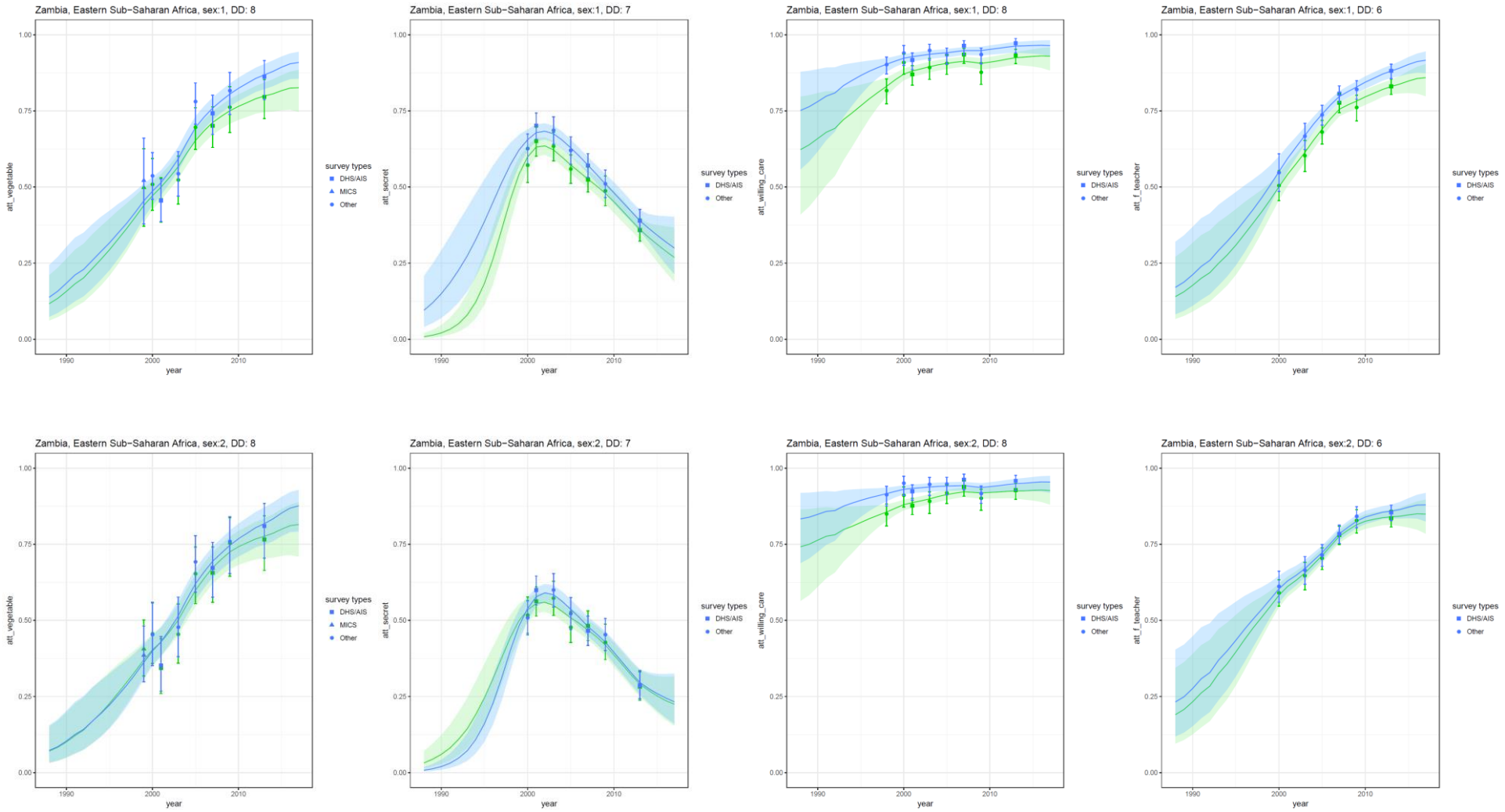
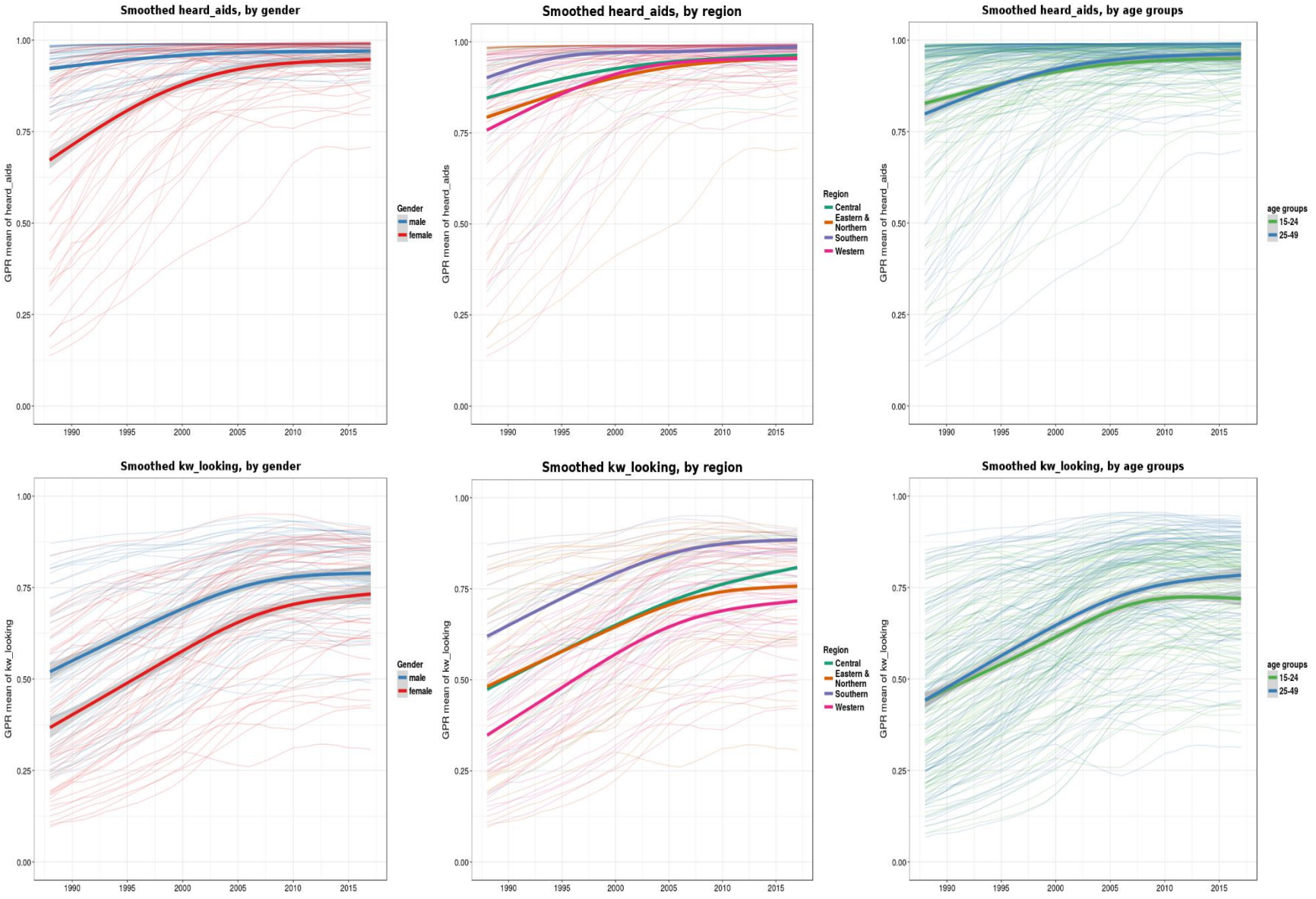
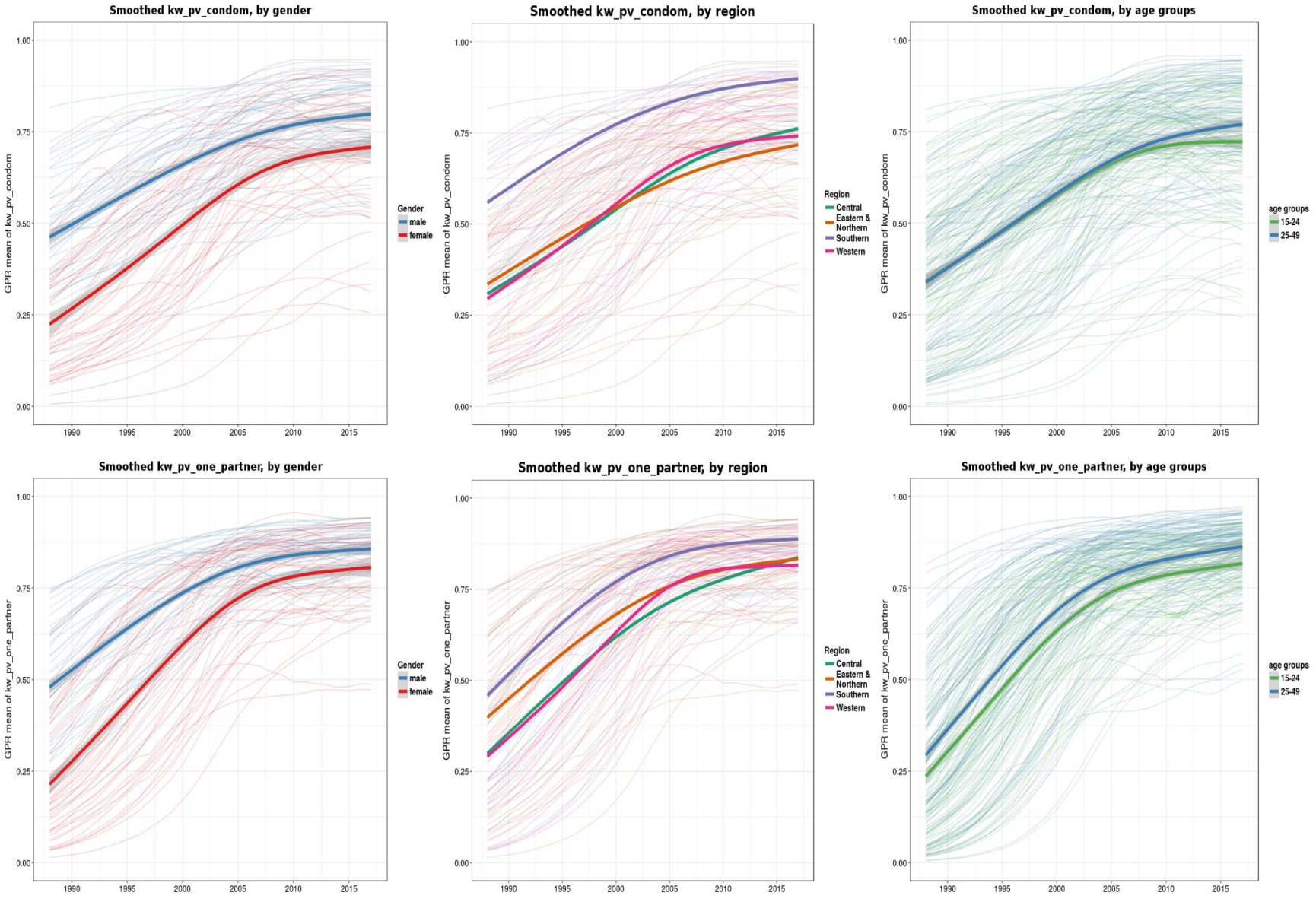
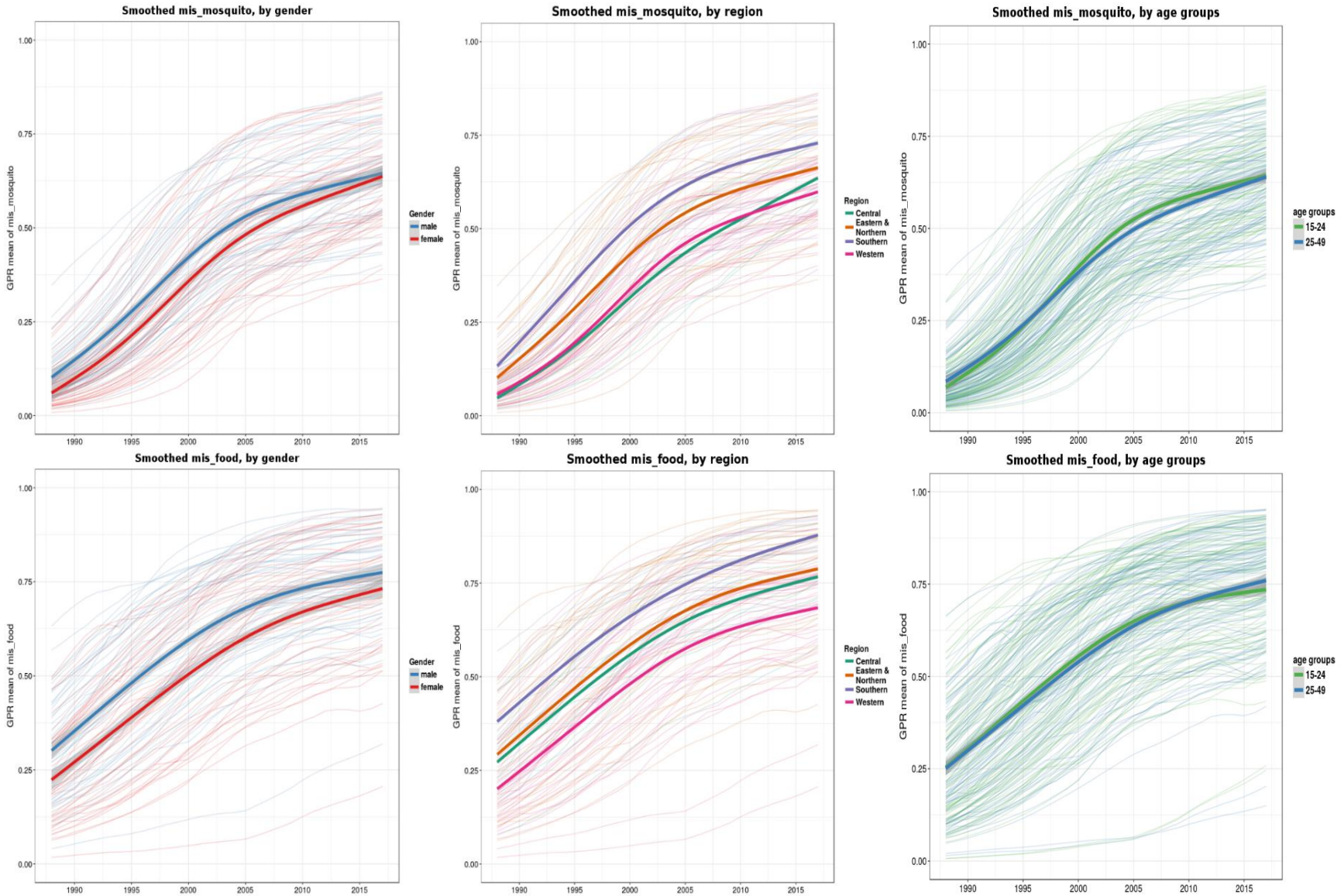
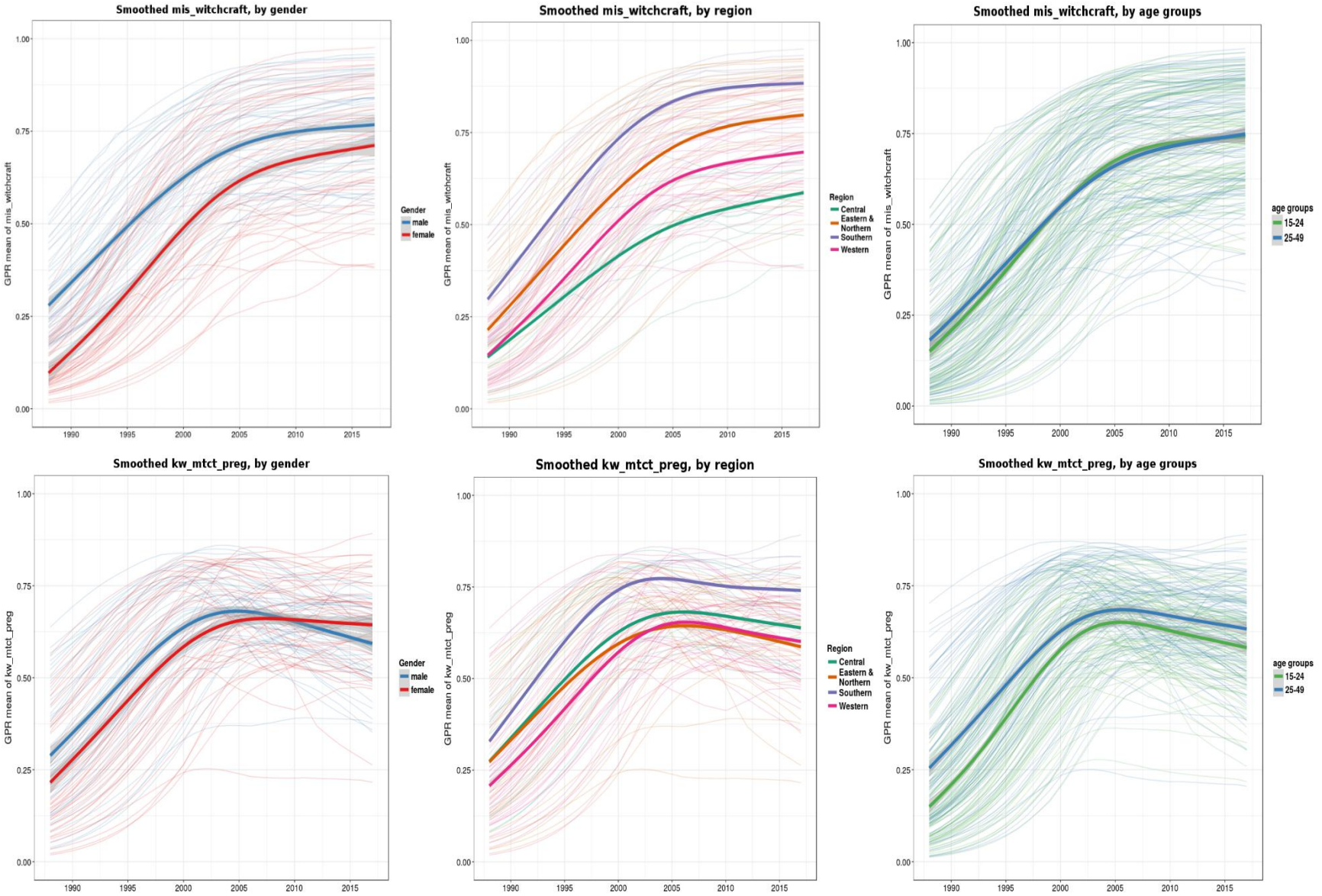


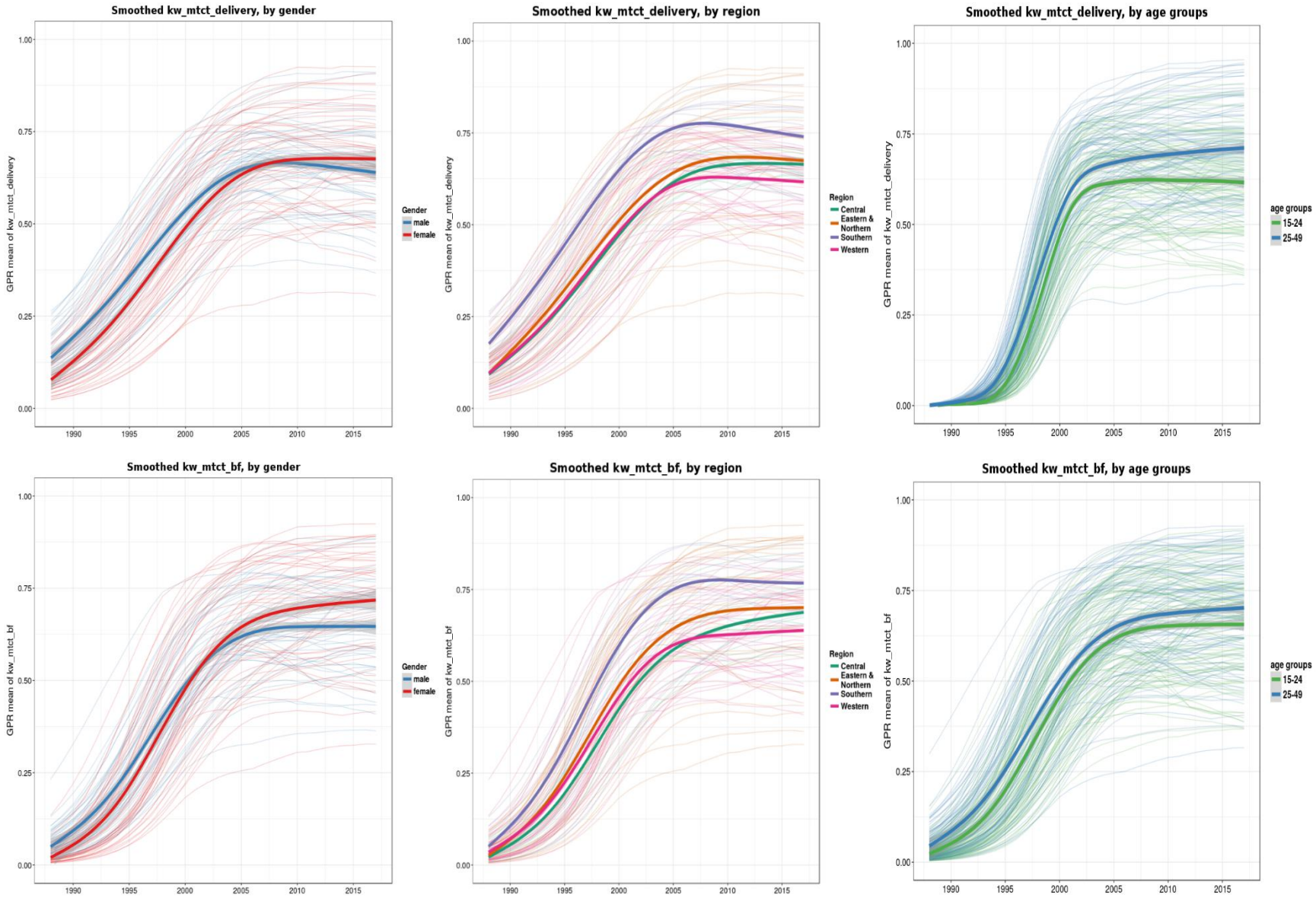
Figure 2-11 Trends of the indicators for age groups 15-24 and 25-49 in Zambia

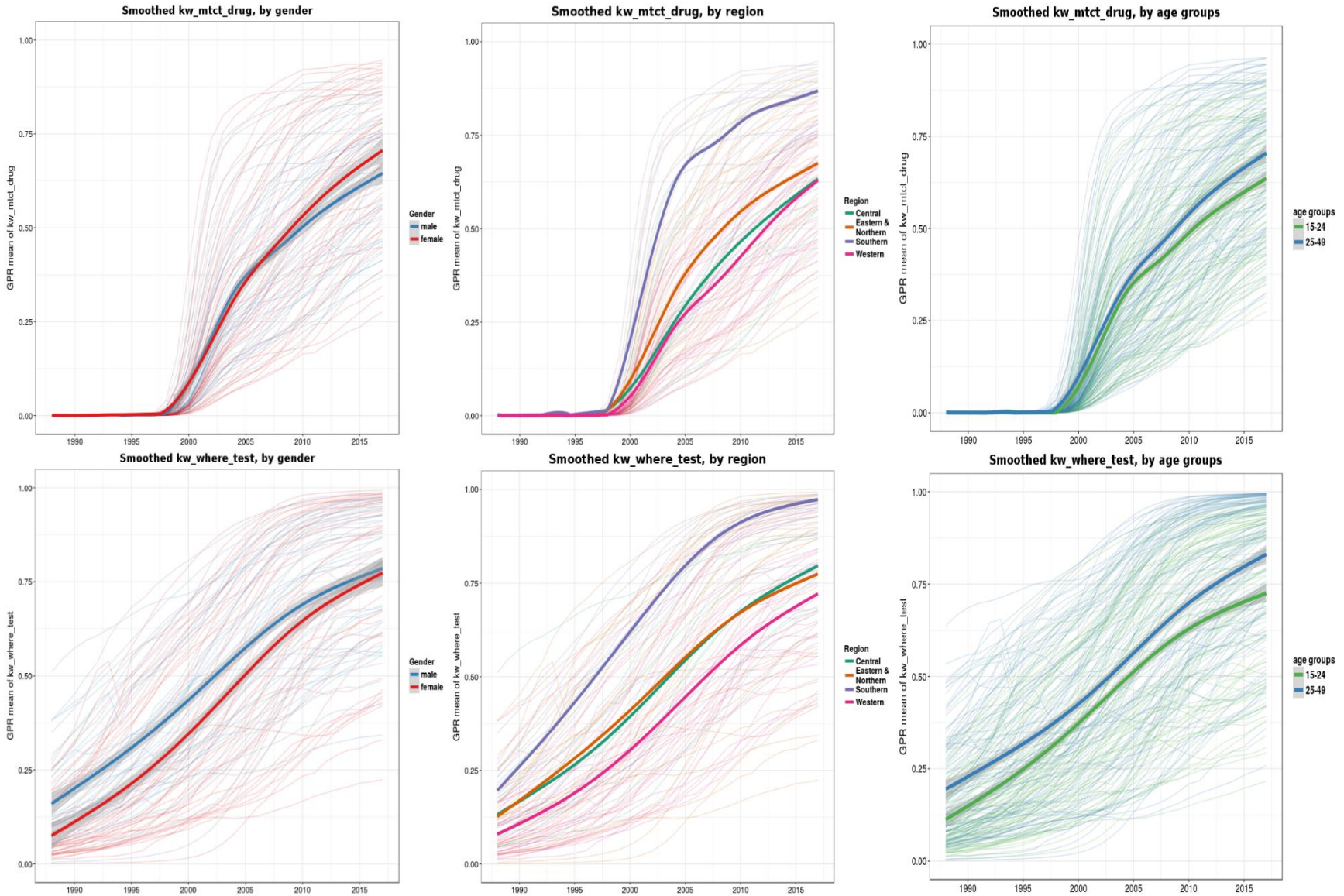

Smoothed heard_aids, by gender
GPR mean of heard_aids
1.00
0.75
0.50
0.25
0.00
1990
1995
2000
2005
2010
2015
Gender
male
female
Smoothed heard_aids, by region
Region
Central
Eastern &
Northern
Southern
Western
Smoothed heard_aids, by age groups
age groups
15-24
25-49
Smoothed kw_looking, by gender
GPR mean of kw_looking
Smoothed kw_looking, by region
Smoothed kw_looking, by age groups

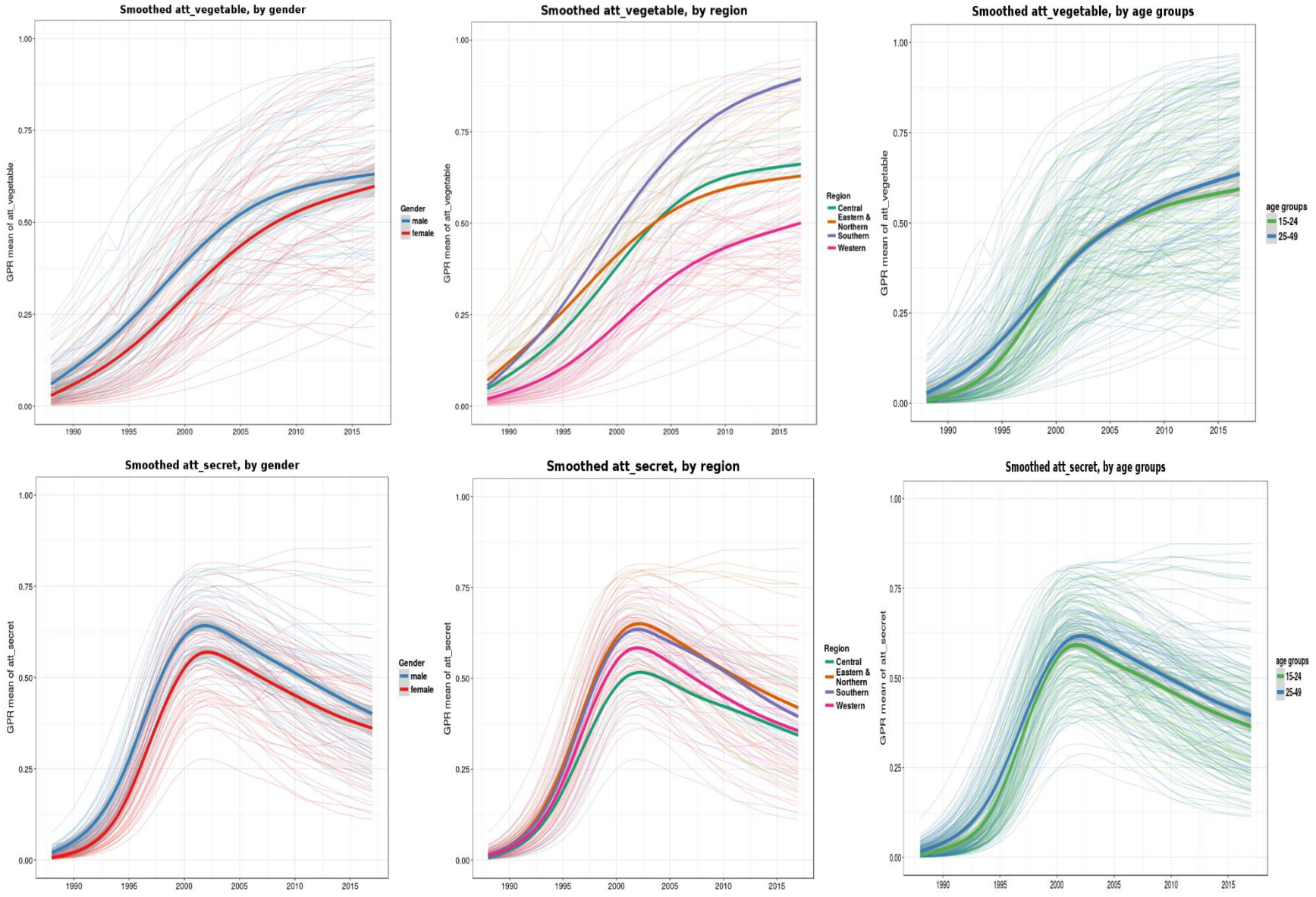

Smoothed kw_pv_condom, by gender
GPR mean of kw_pv_condom
Gender
male
female
Smoothed kw_pv_condom, by region
Region
Central
Eastern &
Northern
Southern
Western
Smoothed kw_pv_condom, by age groups
age groups
15-24
25-49
Smoothed kw_pv_one_partner, by gender
GPR mean of kw_pv_one_partner
Smoothed kw_pv_one_partner, by region
Smoothed kw_pv_one_partner, by age groups
1.00
0.75
0.50
0.25
0.00
1990
1995
2000
2005
2010
2015


Smoothed mis_mosquito, by gender
GPR mean of mis_mosquito
Gender
male
female
Smoothed mis_mosquito, by region
Region
Central
Eastern &
Northern
Southern
Western
Smoothed mis_mosquito, by age groups
age groups
15-24
25-49
Smoothed mis_food, by gender
GPR mean of mis_food
Smoothed mis_food, by region
Smoothed mis_food, by age groups
1.00
0.75
0.50
0.25
0.00
1990
1995
2000
2005
2010
2015


Smoothed mis_witchcraft, by gender
GPR mean of mis_witchcraft
Gender
male
female
Smoothed mis_witchcraft, by region
Region
Central
Eastern &
Northern
Southern
Western
Smoothed mis_witchcraft, by age groups
age groups
15-24
25-49
Smoothed kw_mtct_preg, by gender
GPR mean of kw_mtct_preg
Smoothed kw_mtct_preg, by region
Smoothed kw_mtct_preg, by age groups
1.00
0.75
0.50
0.25
0.00
1990
1995
2000
2005
2010
2015


Smoothed kw_mtct_delivery, by gender
GPR mean of kw_mtct_delivery
Gender
male
female
Smoothed kw_mtct_delivery, by region
Region
Central
Eastern &
Northern
Southern
Western
Smoothed kw_mtct_delivery, by age groups
age groups
15-24
25-49
Smoothed kw_mtct_bf, by gender
GPR mean of kw_mtct_bf
Smoothed kw_mtct_bf, by region
Smoothed kw_mtct_bf, by age groups
1990
1995
2000
2005
2010
2015
0.00
0.25
0.50
0.75
1.00


Smoothed kw_mtct_drug, by gender
GPR mean of kw_mtct_drug
Gender
male
female
Smoothed kw_mtct_drug, by region
Region
Central
Eastern &
Northern
Southern
Western
Smoothed kw_mtct_drug, by age groups
age groups
15-24
25-49
Smoothed kw_where_test, by gender
GPR mean of kw_where_test
Smoothed kw_where_test, by region
Smoothed kw_where_test, by age groups
1.00
0.75
0.50
0.25
0.00
1990
1995
2000
2005
2010
2015


Smoothed att_vegetable, by gender
GPR mean of att_vegetable
Gender
male
female
Smoothed att_vegetable, by region
Region
Central
Eastern &
Northern
Southern
Western
Smoothed att_vegetable, by age groups
age groups
15-24
25-49
Smoothed att_secret, by gender
GPR mean of att_secret
Smoothed att_secret, by region
Smoothed att_secret, by age groups
1990
1995
2000
2005
2010
2015
0.00
0.25
0.50
0.75
1.00

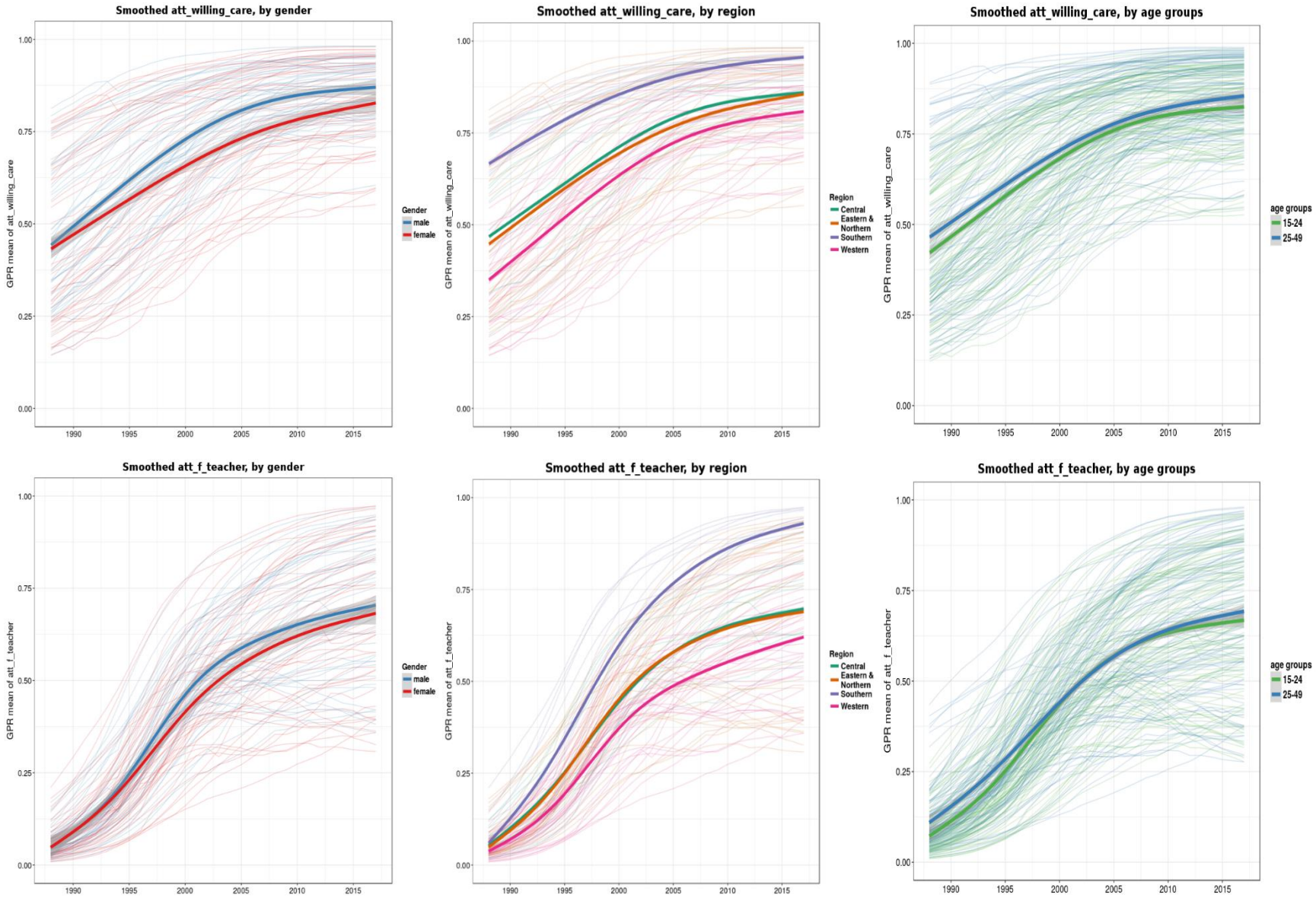


Figure 2-12 Mean trends of the indicators by gender, subregion, and age group

4. Discussion

4.1 Variations across countries and subregions

Although people's knowledge of and attitudes about HIV/AIDS have improved significantly in SSA over the past three decades, there are still marked variations across SSA countries. For example, in 2017 the proportion of people knowing that using a condom can reduce the risk of HIV infection ranged from 21% in Somalia to 96% in Swaziland.

By subregion, people in southern SSA tend to have better knowledge and attitudes about HIV/AIDS than people in other subregions. People in western SSA tend to have lower knowledge of HIV/AIDS compared with people in other subregions, and the discrimination against PLWHs and their family members in western SSA seems to be the greatest of all subregions. People in Somalia, South Sudan, Sudan, Chad and Mauritania tend to have poor knowledge and attitudes about HIV/AIDS, while people in Swaziland, Rwanda, Botswana, and Lesotho tend to have good knowledge and attitudes about HIV/AIDS. Broadly, places with very high HIV prevalence have greatest knowledge and best attitudes while those of lowest prevalence have poorest knowledge and worst attitudes.

4.2 The decreasing trends of some knowledge and attitudes indicators

Although most of the HIV/AIDS knowledge and attitudes indicators have been improving significantly over the past three decades, people's knowledge of MTCT and people's willingness to disclose their family members' HIV status have been significantly decreasing since 2000 in many countries, especially for *att_secret*, which has been decreasing significantly in almost all countries in SSA.

The underlying reasons for the deteriorating trends of these indicators are out of the scope of this study. However, we hypothesize that the advent and the scale up of the PMTCT-ART have

played a role in the decreasing trends of the knowledge indicators on MTCT. The effectiveness of ART to prevent MTCT may have made people believe that HIV is no longer transmitted from mother to child particularly during pregnancy. In such case, we suggest that the questions asking about people's knowledge on MTCT include some clarifications such as "*Without any treatment*, can HIV be transmitted from a mother to her baby during pregnancy?" In addition, during antenatal visits, health workers may take the opportunity to re-educate HIV-positive mothers about the risk of MTCT if they do not follow the treatment regimen rigorously.

Regarding people's willingness to disclose their family members' HIV status, we think that the decreasing trend of this indicator across the board may suggest that the perceived discrimination of PLWHs and their family members may have been increasing over time in SSA. It seems to be contradictory that the other three attitudes indicators, namely, *att_vegetable*, *att_willing_care,* and *att_f_teacher* have all been increasing significantly over time. However, we argue that these three indicators are different from *att_secret* because they reflect people's actual discrimination against PLWHs while *att_secret* reflects people's perceived discrimination as family members of PLWHs, reflecting more HIV/AIDS stigma and its internalization. The gap between the actual and the perceived discrimination against PLWHs are clearly exemplified by southern SSA, where people tend to do much better on all the indicators than people in the other subregions, except for *att_secret,* of which they have the same, if not lower, level of attitudes as people in eastern SSA.

Although the underlying causes of the gap between the actual and the perceived discrimination is unknown and is beyond the scope of this study, we discuss a few hypotheses here for future studies to explore. The first hypothesis is that the actual discrimination is in fact also increasing but due to increased knowledge of HIV/AIDS, people tend to give the "correct" answer to

questions about their attitudes towards PLWHs to avoid embarassment before the interviewer. The rationale is that people usually do not want to "look bad". If a person knows enough about the transmission, prevention, and treatment of HIV/AIDS, the person may feel embarrassed to show any discrimination against PLWHs in front of the interviewer. However, when asked about themselves, they are more likely to show their true opinion on PLWHs. The second hypothesis is that *att_secret* has been decreasing simply because people care more about their privacy than before as the society develops. In this case, *att_secret* does not measure the perceived discrimination well because it may have been confounded by people's increasing awareness of their privacy. The third hypothesis is that people now may think it unnecessary to tell others about their or their family members' HIV status because they know what to do if they are infected with HIV and know that they can live a normal life with the help of ART even if they are infected with HIV. In comparison, people in earlier years were more likely to share their family member's status in order to seek help from others because they did not know what to do and felt hopeless. No matter what has caused the decreasing trend in *att_secret*, we suggest that the DHS and other surveys add additional questions to probe the underlying reasons why people want to keep their family members' HIV status a secret.

4.3 The gap between men and women in HIV/AIDS knowledge and attitudes

Men in SSA generally have better knowledge of and attitudes about HIV/AIDS than women. However, we are pleased to see that the gap between men and women is narrowing over time, especially for knowledge on MTCT, in which women have surpassed men in the early 2000s. The contributors to this encouraging achievement of women are multifaceted and future studies are needed to fully understand them. However, this significant achievement of women coincides with the advent and upscale of PMTCT-ART programs in many SSA countries. [60,66–68] We

believe that the PMTCT programs, which provide voluntary testing and counseling (VTC) services for HIV/AIDS and were started and scaled up in many SSA countries during the 2000s [60,66–68], have played an important role in it. Being linked to the antenatal clinic, the PMTCT programs often serve as a key point of entry for women and children for HIV/AIDS education and services. [69] If future studies prove that PMTCT programs indeed greatly contribute to the increase of women's knowledge on MTCT of HIV, there is an opportunity for us to further improve women's knowledge and attitudes about other HIV/AIDS indicators by incorporating a more comprehensive HIV/AIDS education in the PMTCT programs. Therefore, this would further narrow the gap between men and women.

4.4 Knowledge and attitudes about HIV/AIDS of young people in SSA

Demographically, sub-Saharan Africa has the youngest population in the world, which is also growing fast. In 2012, more than 70% of the population in SSA were young people under 30 years old. [70] By 2055, the population aged 15 to 24 in SSA is projected to more than double the 2015 number.[71] Young people, especially young women, are the key population for HIV/AIDS prevention because they are disproportionately affected by HIV/AIDS. The AIDS-related mortality among adolescents tripled between 2000 and 2015, the only age group for which ADIS-related mortality had increased during the period.[72] In the UNAIDS 2016 Prevention Gap Report, lack of comprehensive knowledge about HIV/AIDS is identified as one of the major barriers for HIV prevention among young people in SSA.[73] By estimating the trends of 16 key indicators of HIV/AIDS knowledge and attitudes by gender and by age group in all SSA countries, our study provides useful evidence for each SSA country to identify their specific gaps among young people.

In this study, we find that young people's knowledge and attitudes about HIV/AIDS, except for *kw_mtct_preg* and *att_secret*, have improved over the past decades. However, compared with older people, young people aged 15-24 in SSA generally have lower knowledge about HIV/AIDS, especially for *kw_looking*, *kw_pv_condom*, *kw_pv_one_partner*, *kw_where_test*, and the four MTCT indicators. It is more concerning that the gaps in some knowledge indicators such as *kw_pv_condom*, *kw_looking* and *kw_mtct_drug* between older and young people have been widening in many countries in recent years.

Africa is in the middle of big demographic shift, with a huge adolescent bubble coming. There is also a large fraction of new HIV cases in adolescents and young adults.[74] However, our results suggest that the rising generation has not really known the illness. Meanwhile, the global financing of HIV/AIDS has shifted from primary prevention to treatment of HIV/AIDS,[75] which may have contributed to the widening knowledge gap between the young and the old. No matter what is the underlying cause of the widening knowledge gap between the young and the old, our results clearly suggest that we should at least continue, if not increase, the efforts to improve young people's knowledge of HIV/AIDS in SSA.

4.5 Comparison with other similar studies

Most previous studies on knowledge and attitudes about HIV/AIDS in sub-Saharan Africa focus on a specific geographic region [76,77] and/or a specific group of people, such as health care professionals [78,79], HIV-infected adults [80], students [18,57,81], and pregnant women [82]. Although DHS, MICS, and other national surveys have been collecting nationally representative data on knowledge and attitudes about HIV/AIDS in SSA for a long time, studies fully utilizing these data across countries are extremely rare. So far, we have found only four studies that examine the levels and/or the trends of HIV/AIDS knowledge and attitudes across countries using these data.

In 2008, Burgoyne and Drummond conducted a comprehensive literature review on the knowledge of HIV/AIDS among women in 18 SSA countries. They found that although most people had heard of AIDS, their knowledge of HIV transmission and prevention was still limited. In addition, men in SSA had greater knowledge of HIV/AIDS than women did across all studies, except for knowledge of MTCT of HIV, of which women had better knowledge than men did in half of the studies they reviewed [83]. Our study reaffirms their findings. Although Burgoyne and Drummond conducted a comprehensive literature review, they simply summarized the findings from all these studies and did not examine the change of women's knowledge of HIV/AIDS over time.

In 2009, USAID published DHS comparative report No.24, which examined the change in HIV-related knowledge and attitudes in 23 SSA countries between two rounds of DHS conducted between 1992 and 2000.[84] The major findings of this report on are reaffirmed in our study, including higher knowledge of MTCT among women than among men and the decrease of *att_secret* across the board. Although this report included 46 DHS surveys in 23 countries and tried to examine the change of people's knowledge and attitudes about HIV/AIDS for both genders over time, the researchers drew their conclusions by simply comparing the numbers between two rounds of surveys. The methods used were simple and the number of data points for each country were limited, which are the major limitations of this study.

In 2016, Teshome *et.al.* published a study in which they compared people's knowledge of and attitudes about HIV/AIDS for women aged 15-49 in Burundi, Kenya, and Ethiopia. Their study suggested that elder women were more likely to have better comprehensive knowledge and attitudes about HIV/AIDS than young women and that women in Burundi and Ethiopia were the most and the least likely to have comprehensive knowledge and attitudes about HIV/AIDS,

respectively. [59] These findings are also supported by our results qualitatively, though the results of the two studies are not fully comparable due to different outcomes. Although Teshome *et.al.* used more solid methods for their analysis, they did only a cross-sectional comparison between three countries but did not examine the change in people's knowledge and attitudes about HIV/AIDS in the three countries.

In 2018, Chan and Tsai from Harvard published the most recent study examining the trends of HIV knowledge in 33 SSA countries using 75 DHS and AIS surveys from 2003 to 2015. This was the first study that comprehensively examined the trends of HIV knowledge across SSA countries using a fair number of DHS and AIS surveys. The outcome variables in their study are a knowledge score of HIV comprising five knowledge indicators, namely, *kw_pv_condom*, *kw_pv_one_partner*, *kw_looking*, *mis_mosquito,* and *mis_food*, and a composite indicator of comprehensive HIV/AIDS knowledge defined by answering all five knowledge indicators correctly. Chan and Tsai used regression methods with country fix effect to examine the effect of time on the outcomes, and found that the HIV knowledge score and the probability of comprehensive HIV/AIDs knowledge had significantly increased over time between 2003 and 2015 but the increase was modest. They claimed to have found no evidence of substantial improvements in HIV knowledge from 2003 to 2015,[50] which is contradictory to our findings. Although we do not use the HIV knowledge score or indicator of comprehensive HIV knowledge as the outcomes in our study, our results show that the five knowledge indicators comprising the HIV knowledge score and the indicator of comprehensive HIV knowledge have all substantially improved since 2000 in SSA. Nonetheless, their finding that men on average had a higher HIV knowledge score and a higher level of comprehensive knowledge of HIV/AIDS than women did across 33 SSA countries is the same as our findings.

Although Chan and Tsai claimed to examine the trends of people's HIV knowledge in SSA, they failed to adjust for socio-demographic variables (SDVs), especially educational attainment and household wealth, which greatly contribute to people's HIV knowledge and are changing significantly over time in SSA.[85] Educational attainment, for example, has been improving significantly since 2000 in SSA.[85,86] By adjusting for the SDVs, they teased out a large portion of the effect of time on people's HIV knowledge mediated by the SDVs, and what is left for time to explain is of course little. Secondly, they did not carefully describe how they cleaned the DHS datasets. They mentioned that they dropped about 100,000 observations due to missingness in one of the five knowledge questions. However, when cleaning the DHS data, we find that many DHS raw datasets do not account for skip patterns in the questionnaire properly. For example, people answering no to the *heard_aids* question should have 0 for all the following knowledge indicators. However, in the DHS raw datasets, these people often simply have missing values for the following knowledge indicators. Excluding those people having not heard of AIDS may have seriously biased the estimated trends upwards especially in early years when more people had not heard of AIDS. Some DHS even have more complex skip patterns in the questionnaire, which requires additional study. In addition, Chan and Tsai modeled the effect of time linearly, which may not capture the potential non-linear trends in HIV knowledge across countries. Lastly, their study did not provide visual presentations of the trends in the SSA countries, making it harder for cross-country comparisons and less useful for policy makers and other health practitioners.

Compared with the existing studies of similar topics, our study has the following advantages. First, the scope of our study is much larger than existing studies. By utilizing 249 DHS, AIS, MICS, and other national surveys, our study estimates the trends of 16 key indicators of

HIV/AIDS knowledge and attitudes by gender and age group from 1988 to 2017 in all 47 SSA countries. By estimating the trends of each indicator by gender and age group, our study provides useful and unique evidence to inform HIV/AIDS policies and/or interventions targeting women and young people aged 15-24. Second, we cleaned the raw dataset of each survey by carefully going through the questionnaires and accounting for skip patterns in the questionnaires. Third, our study utilizes appropriate and state-of-the-art statistical techniques including bias adjustment, crosswalking, natural splines, bootstrapping, GAMM, and ST-GPR, to better model the trends of the indicators in each country and quantify the uncertainty of our estimates. Lastly, our study provides good visual presentations of the estimated trends of the indicators by countries, by subregion, gender, and age group. These visual presentations make the results of the study more accessible to non-technical people, including policy makers and other health practitioners, and make the comparisons between countries, subregions, genders, and age groups much easier.

4.6 Limitations

Although we conducted this study carefully, it is not without limitations. First of all, the survey data used in this study were collected by different people, using different questionnaires and under different settings. There can be huge heterogeneity in the quality of data across surveys due to these discrepancies. However, we cannot verify the quality of each survey because we do not know what exactly happened during the data collection procedure. For report data, the data quality is even more concerning because it is not uncommon for people to make mistakes when calculating the national estimates. To mitigate this limitation, we have added NSV to the data to account for the heterogeneity due to random errors and to adjust the bias due to survey type. In addition, we examine each survey carefully and remove extremely unusual data points especially if they are report data.

Another limitation of this study is that the number of data points in the early years is very limited, which affects the estimates in the early years. Only *heard_aids* and *kw_looking* have a few data points before 1995 and some indicators, such as *att_secret* and *kw_mtct_drug*, only have data after 2000. Therefore, the estimated trends of most indicators in early years heavily depend on extrapolations of the first-stage model in ST-GPR. However, we fit the first-stage model using a wide range of country-level covariates (see **Table 3**), including mean years of education and GDP per capita. These covariates, which are complete time series over the estimation period, greatly help to inform the level of the indicators in the early years when there were limited or no data. In addition, we use natural splines to model *year* in the first-stage model and carefully select the knots of the splines to make sure that the model extrapolations in the early years make sense in reality. For example, there have been decreasing trends for *kw_mtct_preg* and *att_secret* in many countries since 2000, when data began to become available for these two indicators. If we had linearly extrapolated the decreasing trends into early years, the level of these two indicators would be unreasonably high. By carefully studying the early history of HIV/AIDS,[87–90] we believe that although the trends are decreasing, the level of these two indicators must have been low in the beginning of the epidemic and increased to a certain level over time before they began to decrease after 2000 as the data show. The parabolic shape of the trends is also supported by the observed trends in some countries such as Zambia where more data are available. Therefore, we carefully tune the knots of the natural splines for each indicator to make sure that the extrapolation of the indicator in the early years is reasonable judged by the early history of HIV/AIDS. For example, WHO confirmed that HIV could be transmitted from mother to child during breastfeeding in July 1987,[89] which makes us believe that the level of *kw_mtct_bf* should be very low across countries in 1988. Admittedly, no matter how careful we

are, the knots selection is subjective and the extrapolations can be off. However, we find that the knots selection has little impact on the estimated trends for years when data are available. Even in early years, as long as there is a data point, the estimated trends will follow the data closely. Therefore, we think that our estimated trends are trustworthy for years when there are data available (e.g. after 2000) regardless of knots selection or early year extrapolations.

Lastly, in addition to the knots selection in the first stage prediction, the parameter selection of ST-GPR in the second and the third stage are also subjective. We do not use cross validation or other statistically more rigorous methods to select the parameters for two reasons. First, based on experience from previous rounds of GBD studies, the parameters selected using cross validation are often not optimal in terms of the plausibility and the smoothness of the resulting trends, particularly if the data are sparse across countries, which is the case for this study. For instance, we tried to use generalized cross validation (GCV) to select the knots and found the predictions imprecise and the extrapolations in the early years unrealistic. In addition, since we estimate the trends of 16 indicators, we prefer to make the parameters as consistent as possible across all indicators for comparison purposes. Therefore, we decided to select the parameters for ST-GPR following the newest guideline used in the GBD 2017.[39]

5. Conclusion

By including 248 national surveys of 47 SSA countries from 1988 to 2017, our study produces solid evidence on people's knowledge and attitudes about HIV/AIDS in SSA countries over the past decades. According to our findings, although there have been substantial improvements in people's knowledge and attitudes about HIV/AIDS in SSA, there are still efforts to be made. Specifically, we need to further improve people's knowledge on MTCT of HIV, especially knowledge of MTCT during pregnancy and to further reduce HIV/AIDS stigma and its

internalization. In addition, our findings suggest that women and the young tend to have lower knowledge and poorer attitudes about HIV/AIDS than men and the old. Therefore, we urge that more resources should be used to improve knowledge and attitudes about HIV/AIDS among women and the young in the SSA countries. Especially, given the booming young population and the large fraction of new HIV cases among adolescents and young adults in SSA, policy makers as well as researchers should pay more attention to changes in young people's knowledge and attitudes about HIV/AIDS in this region.

Appendix

Removed surveys, indicators and outliers

1. Removed surveys

Among the 3,682 surveys in the 47 SSA countries, 248 surveys are accepted in the study including 231 surveys with microdata and 17 surveys with report data only. The rest of the 3,434 surveys are excluded for various reasons. Most commonly, a survey is excluded because it does not have data on HIV/AIDS knowledge and attitudes. A few surveys are excluded because the data are not nationally representative – focusing on subnational areas or a specific group of people, e.g. people who have children, or senior people. Four surveys are excluded due to serious concerns about the quality of the data in the survey. The full details of the 3,682 surveys are included in Supplementary Materials.

2. Removed indicators

In addition to removed surveys, indicator(s) in an accepted survey may also be removed due to various issues including having missing values for many people, being calculated for a subgroup of people, having a unique or strange skip pattern, and being contradictory to other indicators in the same survey. The full details of the removed indicators are included in Supplementary Materials.

3. Removed outliers

After carefully vetting each data point used in the estimation, there are a few data points which are correct but have extreme or unusual values compared with data in adjacent years and in adjacent countries. After carefully reviewing the data and consulting experts, we removed these outliers from the estimation. Full details of the removed outliers are included in Supplementary Materials. Compared with the number of data available, the number of outliers is extremely small (<1%) and they only affect the estimated trends in a few countries for a few indicators. Moreover, removing the outliers has almost no impact on the estimated mean trends of the indicators but makes the affected country-specific trends more reasonable. Therefore, we are not concerned about removing these outliers.

Strategies for data cleaning

1. General skip patterns applied to all surveys

After extracting the individual level survey data, we carefully cleaned the individual level data to account for the skip patterns for each survey. Several general skip patterns are applied to all the surveys.

First, if a person had not heard of AIDS, i.e. *heard_aids* is 0, we set all the other knowledge and misconception indicators for this person to be 0 and all the attitudes indicators to be missing.

In addition to *heard_aids*, there are other gateway questions in some surveys. For example, some surveys have a gateway question for MTCT questions, i.e. *kw_mtct*, asking people whether they

knew that HIV could be passed from mother to child. Therefore, if *kw_mtct* is in the survey and is 0, we set all the MTCT indicators, namely *kw_mtct_preg*, *kw_mtct_delivery*, *kw_mtct_bf*, and *kw_mtct_drug*, to be 0.

For surveys that have all four MTCT indicators, the question on *kw_mtct_drug* is often skipped if the person did not know any of the routes of MTCT of HIV. Therefore, when all four MTCT indicators exist in the survey and *kw_mtct_preg*, *kw_mtct_delivery*, and *kw_mtct_bf* are all 0, we set the kw_mtct_drug to be 0 as well.

Lastly, in most surveys that ask both "Do you know a place for an HIV test? (*kw_where_test*)" and "Have you ever been tested for HIV? (*ever_tested*)", if the person has been tested before, i.e. *ever_tested* is 1, then the *kw_where_test* is often skipped and is assumed to be 1 as well. For women, more testing questions are usually asked, e.g. "Were you tested for HIV as part of your antenatal care? (*test_at_anc*)" and "Were you tested during the delivery? (*test_at_delivery*)". If any of the testing questions is positive, the question on *kw_where_test* is skipped and assumed to be 1. Therefore, if the person has been tested for HIV before, i.e. any of the testing indicators is 1, we set the *kw_where_test* to be 1 for the person.

Although the aforementioned general skip patterns are actually applied in most surveys, there are still exceptions, especially for *kw_where_test*, whose skip patterns for women are quite complex and vary for some surveys. However, we think that these general skip patterns are reasonable and thus we apply these general skip patterns to all the surveys for consistency across surveys.

2. Survey specific skip patterns applied only to certain surveys

In addition to the general skip patterns, there are also survey-specific skip patterns applied only to certain surveys. In some surveys, there is a gateway question asking people whether there is anything a person can do to protect themselves from HIV/AIDS (*kw_pv_any*)? Although *kw_pv_any* is usually the gateway question for spontaneous questions on knowledge of HIV prevention, it is also the gateway question for the probed questions on knowledge of HIV prevention and/or HIV misconceptions in some surveys, where these probed questions are skipped if *kw_pv_any* is 0. **Table S1** lists the node identifiers (NIDs) of the affected surveys. For these surveys, if *kw_pv_any* is 0, we set the skipped probed indicators to be 0 as well.

There is another specific skip pattern in a few surveys, where the probed indicator *kw_pv_condom* and *kw_pv_one_partner* are skipped and assumed to be 1 if people spontaneously mentioned the two prevention measures of HIV in the previous spontaneous question. We apply this specific skip pattern only to these surveys.

Lastly, in a few surveys, there is a skip pattern for *heard_aids*, namely, if people spontaneously mention HIV/AIDS when asked about knowledge of any STD, the *heard_aids* question is skipped and assumed to be 1. We apply this specific skip pattern only to these surveys.

Table S1: Surveys where *kw_pv_any* is also the gateway for probed questions

Issues	NIDs of the affected surveys
kw_pv_any is the gateway for probed prevention indicators: *kw_pv_condom* and *kw_pv_one_partner*	18950, 19019, 19088, 19571, 19579, 20145, 20223, 20252, 20263, 20315, 20322, 20394, 20417, 20567, 20722, 20796, 20865, 20993, 21102, 21151, 20998, 11639, 12232, 12243, 12320, 12886, 1404, 1994, 2053, 2209, 2244, 26444, 27020, 27044, 27055, 27215, 3114, 3655, 3922, 4808, 687, 7387, 7721, 9439, 21442
kw_pv_any is the gateway for probed misconception indicators: *mis_mosquito, mis_food and mis_witchcraft*	18950, 19088, 19579, 20145, 20223, 20252, 20263, 20315, 20322, 20394, 20417, 20567, 20722, 20865, 20993, 21102, 11639, 12232, 12243, 12320, 12886, 1994, 2053, 2209, 2244, 26444, 27020, 27044, 27055, 3114, 3655, 3922, 4808, 687, 7387, 7721, 9439, 21442
When people mention using condom or having one partner can reduce the risk of HIV spontaneously, the kw_pv_condom or kw_pv_partner is set to be 1	27215, 27924 (*kw_pv_condom* only), 21151, 19571, 22112 (*kw_pv_condom* only)
When people mention HIV/AIDS spontaneously when asked about knowledge on any STD, the heard_aids is set to be 1	27511, 104316, 80790, 20786, 20767

3. Other survey specific issues

Table S2 summarizes other survey-specific miscellaneous issues during data extraction and cleaning. The codebook and the codes used for extracting and cleaning the survey data are included in Supplementary Materials.

Table S2: survey-specific miscellaneous issues

nid	sex	Indicator	Issues
22114	both	kw_mtct_drug	The survey asks people whether they know any way to avoid MTCT and then asks the specific way, including AZT. The indicator kw_mtct_drug is obtained from the two questions.
22116	both	kw_mtct_drug	The survey asks people whether they know any way to avoid MTCT and then asks the specific way, including AZT. So the indicator kw_mtct_drug is obtained from the two questions .
21442	both	kw_mtct_drug	The survey asks people whether they know any way to avoid MTCT and then asks the specific way, including AZT. So the indicator kw_mtct_drug is obtained from the two questions.
22116	both	kw_where_test	If people mention any place to test HIV, their kw_where_test is 1

27952	both	kw_mtct_drug	The survey asks people whether they know any way to avoid MTCT and then asks the specific way, including AZT. So the indicator kw_mtct_drug is obtained from the two questions.
12102	both	N/A	Only keep the completed interviews (drop if creal >2)
228102	both	N/A	Only keep the completed interviews (drop if fresp >2)
313076	both	N/A	Only keep the completed interviews (drop if fresp >2). In addition, drop those with missing value in age
134753	both	kw_mtct_drug	The survey asks people whether they know any way to avoid MTCT for the unborn and newborn searately and then asks the specific way, including AZT. So the indicator kw_mtct_drug is obtained from the four questions.
27924	both	kw_mtct_drug	The survey asks people whether they know any way to avoid MTCT and then asks the specific way, including AZT. So the indicator kw_mtct_drug is obtained from the two quesitons.
325046	both	att_willing_care	People were asked about their willingness to take care of their female and male relatives separately. People answering yes to either question have att_willing_care =1.
324443	both	att_willing_care	People were asked about their willingness to take care of their female and male relatives separately. People answering yes to either question have att_willing_care =1.
1404	women	N/A	The sample in this survey is equally weighted. The sample weight for all people is 1
1994	women	N/A	The sample in this survey is equally weighted. The sample weight for all people is 1
7721	women	N/A	The sample in this survey is equally weighted. The sample weight for all people is 1

Knots selection for the natural splines of year

In the first-stage prediction of ST-GPR, we model year using natural splines to capture the nonlinear trends of the indicators over time. To make the model flexible enough to fit different trends of the indicators in different countries but not too flexible to overfit the scarce data in each country, we choose the minimum of 3 knots for the spline, including two boundary knots and one interior knot. To select the best locations for the three knots, we try different combinations of the locations, predict the time series of the indicator for all countries, evaluate the fit of the model to the data using AIC and BIC and visually examine the plausibility of the trends. Eubank, in his book *Nonparametric Regression and Spline Smoothing*, recommends visual inspection as the simplest method for selecting spline knots and stats that selecting spline knots through visual inspection "often tends to work quite well in terms of giving a visual pleasing fit to the data and has a definite computational advantage over other methods".[47] Using visual inspection to select spline knots is also common in research.[48,49] In our study, selecting knots through visual inspection is even more important and useful because it helps provide plausible extrapolations in the early years when there were few or no data. Following Eubank's recommendation, we place the three knots for all indicators between 1995 to 2015 when the trends change significantly and the most data are available.[47]

Based on the early history of HIV/AIDS (**Box**)[87,89,90], we believe that people's knowledge and attitudes about HIV/AIDS should be generally low but the level might vary across indicators.

Specifically, Richard Berkowitz and Michael Callen had advocated condom use among gay men in 1983 and by September of 1983, the CDC identified all major routes of transmission and ruled out transmission by casual contact, food, water, air, or surfaces.[89] Therefore, we believe that people in SSA should have low to moderate knowledge of indicators *kw_looking, kw_pv_condom, kw_pv_one_partner,* and *mis_food* in 1988. Limited data available for some of these indicators in early years also support our belief. For instance, a Botswana DHS in 1988 shows that about 60% of women aged 15-49 had heard other people used condoms to avoid HIV/AIDS and about 25% of them rejected the misconception that HIV can be transmitted by sharing food. In 1985, USFDA approved the first HIV blood test and in December, the U.S. Public Health Service issued the first recommendations for preventing mother to child transmission of the virus.[89] Therefore, we believe that people in SSA should generally have low knowledge of *kw_where_test, kw_mtct_preg,* and *kw_mtct_delivery* in 1988. Lastly, WHO confirmed that HIV could be passed from mother to child during breastfeeding in 1987 and the AZT was first recommended to prevent MTCT of HIV in 1994. We believe that people's knowledge of *kw_mtct_bf* and *kw_mtct_drug* should be very low in 1988 and 1995, respectively. For people's attitudes about HIV/AIDS, we believe that the discrimination against PLWHs in SSA was overwhelming in the early years but many people would still be willing to take care of their family members if they are infected.

Box 1: Early History of HIV/AIDS [89,90]

1982:

- In September, the CDC used the term 'AIDS' (acquired immune deficiency syndrome) for the first time.

1983:

- May: Richard Berkowitz and Michael Callen—both men living with AIDS—published a booklet on "safer sex" titled How to Have Sex in an Epidemic: One Approach. It advocates condom use for gay men and focuses on self-empowerment for those living with AIDS.
- By September, the CDC identified all major routes of transmission and ruled out transmission by casual contact, food, water, air, or surfaces.

1985:

- In March, the U.S Food and Drug Administration (FDA) licensed the first commercial blood test, ELISA, to detect antibodies to the virus. Blood banks began to screen the USA blood supply.
- In April, the U.S. Department of Health and Human Services (HHS) and the World Health Organization (WHO) hosted the first International AIDS Conference in Atlanta Georgia.
- Ryan White, a teenager from Indiana, USA who acquired AIDS through contaminated blood products used to treat his haemophilia, was banned from school.
- In December, the U.S. Public Health Service issued the first recommendations for preventing mother-to-child transmission of the virus.

1986:

- In May, the International Committee on the Taxonomy of Viruses said that the virus that causes AIDS will officially be called HIV (human immunodeficiency virus).
- October 22: Surgeon General C. Everett Koop issued the Surgeon General's Report on AIDS. The report makes it clear that HIV cannot be spread casually and calls for a nationwide education campaign (including early sex education in schools), increased use of condoms, and voluntary HIV testing.

1987:

- In February, WHO launched The Global Program on AIDS to raise awareness; generate evidence-based policies; provide technical and financial support to countries; conduct research; promote participation by NGOs; and promote the rights of people living with HIV.
- In March, the FDA approved the first antiretroviral drug, zidovudine (AZT), as treatment for HIV.
- In July, WHO confirmed that HIV could be passed from mother to child during breastfeeding

1988:

- WHO declared 1st December as the first World AIDS Day.

1994:

- In August, the USA Public Health Service recommended the use of AZT to prevent the mother-to-child transmission of HIV.

Using the AIC and BIC of the models, and following Eubank's recommendations and the aforementioned general rules, we carefully selected the locations of the knots to make sure that the extrapolated trends of the indicators in the early years when there were no data available are reasonable. Specifically, we tested 24 combinations of different locations of the three knots and chose the best combinations based on small AIC/BIC and the plausibility of the curves. **Table S1** shows the best locations of the knots for each indicator. The last column indicates whether the curves are sensitive to the locations of the knots. For indicators that have a few data points in the early years (e.g. before 2000), the impact of the locations of the knots on the trends is limited. In addition, during the knots selection process, we find that the location of the knots has very little impact on the trends when more data are available, which assures the credibility of the estimated trends in more recent years (e.g. after 2000). **Figure S1** demonstrates the effect of the locations of the knots on *att_secret*, the most sensitive indicator to the locations of the knots. We can see that although the estimated trend in the early years changes significantly with different selection of the locations of the knots, the estimated trend after 2000, when data became available for the indicator, changes very little.

Table S3. Locations of the knots for each indicator

Indicator	Boundary knot 1	Interior knot	Boundary knot 2	Sensitivity
heard_aids	1995	2000	2010	No
kw_looking	2000	2008	2010	No
kw_pv_one_partner	1995	2007	2010	No
kw_ov_condom	1995	2008	2010	No
kw_mtct_preg	1995	2000	2008	Yes
kw_mtct_delievery	1995	2005	2010	Yes
kw_mtct_bf	1995	2000	2008	Yes
kw_mtct_drug	1995	2000	2010	Yes
kw_where_test	1995	2008	2010	Yes
mis_mosquito	1995	2005	2010	No
mis_food	1995	2008	2010	No
mis_witchcraft	1995	2000	2010	Yes
att_vegetable	1995	2008	2010	No
att_secret	1995	2000	2004	Yes
att_willing_care	1995	2008	2010	No
att_f_teacher	1995	2008	2010	No

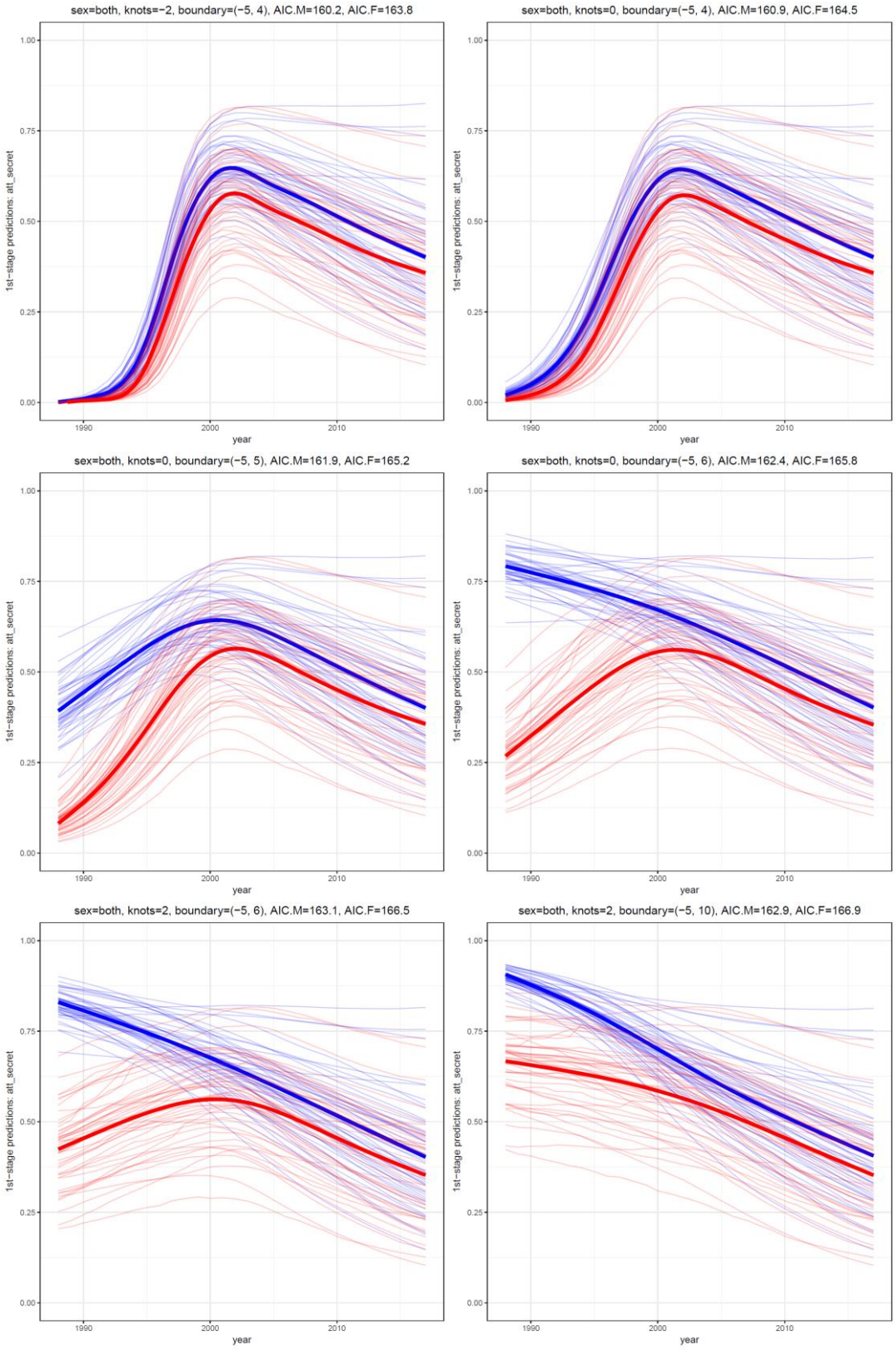


Figure S1: The impact of knots locations on the predicted trends of *att_secret*

Chapter 3 : Evaluating the Performance of Different Multiple Imputation Methods When Imputing Missing variables of Knowledge and Attitudes about HIV/AIDS in the Surveys

Abstract

Background

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

Methods

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

Results

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

Conclusion

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

1. Introduction

As in many other fields such as political science and economy, the time series cross-sectional (TSCS) data are an important type of data for global health research, especially for comparative studies of health indicators across countries and over time, such as the Global Burden of Disease (GBD) Study.[91] Large-scale national health surveys, such as Demographic Health Survey (DHS), Multiple Indicators and Cluster Surveys (MICS) and other national surveys conducted regularly by each country, are important sources of TSCS data for global health research. Although these TSCS data provide numerous measurements that open doors to many research opportunities, when used together, they often face serious missing data problem simply because some questions are not asked in some surveys (a.k.a. missing variables).[23–26] Even for DHS or MICS which are designed to be as consistent as possible over time and across countries, the questionnaires used in different countries or in different rounds can be significantly different from each other due to local adaptation and changes in the health priorities over time.[27] When surveys from different sources are used, missing variables usually become more prevalent.

Multiple imputation (MI), a Bayesian model-based approach first introduced by Rubin[28], has become a major principled method for estimating missing data across different research fields.[24,29] The basic idea and intuition of MI is to estimate the missing values in a dataset by making use of all the observed data. The estimation of the missing values is usually repeated *m* times to produce *m* different complete datasets in which the observed data are the same but the imputed data are different across the *m* complete datasets. After obtaining the *m* complete datasets using MI, one can perform the identical analyses on each of the *m* complete datasets and combine the results (estimates and standard errors) using simple rules provided by Rubin[28] to produce an overall estimate and its standard error. The major benefits of MI are that it results in

unbiased estimates, increase statistical power by using all available data and account for the uncertainty due to missing data.[28–30] Nowadays, there are two major families of MI approaches, namely, MI using joint modeling (MIJM) and MI using chain equations (MICE).[29,31,32] MIJM draws missing values simultaneously for all incomplete variables using a multivariate distribution (e.g. multivariate normal)[30–32] while MICE imputes incomplete variables one at a time, drawing missing values sequentially from a series of univariate distributions (e.g. regression models).[29,32,92,93] When first introduced by Rubin and others, MI was mainly used for imputing missingness in single-level cross-sectional data[28,33]. However, after more than four decades, MI has been developed to be able to properly impute missing data for multiple-level data[31,94,95], longitudinal or panel data [96,97] and TSCS data.[34]

In our last study on people's knowledge and attitudes about HIV/AIDS in sub Saharan Africa, we estimated the trends of 16 key indicators of HIV/AIDS knowledge and attitudes across 47 SSA countries. Although we found 248 national surveys asking key indicators of HIV/AIDS knowledge and attitudes, only a few of them asked all the 16 key indicators and many missed one or more indicators. In addition, some surveys only collected women's but not men's data. Therefore, after we stacked all the country-level survey estimates together into one dataset, there were many missing values due to missing variables across surveys. To attenuate the impact of missing variables on our results, we used regression method to impute men's indicators using women's when men's indicators were not collected and we estimated the trends of the 16 key indicators separately to avoid missingness due to indicators not collected in some surveys.[98] However, both measures we took had limitations. First, although the indicators for women and men are highly correlated and the linear mixed model we used gave small in-sample prediction error, regression imputation has long been considered inappropriate for imputation due to its

inability to account for uncertainty of the imputed data.[30,99,100] Buuren even thought that regression imputation is the most dangerous of all imputation methods because it artificially strengthens the correlations in the data and thus leads to false positive and spurious relations.[99] Second, although we avoided the missing variables problem by estimating the 16 key indicators separately, we also lost the opportunity to improve the estimated trends of the indicators by borrowing information from available indicators in the survey. Given the scarcity of the data on most indicators, borrowing information from available indicators could significantly improve the estimated trends. Lastly, the missing variables in some surveys prohibited us from constructing more informative composite measurements of people's knowledge and attitudes.

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

2. Literature Reviews

2.1 Types of Missing Data

The problem of missing data is ubiquitous in research of all fields including global health research. Missing data, if handled inappropriately, can pose great threat to the validity of research findings.[28,29,31] According to the literature, missing data can be categorized either by the causes of the missingness or by the mechanisms of the missingness.

2.1.1 based on the causes of the missingness

When thinking of missing data in surveys, people usually think of missingness due to nonresponse to certain questions asked in a survey. However, there are more causes of missingness, especially for time series cross-sectional (TSCS) data, which are data collected from multiple sections (e.g. countries) and across multiple time periods (e.g. years).[23,34,101]

In a study of multi-format and multi-wave surveys, He et al. categorized missing data into three types based on the causes of the missingness, namely, *unit nonresponse, block nonresponse* and *item nonresponse*.[24,102] According to He et al. *unit nonresponse* refers to missingness of all variables of patients due to sampled patients not participating in the study; *block nonresponse* refers to missingness of blocks of variables due to early drop out or use of different forms of survey asking different questions; and *item nonresponse* refers to missingness of certain variables of a patient due to skip patterns of the survey or questions being refused or answered "don't know" by the patient.

Using different terminology, Denk and Weber categorized missingness of TSCS data into 6 types, namely, *missing items*, *missing variables within sections* (e.g. countries), *missing periods (e.g. years) within sections, missing periods*, *missing variables and missing sections*.[23] According to Denk and Weber, *missing items* refers to missingness of one or multiple indicators in one or multiple periods (years) for one or multiple sections (countries), e.g. an indicator is not collected for one country in one yea; *missing variables within sections* refers to missingness of one or multiple indicators in one section across all periods, e.g. an indicator is not applicable for one country and thus has never been collected for this country; *missing periods within sections* refers to missingness of all indicators in one or multiple sections for one or multiple periods, e.g. no survey is conducted in one or multiple years in a country and thus there is no data at all for these

years in this country; *missing periods* refers to missingness of all indicators in one or multiple periods for all sections, e.g. no survey is conducted in one or multiple years for all countries and thus there is no data at all in these years for all countries; *missing variables* refers to missingness of one or multiple variables in all periods across all sections, e.g. some variables have never been collected for all countries; lastly, *missing sections* refers to missingness of all variables in all periods for one or multiple sections, e.g. there is no data at all for one or multiple countries of interest in the dataset. ***Figure 3-1*** visually represents the six different types of missingness described by Denk and Webe.[23]

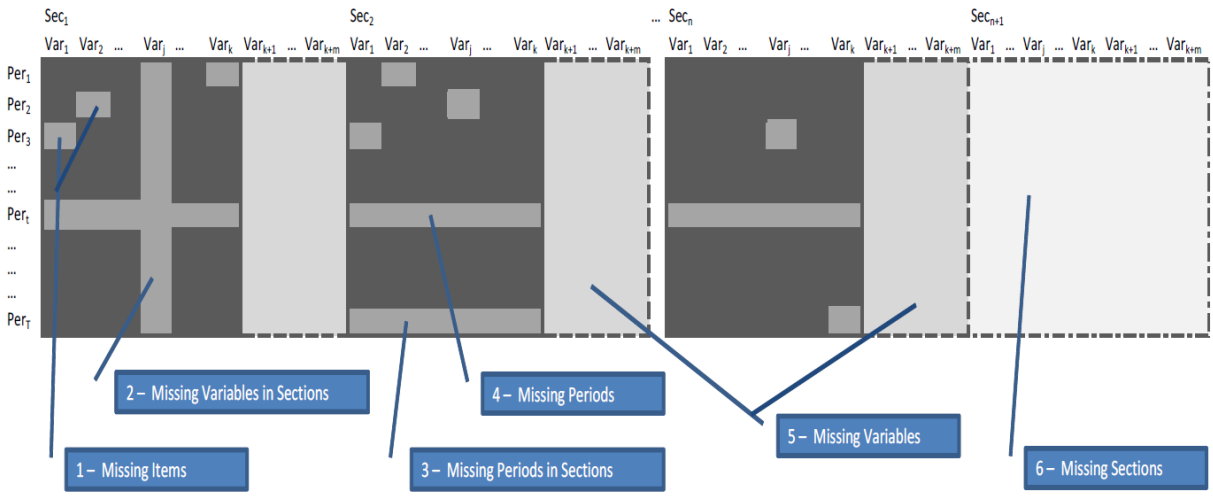


Figure 3-1 Missing data patterns for multivariate time series cross-sectional data [23]

2.1.2 based on the mechanisms of the missingness

In 1976, Rubin first described the mechanisms of missing data and categorized missing data into three types based on the reasons for the missingness.[103] According to Rubin, missing data can be missing completely at random (**MCAR**), missing at random (**MAR**) and missing not at random (**MNAR**).

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

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

where M and D represent missingness and data (both observed and missing data) respectively. MCAR implies that there is no relationship at all between the data and missingness. If the missingness is by design or planned, or because the samples are lost in transit or due to equipment failure, the missingness is considered to be MACR.[32,104] Under MCAR, the analyses are not biased by the missingness and listwise deletion produces valid estimates and inferences though power may be lost due to exclusion of observations with missing data.[99,104]

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

Mathematically, MAR holds if

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

Under MAR, any relationship between missingness and the unobserved data disappears after conditioning on the observed data.[32] In statistics, the missingness is "ignorable" if the missing data are considered MCAR or MAR.[105] However, it is worth noting that the "ignorability" of missingness does not mean that one can ignore the missingness, especially if the missingness is MAR. Instead, one needs to handle the problem of missingness using appropriate methods so that the missingness does not bias the estimates and the inferences.

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

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

which indicates that the missingness is related to the missing values themselves and the relationship between the missingness and the data remains even after conditioning on the observed data. For example, MNAR could occur if a respondent refuses to answer a question asking about his/her income because the respondent has very large or very small income.[32] MNAR is the real bane of missing data problem because one has to make untestable assumptions about the specific missing mechanisms to model the missing data in order to obtain unbiased estimates and inferences.[30,104]

2.2 Traditional Ad Hoc Methods for Missing Data Problem

Missing data is a very common issue in many studies across all fields. Many techniques have been developed to deal with missing data. Listwise deletion (LD), also referred to as complete-case analysis, is arguably the most commonly used ad hoc method to deal with missing data across fields. It is also the default way of handling missing data in many statistical packages, including R, Stata and SAS.[99] LD removes all the observations with missingness on the analysis variables and the following analyses are conducted using the complete cases only. The biggest advantage of LD is the convenience. Under MCAR, LD can produce unbiased estimates and inferences though at the cost of lower power due to exclusion of the incomplete observations.[99,106] However, the major limitation of LD is that it leads to biased estimates and inferences if the data are not MCAR, which is often the case.[100,107] In addition, LD often causes inconsistencies in the analyses. Since different analyses usually involve different subsets of variables and LD applies only to the active variables, different analyses are usually based on different subsamples of the dataset.[99]

Pairwise deletion (PD), a.k.a. available-case analysis, is a remedy to the data loss problem of LD by using the means, variances and covariances of all available data for the analyses. When calculating the means, variances and covariances of the data, PD uses all available cases, thus avoiding loss of data problem. After obtaining the moments of the data, one can use estimation method, such as method of moments (MOM), to estimate the coefficients of interest. Although being simple and avoiding the loss of data problem, PD still produces biased estimates and inferences if the data are not MCAR, which is the major shortcoming of the method. In addition, the covariance matrix may not be positive definite especially if the variables are highly correlated.[108] Moreover, due to missing data, using the average sample size for the estimates may yield over confident inferences. In short, PD only works well if the data are approximately multivariate normal, if the correlation between variables are low and if the missing data are MCAR.[99]

In addition to LD and PD, single imputation, including mean imputation, regression imputation (RI) and stochastic regression imputation (SRI), is another category of methods handling missing data. Mean imputation is to replace the missing data using the mean of the observed data. Although the method is simple, it seriously underestimates the variance, distorts the distribution of the data and produces biased estimates and inferences even under MCAR.[99] Compared with mean imputation, RI produces smarter imputations of missing data by incorporating information of the covariates in the imputation. However, RI artificially strengthens the relationships between the variables and systematically underestimates the variability of the imputed data.[99] SRI, which accounts for variability in the imputation by adding a random draw from the residual to the prediction, attempts to address correlation bias. However, the method still cannot fully capture the variability in the missing data and can produce implausible imputations.[99]

Another category of imputation method is donor-based imputation (DoBI), including hot-/cold-deck imputation and nearest neighbor methods. The general idea of DoBI is to impute the missing value of a "recipient" by finding a "donor" who is completely observed and has similar characteristics with the "recipient" and replacing the recipient's missing value with the donor's observed value. Hot-deck method groups the complete observations of a dataset into subsets which share the same values of some matching variables (e.g. age, sex, race etc.). Then, to each observation in the dataset with missing data, a donor is randomly assigned. The cold-deck method is only different from the hot-deck counterpart in that the donors are selected from other comparable data sources instead of the same dataset being filled in.[23] Nearest neighbor (NN) method measures the "distance" between complete and incomplete observations and matches the recipients with donors based on the distance between them. The distance can be calculated using multiple methods but is usually based on the metric matching variables. Usually, the nearest neighbor or one of the k nearest neighbors (KNN) randomly selected is used as the donor for the missing data. The benefits of NN/KNN are that it produces realistic imputations, better reflects the distributional property and can deal with missing data of any type.[23] However, the major limitation of NN/KNN method lies in its heuristic nature since the analyst needs to make many influential but subjective decisions such as the selection of matching variables and the choice of distance measures when using the method.[23,109]

The next category is distribution-based imputation (DtBI), which randomly draws imputation from the empirical (non-parametric) or probabilistic (parametric based on distributional assumptions) distribution of the observed data. Although univariate distributions are most often used for imputation of missing data, a multivariate approach may produce more reasonable combinations of imputed data for multiple incomplete variables.[23] DtBI is the foundation of more

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

Lastly, last observation carried forward (LOCF) and baseline observation carried forward (BOCF) are two ad hoc imputation methods specific for missing data in time series and longitudinal data. The idea of these methods are simple—the last period/ baseline observed value is used to replace the missing data in the following period(s). LOCF and BOCF are commonly used in clinical trials due to its simple and convenient nature. However, LOCF and BOCF produce biased estimates and inferences even under MCAR[110] and thus are not recommended for handling missing data in longitudinal and panel data.[111]

2.3 Multiple Imputation

Since most traditional ad hoc methods handling missing data result in biased estimates and inferences, more appropriate methods have been developed to handle missing data, including full information maximum likelihood (FIML) and multiple imputation (MI).[104,112] Compared with LD and single imputation method, FIML is a more appropriate method coping with missing data as it incorporates information of both observed and missing data into the likelihood function and finds the estimates that maximize this likelihood function.[112–116] However, FIML is available only for certain models such as structure equation models (SEM) and can only be implemented by special software packages.[113] Also, FIML, like listwise deletion, does not impute missing values [112] and thus is not very useful for this study, where the missing country-level proportions of key indicators, along with their uncertainty, need to be imputed.

Among all the techniques handling missing data, MI is the most appropriate method for my purpose in that it imputes the missing values and also accounts for the uncertainty inherent in the imputation.[113,117,118] Multiple imputation was first proposed by Rubin[28] to deal with nonresponses

in surveys. It uses information (e.g. distribution, correlation etc.) of the observed data to estimate likely values of the missing data. MI estimates the missing values m times with each time incorporating a random component to account for uncertainty about the missing values. In the end, we obtain m completed datasets with the same observed data but different imputed missing data. Once m completed datasets are imputed, one can perform identical analyses he/she wants using each of the m completed datasets and then pools the estimates from each dataset together using Rubin's rule.[28,118] Under the assumption of MCAR or MAR, the pooled estimates have been proved to be unbiased and the standard errors are appropriately adjusted.[28,115,117–119] ***Figure 3-2*** presents a graphical flow chart of MI procedure.

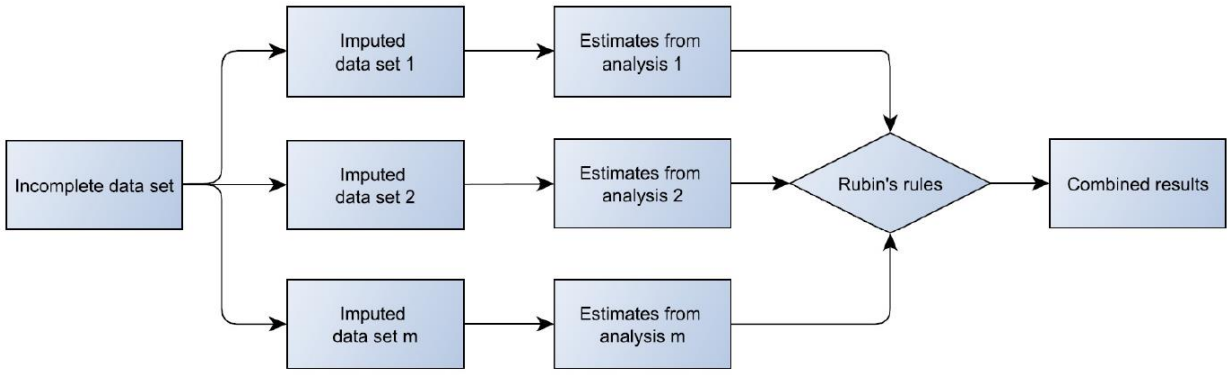


Figure 3-2 The prodecure of multiple imputation [28,118]

Although the MI process is always the same as shown in ***Figure 3-2***, there are different ways of imputing the missing data. Based on the imputation algorithm, there are two major families of MI, namely, MI by joint modeling (MIJM) and MI by Chained Equation (MICE).[32,92,94]

2.3.1 multiple imputation by joint modeling

MIJM assumes that the variables in the imputation model follow a joint distribution, such as a multivariate normal (MVN) distribution.[28,31,117] Under this assumption, the missing values are treated as random draws from the posterior predictive distribution given the observed data.[117]

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

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

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

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

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

This process is iterated and when it is converged, the draws of $\widetilde{D}_{mis}, and \; \tilde{\mu} \; and \; \tilde{\Sigma}$ are from the true posterior independent of their starting values [117].

The advantage of IP is that the algorithm is theoretically justified, which means we are confident that once converged, the draws from the conditional posterior distribution are draws from the joint distribution. Therefore, the predictive posterior distributions are exact. However, the downside of IP is that this algorithm is computationally intensive. In many cases, MCMC can only converge after an infinite number of iterations, which means a long "burn-in period"[117] and the diagnosis of convergence needs expert assessment. In addition, the multiply imputed data needs to be independent in order to use Rubin's rule to pool them together.[28] However, the draws from MCMC are auto-correlated by nature. In practice, people reduce the dependence by using every r^{th} (e.g. 100th) random draws from IP, which further increases the computational burden of the algorithm.[117] Lastly, when the dataset contains different types of variables (e.g.

continuous, binary, categorical and count variables), the current classes of joint models (e.g. MVN model, log linear model, general location model),[120] may not be appropriate for the joint distribution of the data.[121,122]

Expectation Maximization (EM) is a deterministic version of IP. Instead of randomly drawing from the posterior distribution, EM calculates the posterior means and use them as the imputed values. In the E step, missing cells ($\widetilde{D}_{mis}$) are filled with their predicted values, and in the M step, the random draw of $\tilde{\mu}, \tilde{\Sigma}$ are replaced with the maximum posterior estimate.[117,123] The advantages of EM are that it is simple and fast, converges deterministically and can account for fundamental variability in the imputations. However, the serious shortcoming of EM is that it does not account for the uncertainty inherent in estimation of $\tilde{\mu}, \tilde{\Sigma}$.[117] To account for the uncertainty in estimation of $\tilde{\mu}, \tilde{\Sigma}$, King et al[117] propose EMs, EMis and EMB.

Expectation Maximization sampling (EMs) accounts for uncertainty in estimation of $\hat{\theta}$, (i.e. $\tilde{\mu}, \tilde{\Sigma}$) using the asymptotic approximation.[117] After running EM to find the maximum posterior estimates of $\hat{\theta}$, King et al use the outer product gradient or inverse of the negative Hessian to calculate the variance of $\hat{\theta}$, $V(\hat{\theta})$ and then draw $\hat{\theta}$ from a $MVN\left(\hat{\theta}, V(\hat{\theta})\right)$. They then use $\hat{\theta}$ to compute $\tilde{\beta}$ deterministically and use the equation

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

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

number of variables to that of observations is high or when there are highly skewed categorical data.[117]

Expectation Maximization importance resampling (EMis) builds upon the EMs but includes a round of importance resampling, which is a technique to improve small sample performance but not based on Markov chain.[28,61,117,124,125] In addition to all the EMs steps, EMis uses an acceptance-rejection algorithm by keeping draws of $\tilde{\theta}$ from EMs with probability proportional to the "importance ratio"

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

and discarding the rest. The kept draws are considered independent draws from the posterior distribution. The EMis has all the benefits of EMs, such as easy and fast, dose not rely on Markov chain and thus produces full independent imputation. Besides, it also works well for small sample. The posterior produced by EMis also approximates that produced by IP, which is the gold standard for missing data imputation.[117] However, EMis does not work well for all likelihood functions, especially when the normal density is not a good approximation.[117]

Bootstrapped-based EM algorithm (EMB) draws m samples of size n with replacement from the data. In each sample, EM algorithm is implemented to produce estimate of $\tilde{\mu}, \tilde{\Sigma}$. Then for each set of estimates, use the original sample to impute the missing values in their original position. This process produces m multiply imputed datasets.[34] The benefits of EMB compared with EMis and IP are that, it is much faster (computation can be done in a parallel fashion), has better lower order asymptotics than the parametric approaches used by EMis and IP and is more

robust to distributional and small sample problems.[34] The bootstrapped estimates of $\tilde{\mu}, \tilde{\Sigma}$ are very close empirically to those from posterior distribution in large samples.[126]

2.3.2 multiple imputation by chain equation

Besides MIJM, **Multiple Imputation by Chained Equation (MICE)**, also called "fully conditional specification" (FCS) or "sequential regression multiple imputation",[93,113,127] is another major family of MI. Different from MIJM, where the missing values are treated as random draws from posterior predictive distribution given the observed data, MICE reduces the imputation problem to a series of estimations where each variable takes its turn to be estimated using the other variables. This procedure provides great flexibility as each variables can be assigned a suitable distribution, e.g. linear (for continuous variable), Poisson (for count variable), binomial (for binary variable) or multinomial (for categorical variable).[113,128] MICE runs through an iterative process, the detailed steps can be found in the paper of Azur et al.[113] In step 6, noted that a number of iterations (e.g. 10 iterations) are performed to make sure that the distribution of the parameters governing the model have converged and then one imputed dataset was obtained. The whole process is repeated m times producing m imputed complete datasets.[113] The biggest advantage of MICE is great flexibility. One can specify different models for different types of variables and can easily impose bounds and restrictions, such as skip pattern, upon some variables,[93,93,113] which makes it more suitable than joint modeling for some datasets [121]. However, although widely used in medical research, MICE still lacks theoretical justification,[113,118,128] i.e. the implicit joint distribution underlying the separate models may not always exist, that is, the conditional models may be incompatible.[121] When the conditional models are incompatible, the order in which the missing values are updated in the chained equations may seriously affects the results, which is referred to as "order effect".[121] Although

Hughes, et al.[121] and Liu, et al.[129] independently proves the sufficient conditions under which MICE equates MIJM, these conditions are so strict that they are hard to be met in practice.[121,129] Another major downside of MICE is that correctly specifying the model for each variable is very difficult if not impossible and enough auxiliary variables informing the missingness need to be included in the model.[113,118,130] Misspecification of the models can lead to failure of MICE[131] and is the major contributor to biased results.[118,132,133]

Due to the flexibility of MICE, the model for each variable needs not to be parametric, e.g. regression models. Instead, semi-parametric and non-parametric methods can also be used to estimate the missing values. When people use MICE to impute the missing data, two popular non-/semi-parametric methods, namely, random forest (RF) and predictive mean matching (PMM), are often used to estimate the missing values. **MICE with random forest (MICErf)** is a non-parametric technique well-suited for handling complex non-linear relationships. It reduces bias due to model misspecification when model includes complex interactions and polynomial terms.[118,134] Using MICErf, we only need to worry about including all variables (including auxiliary variables) that informs the missingness but do not need to worry about including non-linear terms, such as interactions and polynomial terms, in the model.[118,133] Compared with MICE, MICErf can handle high dimensional data and highly correlated predictors, and it also runs much faster.[118] However, similar to MICE, the biggest drawback of MICErf is that the conditional models may be incompatible.[118] **MICE with predictive mean matching (MICEPMM)** is a semi-parametric method and produces imputed values that resemble the observed values better than other methods because it uses the predicted value for a given observation to identify similar observations. Then the identified similar observations form a matching set and imputed values are randomly drawn from this matching set.[118,128] Therefore,

one benefit of PMM method is that it prevents unrealistic values.[118,135] The major disadvantage of PMM is also lacking of mathematical justification.[118]

2.3.3 multiple imputation for time series cross-sectional data

When first introduced by Rubin et al., MI could not properly handle missingness in multilevel or time series data because it could not account for the cluster structure or the autocorrelation of the data.[28] However, after decades of development, the state-of-the-art MI techniques can properly handle missingness in more complex data such as panel or even TSCS data.[97,136] Schafer et al. were the first to propose a multilevel imputation methods called PAN, which extends the MIJM for multilevel data by specifying a multivariate mixed model to predict the incomplete variables in level-1 variable using the complete level-1 and level-2 variables.[30,31,94] The multivariate mixed model is as follow:

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

where j denotes cluster j, $\boldsymbol{Y_j}$ contains all the incomplete level-1 variables in cluster j, $\boldsymbol{X_j}$ is the matrix of all the complete level-1 and level-2 variables including a unit vector for intercepts. $\boldsymbol{\beta}$ is the matrix of fixed effects of the complete variables which are the same for all clusters. $\boldsymbol{Z_j}$ is the matrix of a subset of complete level-1 variables which have random effects (slopes and intercepts) on the variables in $\boldsymbol{Y_j}$, as well as a vector of 1 for random intercepts. $\boldsymbol{B_j}$ is the matrix of random effects (level-2 residuals) of the variables in $\boldsymbol{Z_j}$ for cluster j. $\boldsymbol{E_j}$ is the matrix of level-1 residuals for cluster j.[32]

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

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

In equation (9), the multivariate normal distribution is centered at the predicted value from the imputation model with spread of $\boldsymbol{\Sigma}$, which is the level-1 residual covariance matrix. A scaler version of equation (9) is detailed by Mistler[32] and thus not described here. The parameters in equation (9) can be estimated using IP algorithm. The detailed estimation and sampling steps can be found in many literature[94,137–139] and thus are not described here. In equation (9), $\boldsymbol{\Sigma}$ captures the entire relationships between incomplete level-1 variables and is the same across all clusters, meaning that the random effects of the incomplete level-1 variables on other incomplete level-1 variables cannot be captured in the PAN method.[32,95] Therefore, if the analysis model contains the random effects between incomplete level-1 variables, PAN imputation method is said to be uncongenial with the analysis model.[32,120,140]

To allow the PAN method to account for the random effects between the incomplete level-1 variables, Yucel proposed an improved PAN method, called random-covariance and mixed-effect (RCME) model, which allows the residual covariance matrix $\boldsymbol{\Sigma}$ to randomly vary across different clusters and thus preserves the important relationships between the incomplete variables in the imputation model.[95,141] This RCME model can be implemented by R package JOMO[142] and thus we call it JOMO method.

Although the PAN and JOMO methods are not specifically developed for TSCS data, they can be used to handle missingness in TSCS data by including time (e.g. year) in the imputation model and allowing the effect of time on the incomplete variables to vary across the clusters (e.g. countries). In addition to PAN and JOMO, King et al.[34] developed another MIJM method called *Amelia* to handle missingness in TSCS data in particular. Different from PAN and JOMO, which employ a full Bayesian method, i.e. IP, to estimate the parameters of the posterior predictive joint distribution, Amelia uses EMB algorithm to estimate the parameters of the posterior.

Compared with IP, the EMB algorithm is much faster, can handle more variables, and produces empirically similar results to those produced by IP in large samples.[34] Under the assumption that time series variables often have smooth trends over time, Amelia included smooth basis functions, such as polynomials or splines, of time in the imputation model and allows the basis functions of time to interact with country indicators to account for heterogeneity of trends across countries.[34] Another advantage of Amelia is its ability to easily incorporate one's prior knowledge about the missing data into the imputation process. Instead of specifying the priors for the abstract model parameters, Amelia allows one to incorporate prior knowledge of the missing data using either (1) a mean plus a standard error of the missing value or (2) a confidence interval of the missing value.[143]

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

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

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

effects (i.e., level-2 residuals) of the variables in $\boldsymbol{Z}_{jk}$ on $\boldsymbol{y}_{jk}$. $\boldsymbol{e}_{jk}$ is the vector of level-1 residuals for cluster j, variable k.[32]

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

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

In equation (11), the conditional univariate normal distribution is centered at the predicted value from the imputation model for variable k, with spread equal to the level-1 residual variance of k (i.e., σ_k^2).[32] Noted that the set of random effect variables predicting each incomplete variable k can be different, which allows for inclusion and exclusion of certain random effects when necessary. The parameters in equation (11) can be estimated by an MCMC algorithm which is described in many literature[94,99,137–139] and thus is not detailed here.

Built upon on the above model, Carpenter and Kenward[141] recommended to include the cluster means of level-1 variables as predictors in the univariate imputation models. Mistler and Enders[144] confirmed the benefits of such adaptation and found that inclusion of cluster means of level-1 variables greatly improved the performance of MICE in all scenarios. However, a recent study done by Resche-Rigon and White[145] found that including cluster means of level-1 variables to the imputation models had little impact on the performance of MICE though did not hurt either.

Similar to MIJM methods, such as PAN and JOMO, MICE can also handle missingness in TSCS data by including time (e.g., year) in the univariate imputation model for each of the incomplete variables. The MI methods mentioned above, namely, PAN[146], JOMO[142], Amelia[147] and MICE[148], can all be implemented in R.

2.4 Evaluation of the Performance of MI

According to Chambers, there are four types of accuracy of imputations. Ranked from the hardest to the easiest to achieve, the four types of accuracy are predictive accuracy, ranking accuracy, distributional accuracy and estimation accuracy.[149] The strongest predictive accuracy requires the maximal preservation of true values, which implies the other three types of accuracy; ranking accuracy requires maximal preservation of the order of true values; distributional accuracy requires the maximal preservation of distributions of the true values (e.g. the marginal and higher order distributions); and the weakest estimation accuracy only requires the reproduction of lower order moments (e.g., mean and variance) of the true values.[23,149]

Evaluation of the performance of MI depends on one's objective of using the MI method. According to Barnard and Meng[150], the application of MI falls within two categories, the "outside" and "in-house" application. The traditional application of MI is mostly "in-house" where the imputer and the analyst is the same person and the imputation is done for one specific analysis, the so-called "one-analyst-one-goal" studies. However, in "outside" application, the imputer and the analyst are typically different persons and the main objective of MI is to accurately impute missing data and thus to produce multiple complete datasets to fit for the "many-analysts-many-goals".[26,102]

For "in-house" application, due to the nature of "one-analyst-one-goal", we only care about a single or a few parameters of the population, e.g., the mean and variance of some variables or the coefficients of a regression in the population. We only ask how well the MI method helps us estimate these population parameters. Therefore, we only require estimation accuracy of the MI method. Dr. Rubin[151] even stated that the objective of multiple imputation should rather be statistically valid inference (i.e. estimation accuracy) than the optimal point prediction (i.e.

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

However, different from "in-house" application, the "outside" application requires predictive accuracy of MI to achieve "many-analysts-many-goals". Since the predictive accuracy implies the other three types of accuracy, achieving predicative accuracy makes sure that the imputed datasets can be used for a variety of analyses.[149] To some extent, the "outside" application uses MI to accurately "predict" the missing values while accounts for uncertainty of the "predictions" by imputing the missing values multiple times.[26,102] As a results, for "outside" application, the performance is evaluated by how close the imputed data are to the true data.

Although the evaluative targets of MI are different under different applications, the methods used to evaluate the "closeness" of estimated parameters of the population (for "in-house" application) or imputed data (for "outside" application) to the true values are the same. Since the true values are unobserved, a simulation method is widely used to evaluate such "closeness" for imputation methods.[92,152–154] Simulation method artificially "generates" missingness by removing some of the observed data from the original dataset. The missingness can be generated under different assumptions, i.e., MCAR, MAR and MNAR.[92,144] The original dataset can be either artificially generated[32,144,152,153] or derived from a real world dataset.[92,154] Under "in-house" application, an MI method is considered good if the MI estimates and inferences of regression coefficients using the imputed datasets are close to the estimates and inferences using the original complete dataset.[32,92,144,152–154] Whereas, under the "outside" application, an MI method is considered good if the imputed data are close to the observed true data.[155,156] The common performance indicators are (1) bias, which is defined as the difference between an estimator's average value and its true value, (2) root mean squared error (RMSE), which is the square root of the mean squared

difference between average estimates and true values, and (3) the coverage rate of 95% confidence/credible interval, which is the proportion of the confidence/credible intervals covering the true values.[152] These three indicators can be used to evaluate the performance of MI methods under both applications.

3. Methods

3.1 Data Description

3.1.1 country-level estimates of the key indicators

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

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

performance. Therefore, the imputed variables are also in the logit scale and thus need to be back transformed to its original scale afterwards.

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

To improve the performance of imputation, important country-level covariates (**Table 3-1**) of HIV/AIDS knowledge and attitudes are included in the imputation model. The country-level covariates are extracted from IHME's GBD study 2017 and are complete over the period from 2000 to 2017. To improve the performance of imputation models, we logit-transform *contra_prev*, *ASFR*, *prop_urban*, *ANC4* and log-transform *GDP*.

Table 3-1 List of country-level covariates

Covariate	Description	Type
education	Mean years of education per capita (by sex)	Continuous
GDP	GDP per capita base 2010 international dollars	Continuous
ASFR	Age-specific fertility rate	Proportion
contra_prev	Modern contraception prevalence in women by age groups	Proportion
HAQI	Healthcare access and quality index [37]	Continuous
prop_urban	Proportion of population living in urban area	Proportion
Muslim	Binary indicator: value 1 if country is greater than 50% Muslim	Binary
HSA	Health system access: a composite score of immunization, measles immunization, hospital beds, in-facility delivery and skilled birth attendance	Continuous
ANC4	Proportion of pregnant women receiving 4 or more antenatal care from a skilled provider.	Proportion

3.1.3 covariates of survey and estimates

In addition to covariates of HIV/AIDS knowledge and attitudes, covariates of surveys and estimates, e.g., year and country of the survey, sex and gender of the estimates, are also important covariates to be included in the imputation model. **Table 3-2** describes these covariates in detail.

Table 3-2 List of covariates of the survey and the estimates

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

3.2 Multiple Imputation Methods to be Evaluated

In this study, we examine and compare the performance of 7 different MI algorithms for TSCS missing data, namely, ***PAN***, ***JOMO***, ***Amelia*** and four ***MICE*** *algorithms* with different univariate modeling methods. As described in the literature review, PAN, JOMO and Amelia are all MIJM. PAN and JOMO use full Bayesian MCMC (the **IP**) algorithm to estimate the parameters of the posterior and to draw imputed values from the posterior[95,141,157] whereas Amelia uses EMB algorithm to estimate the parameters of the posterior.[136] Different from PAN, which assumes fixed level-1 residual covariance matrix across different clusters, JOMO relaxes the assumption by allowing the covariance matrix of residuals at level 1 to vary across clusters.[95,141] The heterogeneous covariance matrix of residuals helps better capture the random effects between incomplete level-1 covariates across clusers.[32,95] However, the JOMO algorithm is very computationally intensive and is expected to take much longer time to run compared with the PAN algorithm.[32] To implement PAN, JOMO and Amelia, R functions "panImpute" and "jomoImpute" in package "mitml"[158] and function "amelia" in package "amelia"[147] are used.

As described in the literature review, MICE is more flexible than MIJM because each incomplete variable is modeled separately and in turn in MICE. Since the incomplete variables to be imputed

are all two-level continuous variables, four univariate models for two-level continuous variable are chosen to impute the missing data, namely, ***mice.2l.pan***, ***mice.2l.norm***, ***mice.2l.lmer*** and ***mice.2l.pmm***. The methods *mice.2l.pan* and *mice.2l.norm* impute the univariate missing data using a two-level normal model with homogenous and heterogeneous within group variance respectively.[99] They both implement the Gibbs sampler to fit the two-level normal model (see details in ***section 2.3.3***). The *mice.2l.lmer* method uses univariate linear mixed model (using R function "lmer") to predict the univariate missing data. The predictions take into account the uncertainty of the model parameters, the random effects and the model residuals.[99,159] Based on *mice.2l.lmer* ,the *mice.2l.pmm* uses predictive mean matching based on predictions from the linear mixed model above. For each missing value, 5 donors are selected based on proximity to the predicted values and one of the 5 donors is randomly selected as the imputed value.[99,118] All the four MICE methods are implemented using R function "mice" in the "mice" package.[160]

3.3 Imputation Models

When comparing the performance of different MI methods, we use the same imputation model which includes all the 16 key indicators of HIV/AIDS knowledge and attitudes, the important country-level covariates (**Table 3-1**) and the covariates of the survey and the estimates (**Table 3-2**). Among these variables, the 16 key indicators are the target variables which contain missing values and the other variables are all auxiliary variables which are completely observed. Among the variables included in the imputation models, *location_id* is the cluster variable and *year_id* is the time variable whose slope is allowed to vary across clusters. Therefore, the primary imputation model is a two-level model with random slope and random intercept.

As discussed in the literature review, there are still discrepancies among researchers on whether including cluster means of variables improves imputation.[141,144,145] Therefore, building upon the

model above, we further include the cluster means of key indicators and of covariates to examine the impact of adding these cluster means on the imputation performance. In addition, it is recommended that one can include as many auxiliary variables as possible to make the MAR assumption more plausible, which is the so-called inclusive strategy.[161] However, study also shows that including auxiliary variables that have too many missing values harms the efficiency of imputation.[162] Therefore, building upon the primary model, we further include auxiliary variables (other HIV/AIDS knowledge and attitudes indicators) with different proportion of missingness to examine the impact of adding these incomplete auxiliary variables on the imputation.

To examine the impact of two additional modeling strategies mentioned above, we first pick the best-performing MI methods for the primary imputation model and then use these methods to conduct MI with two additional imputation models including cluster means and additional incomplete auxiliary variables respectively.

3.4 Implementation of the MI Methods

For all the MI methods, we use *location_id* and *year_id* as cluster and time variable respectively, model the time effect linearly, allow the time effect to vary across countries, and impute the missing values 1000 times. For **Pan** and **JOMO**, we set the burn-in to be 1000 iterations and draw one imputed value every 100 iterations to make sure that the draws are independent. For **Amelia**, we follow the advice by Honaker et al. and add a small ridge prior (1% of the total number of observations) to stabilize the imputation algorithm.[163] To reduce the running time, we use 1000 machines to run Amelia parallelly and combine the results afterwards. For the four **MICE** methods, we impute the variables in a monotone sequence and use 20 iterations for each imputation. The random seed is set to be 2019 for all the random processes.

3.5 Evaluation Methods

3.5.1 10-fold cross-validation

To evaluate the performance of different MI methods and imputation models, we employ a 10-fold cross-validation (CV) approach. In a 10-fold CV, the observed data are randomly divided into 10 mutually exclusive subsets (the folds) of approximately equal size. The imputation model is trained and tested 10 times; each time, one subset of observed data is left out and used as test set and the imputation model is trained using the remaining 9 subsets of observed data.[164]

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

To simulate the missing data for 10-fold CV, we first randomly divide the observed data of each key indicator into 10 groups and then we randomly select one group from each indicator and remove the observed data in the selected groups. The selection of groups is repeated 10 times, resulting in 10 datasets with different simulated missing data. Since the groups are selected without replacement, each group will only be selected once and no observed data will be removed twice. The missing data are MCAR.

3.5.3 indicators of performance

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

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

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

where y_i^{mis} is the removed true data value for unit i, $\hat{y}_i$ is the imputed value for unit i and n_{mis} is the total number of true values that are artificially removed. The $RMSE$ measures the average deviance of imputed values from true values and takes into account the trade-off between unbiasedness and efficiency of the estimator by combining information about both bias and variance of the estimator.[152] In general, an imputation method is considered to be better if the $RMSE$ of its imputed values is smaller. It is a commonly used performance indicator for imputation methods.[155,156] Although we impute the missing proportions of key indicators in logit scale, we calculate the RMSE in the original scale of the proportions, i.e. between 0 to 1. Therefore, in this study, a RMSE is considered good if it is smaller than 0.05 suggesting that the imputation method has less than ±5% imputation error on average.

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

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

where y_i^{mis} is the removed true data value for unit i, n_{mis} is the total number of true values that are artificially removed, $\hat{y}_{2.5^{th}}$ and $\hat{y}_{97.5^{th}}$ are 2.5th and 97.5th percentile of the imputed values of unit i. As mentioned above, the CR_{95} should be close to 95% by definition. According to Grund et al., the MI method is considered suboptimal if CR_{95} is below 90% or very close 100%, suggesting that the distribution of imputed values are off or the variance of imputed data are too large.[152]

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

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

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

3.5.4 comparison of model performance

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

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

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

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

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

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

4. Results

4.1 The Pattern of Missing Data

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

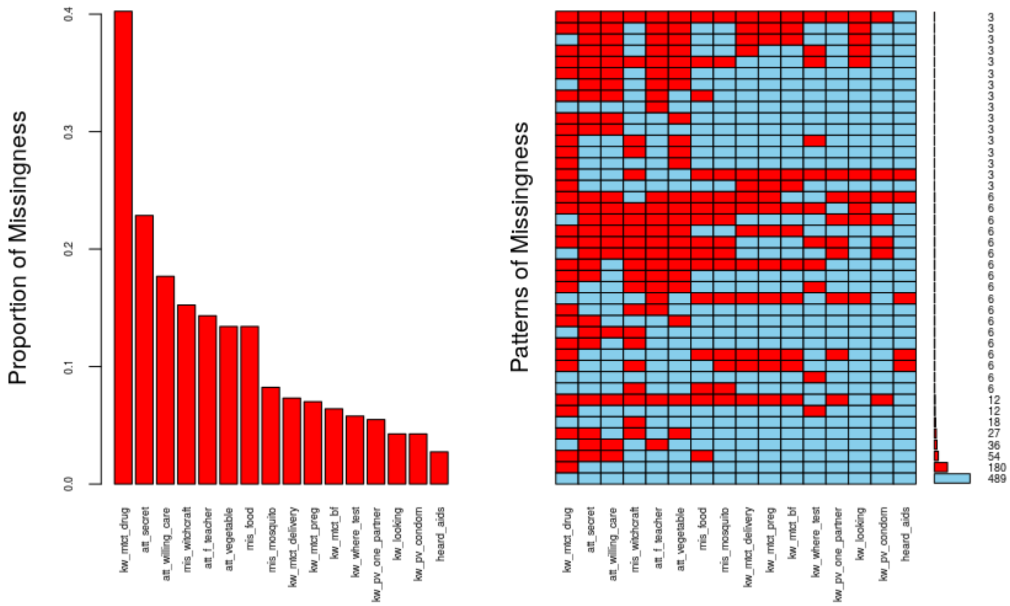


Figure 3-3 Proportion of missingness by variables and patterns of missingness

After removing 10% of observed data for each indicator, the overall missing rate of the simulated dataset becomes 20.7%. The number and proportion of missingness for each indicator in the simulated dataset are summarized in **Table 3-3**. In the simulated dataset, all indicators have more than 10% missingness and 7 indicators have more than 20% missingness.

Table 3-3 Number of missing data by variables

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

att_willing_care	174	17.7%	255	25.9%
att_secret	225	22.9%	301	30.6%
kw_mtct_drug	396	40.2%	455	46.2%

4.2 The Performance of MI Methods Using the Primary Imputation Model

The overall $RMSE$ and CR_{95} for each method are summarized in **Table 3-4** and the variable-specific $RMSE$ and CR_{95} are summarized in **Table 3-5** and **Table 3-6** respectively.

Table 3-4 The overall performance indicators for all MI methods

MI methods	RMSE	CR_{95}	ART (min)	PS	PS^t
Amelia	0.0391	0.9520	5	**0.6688**	0.6155
pan.100[1]	0.0378	0.9465	15	**0.7406**	0.6960
jomo.100	0.0204	0.9556	7053 (5.9 days)	**0.6226**	11.2523
mice.2l.pan	0.0378	0.9475	2475 (1.7 days)	**0.6808**	4.3689
mice.2l.norm	0.0681	0.9583	5585 (3.9 days)	1.4489	9.7793
mice.2l.pmm	0.1050	0.9840	2939 (2.0 days)	3.5074	7.6415
mice.2l.lmer	0.1714	0.9467	1178 (0.8 days)	2.5985	4.1471

Table 3-5 The variable-specific $\boldsymbol{RMSE}$ for all MI methods

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

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

kw_mtct_drug	0.050	0.046	0.029	0.046	0.085	0.109	0.216

Table 3-6 The variable-specific $\boldsymbol{CR_{95}}$ for all MI methods

Indicators	**Amelia**	**Pan.100**	**Jomo.100**	**mice. 2l.pan**	**mice. 2l.norm**	**mice. 2l.pmm**	**mice. 2l.lmer**
heard_aids	0.927	0.930	0.948	0.933	0.951	0.959	0.936
kw_looking	0.954	0.949	0.963	0.948	0.962	0.989	0.948
kw_pv_condom	0.956	0.948	0.952	0.941	0.969	0.983	0.948
kw_pv_one_partner	0.941	0.944	0.951	0.945	0.953	0.980	0.944
kw_where_test	0.947	0.946	0.962	0.946	0.947	0.992	0.945
kw_mtct_bf	0.961	0.956	0.955	0.955	0.956	0.990	0.944
kw_mtct_preg	0.946	0.947	0.954	0.947	0.960	0.969	0.951
kw_mtct_delivery	0.957	0.950	0.953	0.951	0.962	0.995	0.948
mis_mosquito	0.959	0.952	0.948	0.950	0.945	0.981	0.951
att_vegetable	0.961	0.955	0.957	0.959	0.953	0.989	0.947
mis_food	0.961	0.950	0.958	0.957	0.957	0.993	0.946
att_f_teacher	0.961	0.953	0.956	0.954	0.968	0.989	0.949
mis_witchcraft	0.954	0.941	0.956	0.937	0.964	0.988	0.952
att_willing_care	0.949	0.952	0.957	0.953	0.967	0.975	0.942
att_secret	0.937	0.930	0.961	0.938	0.963	0.984	0.947
kw_mtct_drug	0.963	0.949	0.966	0.949	0.961	0.988	0.949

We can see from **Table 3-4** that among the 7 MI methods, JOMO has the smallest $RMSE$ (0.0204) and the smallest PS (0.537), suggesting that JOMO method performs the best among the 7 methods. However, the ART for JOMO method is 7053 minutes or 5.9 days, making this method immensely impractical. Based on the indicator PS^t, which takes into account the running time, JOMO is ranked the last among the 7 methods. Although having $RMSE$ higher than that of JOMO, Amelia, mice.2l.pan and PAN all have $RMSE$ smaller than 0.05 and CR_{95} very close to 95%. The PS scores for Amelia, mice.2l.pan and PAN are 0.642, 0.673 and 0.821 respectively, suggesting that these three methods also perform very well imputing the missing values in the dataset. However, the rest three methods, namely, mice.2l.norm, mice.2l.pmm and mice.2l.lmer, all have $RMSE$ greater than 0.05 and their PS scores are much higher than those of the other

methods, suggesting that these three methods do not perform well imputing missing values in the dataset.

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

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

Table 3-5 and **Table 3-6** summarize the variable-specific $RMSE$ and CR_{95}. The indicators are sorted based on proportion of missingness of the indicators, ranked from the top to the bottom from low missingness to high missingness. Among the 16 key indicators, *heard_aids* has the smallest $RMSE$ and *att_secret, kw_where_test* and *kw_mtct_drug* have the largest $RMSE$. In general, the higher the proportion of missingness, the larger the $RMSE$ is because higher proportion of missingness usually leads to higher variance of the imputation. However, as shown in **Table 3-5**, this is not always the case because $RMSE$ not only depends on the variance but also depends on the bias. For instance, if the distribution of an indicator is far from normal, e.g. highly skewed, the bias of the imputations would be high. ***Figure 3-4*** compares the density of *kw_where_test* and *kw_mtct_bf.* Although the two indicators have similar missing rates of 15.2% and 15.9% respectively in the simulated dataset, *kw_where_test* has higher $RMSE$ than *kw_mtct_bf* does because the distribution of *kw_where_test* is more skewed.

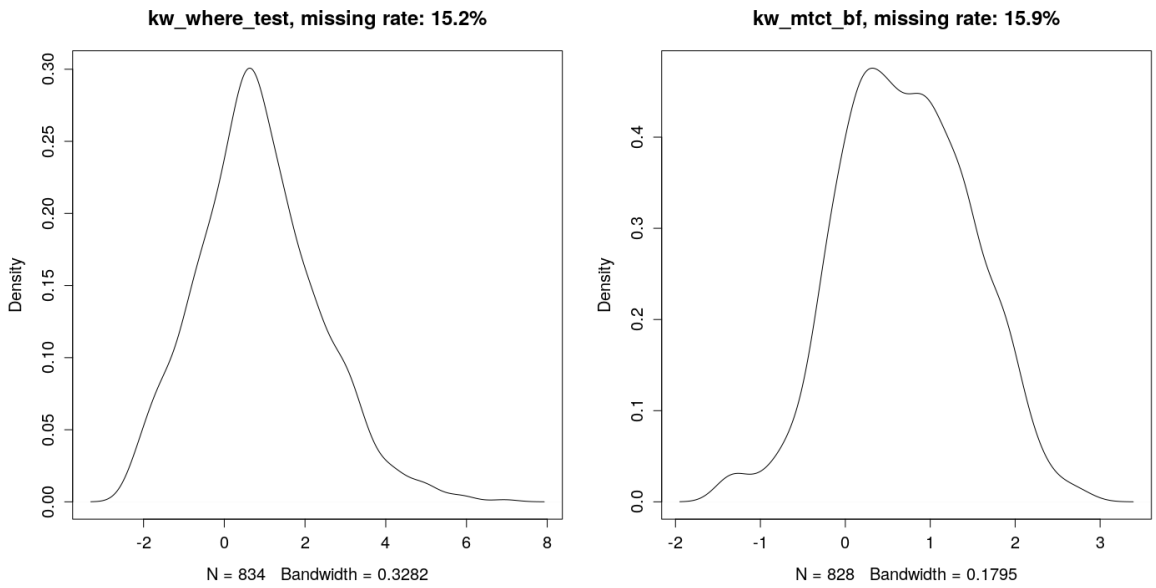


Figure 3-4 The density plots of kw_where_test and kw_mtct_bf

4.3 Diagnostics of the Imputation Methods

Although the $RMSE$ and CR_{95} are already important diagnostics for the imputation methods, other diagnostics are still informative of the performance of imputation methods. There are common diagnostics for all imputation methods such as density plots of observed and imputed values. There are also method-specific diagnostics, such as potential scale reduction factor ($\hat{R}$) and trace plots of the MCMC chain to examine convergence for PAN and JOMO method. In this section, we provide important diagnostics for each imputation method.

4.3.1 diagnostics of Amelia methods

Figure 3-5 shows the density plots of observed (in black) and of imputed (in red) data for each indicator using Amelia. We can see that the distributions of the imputed data are very close to those of the observed data for nearly all the indicators, suggesting that Amelia produces valid imputations. For *kw_mtct_drug,* the mean of the observed and of the imputed data are slightly

different, suggesting that imputations for this indicator are a bit off. This is probably due to the high missing rate (46.2%) of *kw_mtct_drug*.

Figure 3-6 shows the disperse plot which is a visual diagnostic of EM convergence. In the disperse plots, the EM chain is started at 5 different places and we can see that the Amelia EM algorithm converges well in both one and two dimensional spaces.

Figure 3-7 shows the overimputation plots for the key indicators in which each observed value is treated as missing and is imputed using the imputation model. In plots, the dots are the mean imputation and the vertical lines are the 90% confidence intervals. Ideally, 90% of the vertical lines should across the diagonal lines where the imputed values equal to the observed values. Based on the overimputation plots, we think the Amelia method works pretty well.

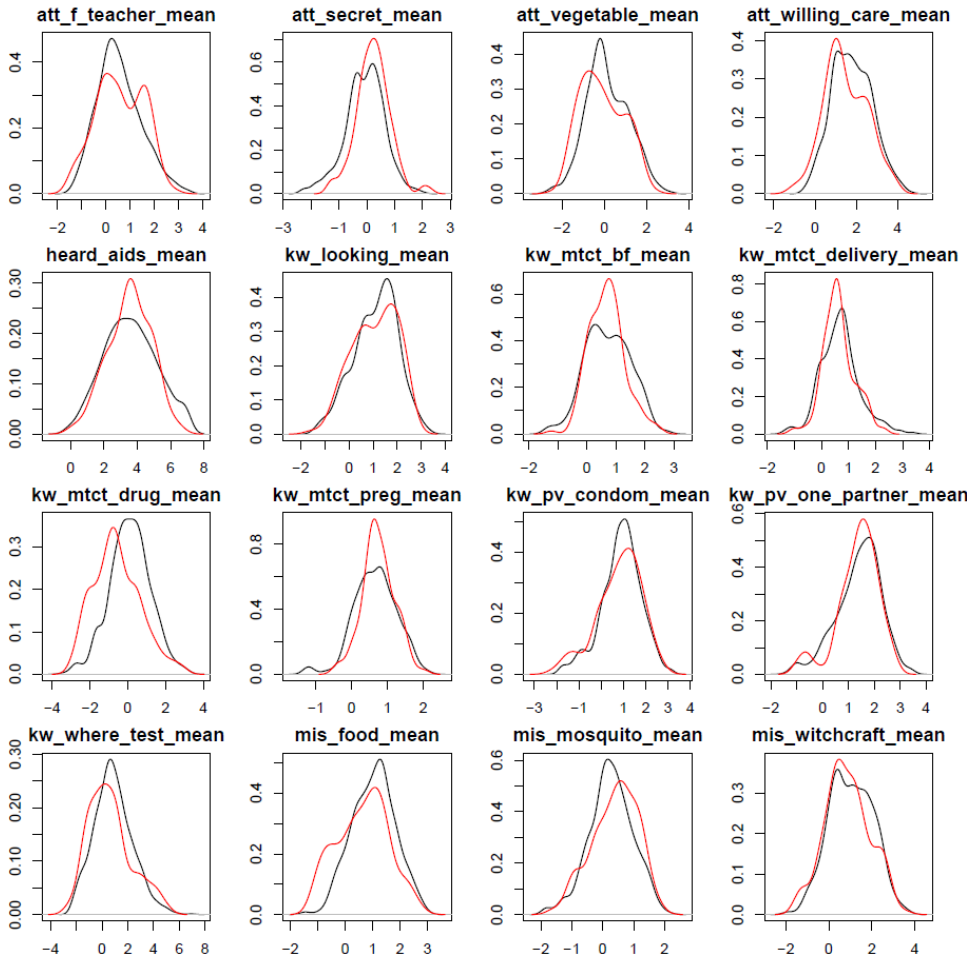


Figure 3-5 The density plots of observed (in black) and imputed (in red) values for each indicator using Amelia method

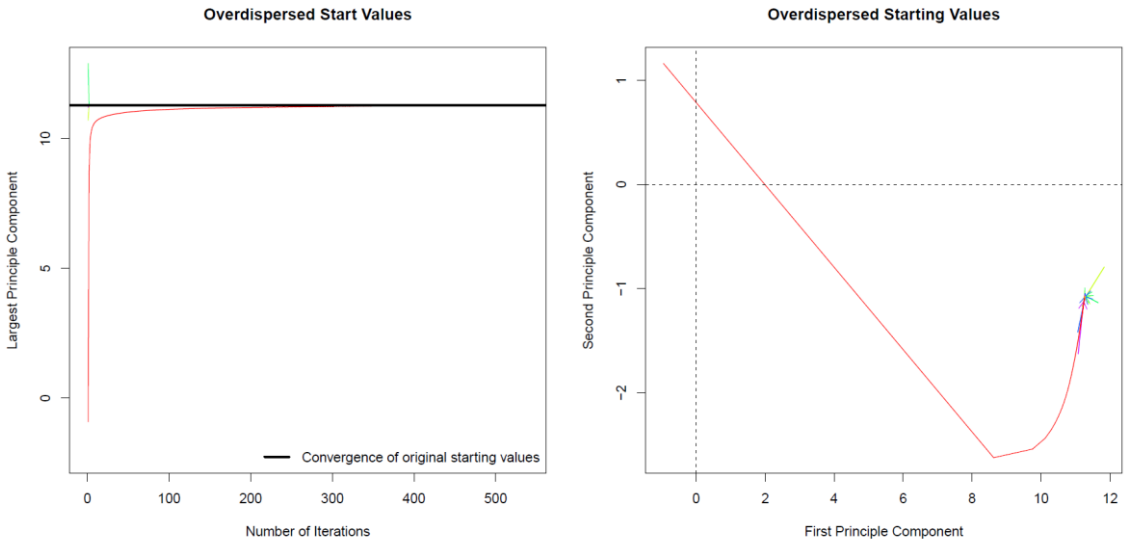
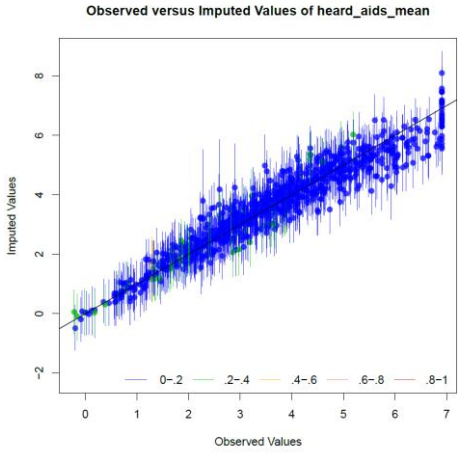
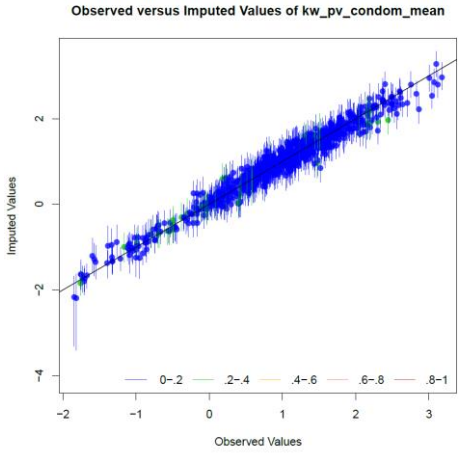
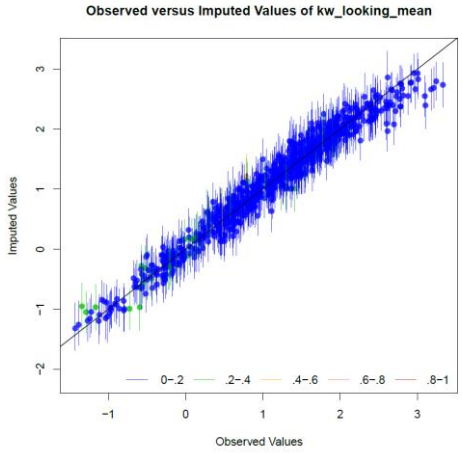
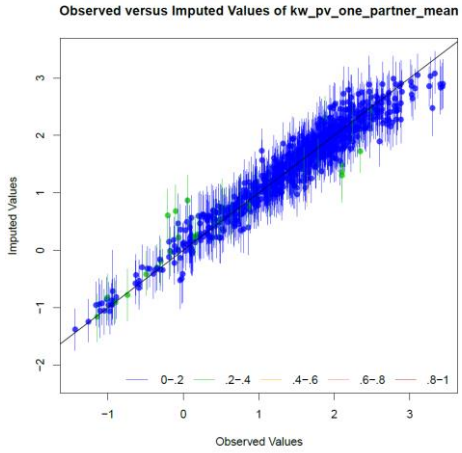


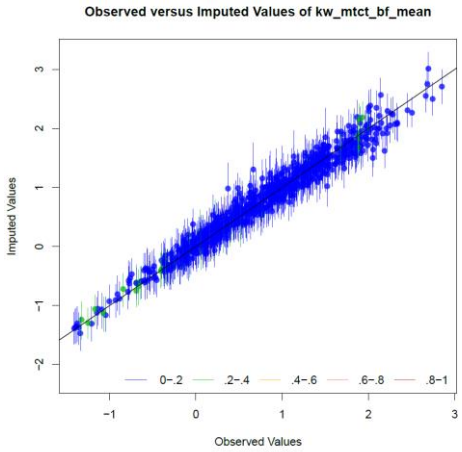
Figure 3-6 Disperse plots of one- and two-dimensional EM convergence

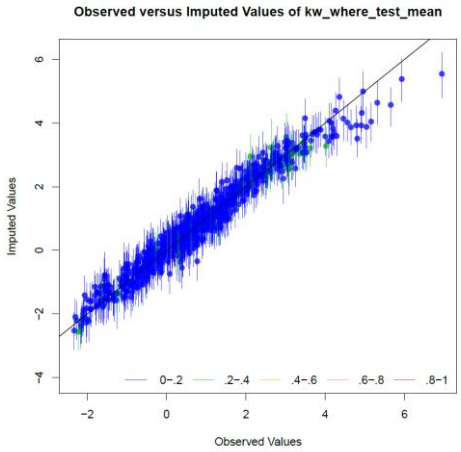

Observed versus Imputed Values of heard_aids_mean
Imputed Values
Observed Values
0−.2
.2−.4
.4−.6
.6−.8
.8−1

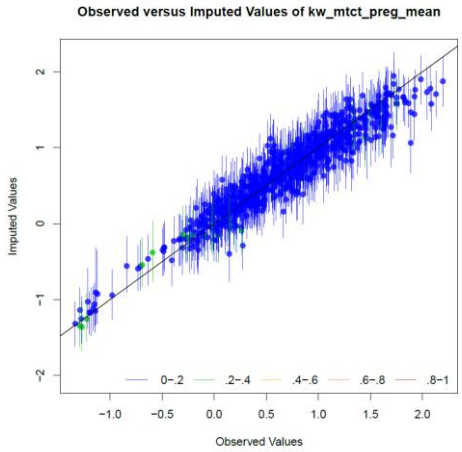

Observed versus Imputed Values of kw_pv_condom_mean
Imputed Values
Observed Values
0−.2
.2−.4
.4−.6
.6−.8
.8−1

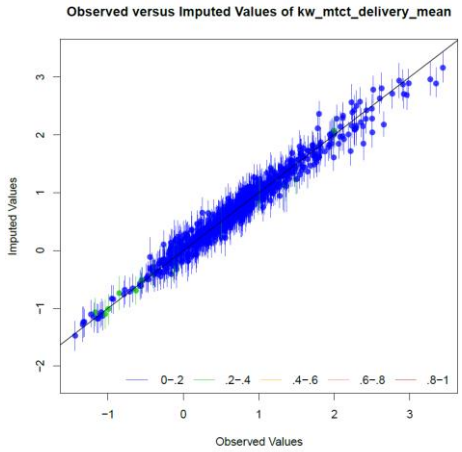

Observed versus Imputed Values of kw_looking_mean
Imputed Values
Observed Values
0−.2
.2−.4
.4−.6
.6−.8
.8−1


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


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


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


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


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

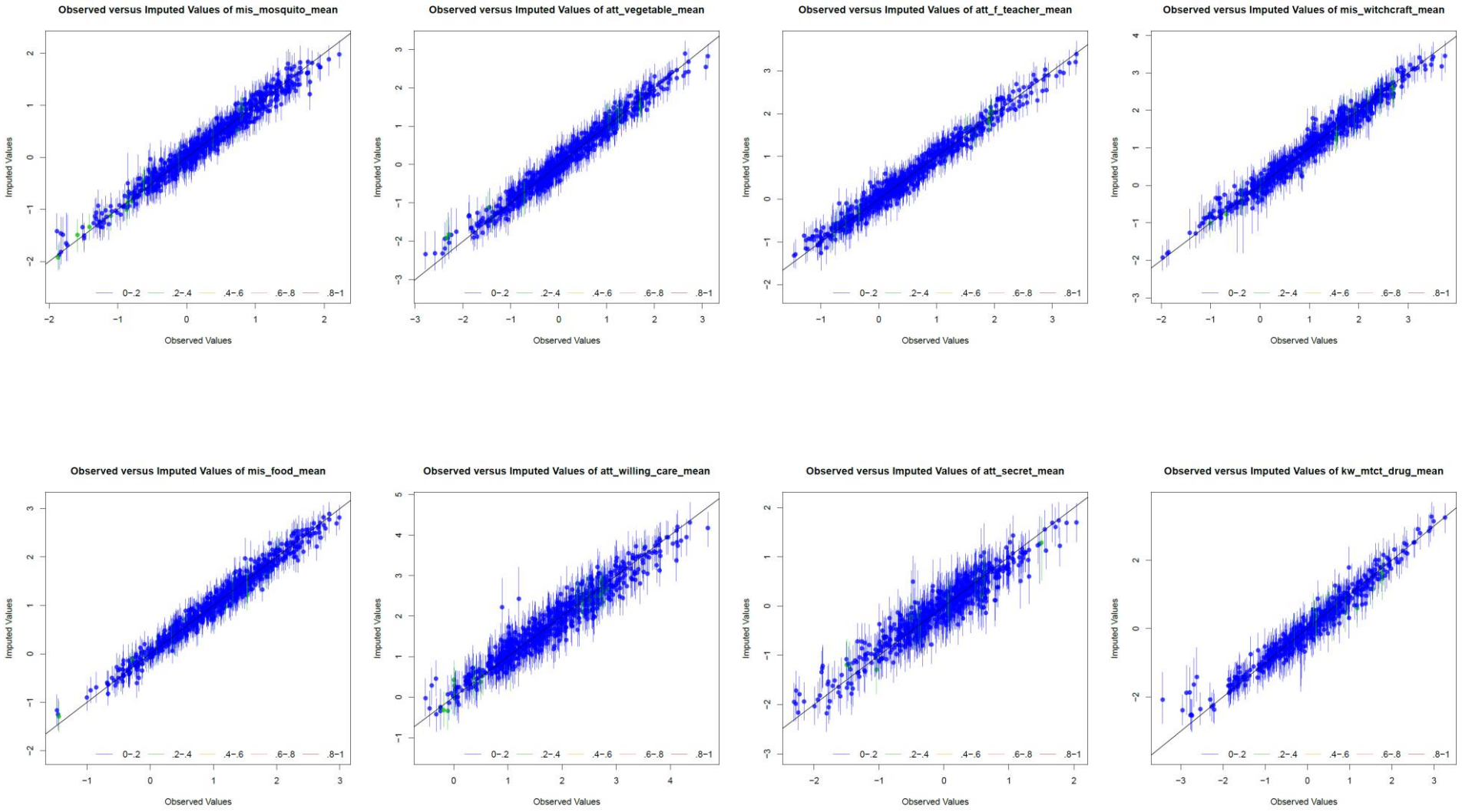


Figure 3-7 Overimputation plots for each key indicator

4.3.2 diagnostics of PAN and JOMO methods

Since PAN and JOMO use full Bayesian MCMC method to impute the missing data, the convergence of MCMC chains is important. In addition, draws from MCMC chain are inherently dependent especially for draws close to each other. However, for each missing value, the multiply imputed values should be independent draws.[117] To overcome the two problems, we discard the first 1000 draws hoping that the chains have converged after 1000 burn-ins and we draw an imputed value every 100 draws in the MCMC chain hoping that the imputed values are independent. After the imputation, we examine convergence of the chains using potential scale reduction factor ($\hat{R}$) and trace plots and examine independence of the imputed values using auto-correlation function (ACF) plots of the imputed values.

To calculate $\hat{R}$ we first discard the burn-in periods and divide the MCMC chain for each parameter into five segments. We then compare the variance within and between segments to detect shift of the chain. If the MCMC chain has converged, the $\hat{R}$ should be very close to 1. Practically, if $\hat{R} < 1.05$ for all parameters, we think the chains have converged.[157,165] **Table 3-7** shows the summary of $\hat{R}$ of different PAN and JOMO methods. In **Table 3-7**, pan.100 (jomo.100) and pan.200 (jomo.200) represent PAN (JOMO) with 1000 burn-ins and thinning factor of 100 and with 2000 burn-ins and thinning factor of 200 respectively; Beta, Psi and Sigma represent parameters of variables in the imputation model, variance and covariance of the random effects and variance of the residuals respectively. The maximum $\hat{R}$ of the parameters is way larger than 1.05 across all models, suggesting that the burn-ins are not enough and none of the models has converged. **Table 3-8** shows the summary of autocorrelation of the parameters in four models. In **Table 3-8**, k represents the number of iterations per imputation (the thinning factor). We can see that the autocorrelation between adjacent draws of the parameters are high.

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

Table 3-7 Summary of $\widehat{\boldsymbol{R}}$ of different PAN and JOMO methods

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

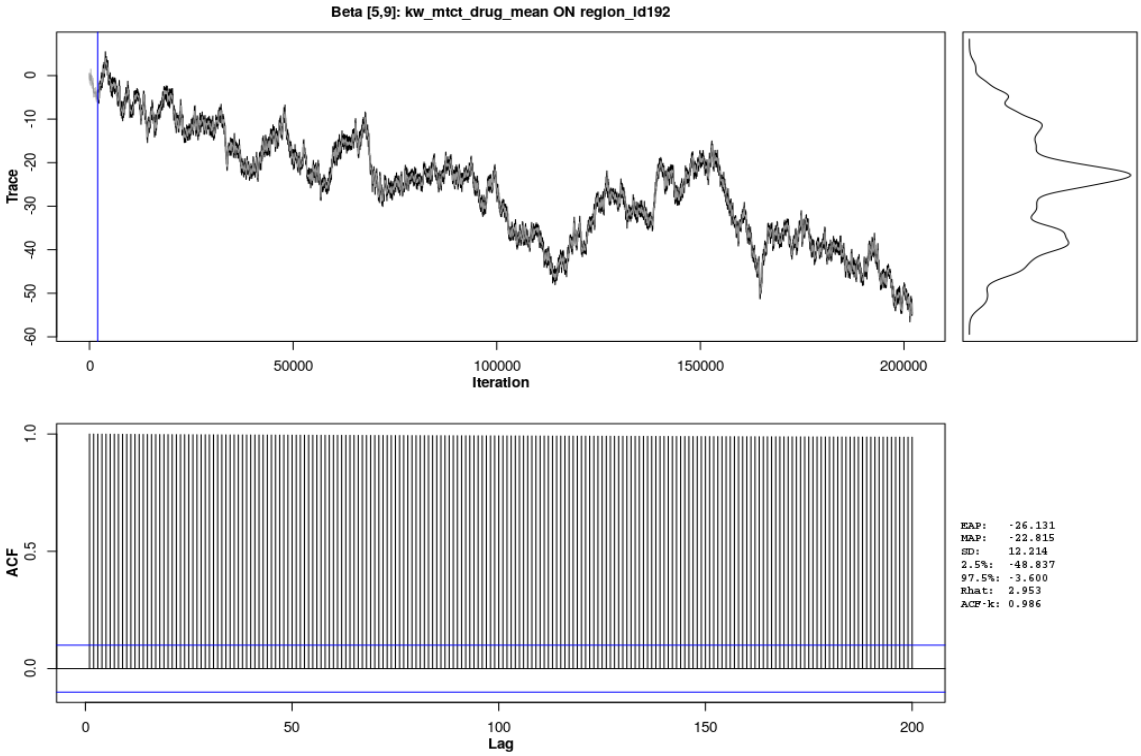
Table 3-8 Summary of ACF of different PAN and JOMO methods

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

Although $\hat{R}$ is useful, Geyer argues that large $\hat{R}$ does not necessarily indicates poor convergence and examining the trace plots of the parameters is still important.[166] ***Figure 3-8*** shows trace and ACF plots of the parameters with the largest $\hat{R}$ for pan.200 (top) and jomo.200 (bottom)

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

Based on $\hat{R}$, ACF and trace plots, compared with PAN, JOMO would require longer burn-ins and larger thinning factor to reach convergence and to produce independent imputations. Given the already long running time of JOMO (over 9 days for jomo.200), it seems impractical for JOMO to reach convergence and to produce independent imputations. Therefore, we run PAN again with larger burn-ins and thinning factor. Inspired by Grund et al. who implemented PAN with 50000 burn-ins and thinning factor of 5000 (pan.5k) and found that the model converged well and produced independent imputaitons,[157] we tried to run PAN with the same parameters hoping that the model will converge.



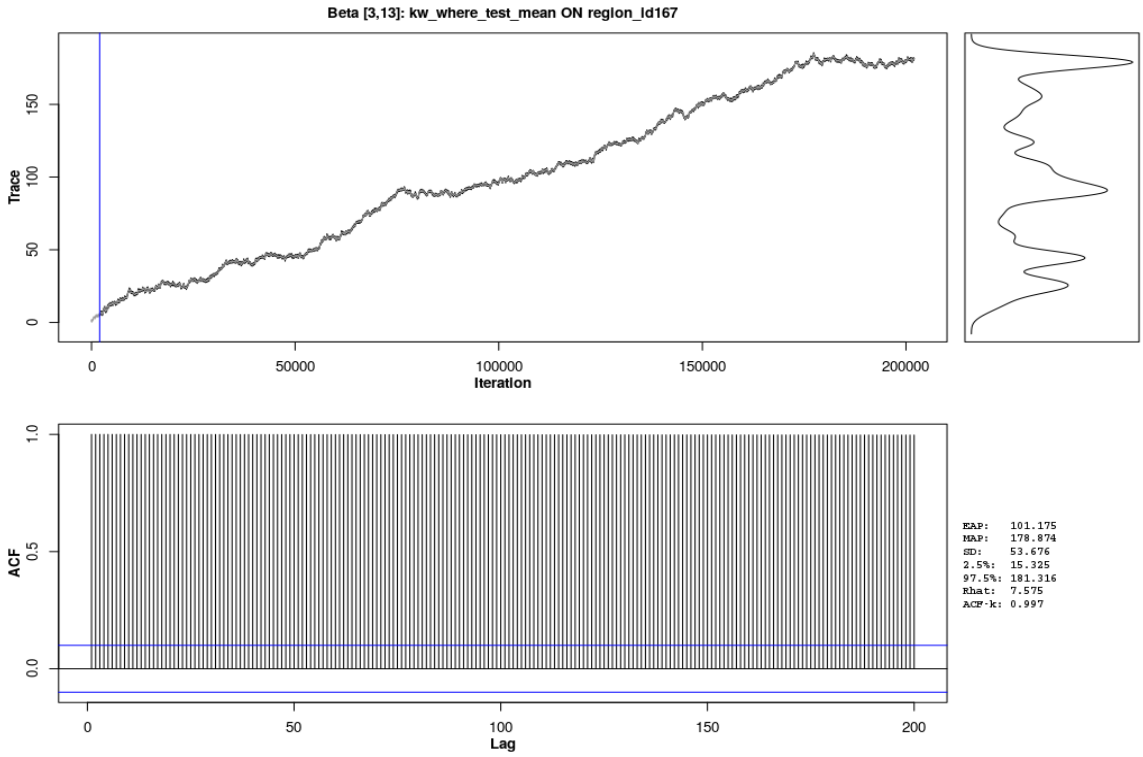


Figure 3-8 Trace and ACF plots of the parameters with the largest $\widehat{R}$ for pan.200 (top) and jomo.200 (bottom)

Table 3-9 Summary of $\widehat{R}$ and ACF of pan.5k

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

Table 3-9 and ***Figure 3-9*** show the summary of $\hat{R}$ and ACF and the trace and ACF plots for *pan.5k* respectively. We can see that although most parameters have converged, there are still a few that do not converge even with such long burn-ins and lag between imputations. Therefore, the usefulness of PAN and JOMO methods is limited due to the convergence issue when imputation model is complex.

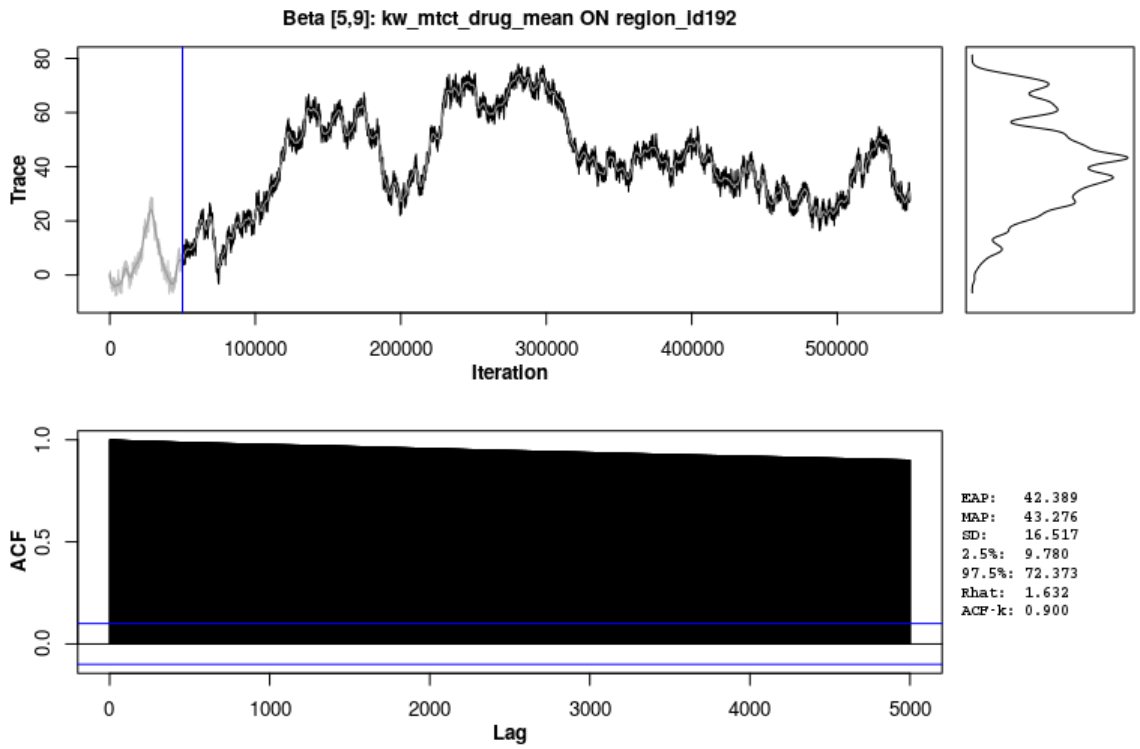


Figure 3-9 Trace and ACF plots of the parameter with the largest $\widehat{R}$ for pan.5k

Figure 3-10 shows the density plots of the observed and imputed values for each indicator for pan.200 method. Although the distributions of the imputed values approximate those of the observed values for many variables, the height of the density curve for a few variables, such as *heard_aids*, *kw_mtct_bf*, *mis_mosquito*, and *att_f_teacher*, is a bit off, suggesting suboptimal imputations for these indicators. More seriously, distribution of the imputed values for *kw_mtct_drug* has a very long tail to the right, suggesting that the imputation model does not converge and the imputations of this indicator are questionable.

Figure 3-11 shows the density plots of the observed and imputed values for each indicator for jomo.200 method. These plots show that jomo.200 has the same problems as the pan.200 method, suggesting that convergence of the full-Bayesian imputation methods (i.e., PAN and JOMO) can be crucial to the validity of imputations.

Figure 3-12 shows the density plots of observed and imputed values for *kw_mtct_drug* using pan.200 (left panel) and pan.5k (right panel), respectively. The distribution's tail of the imputed values shifts from the left to the right, further proving that imputations from full-Bayesian imputation methods can be unstable when the imputation model does not reach convergence.

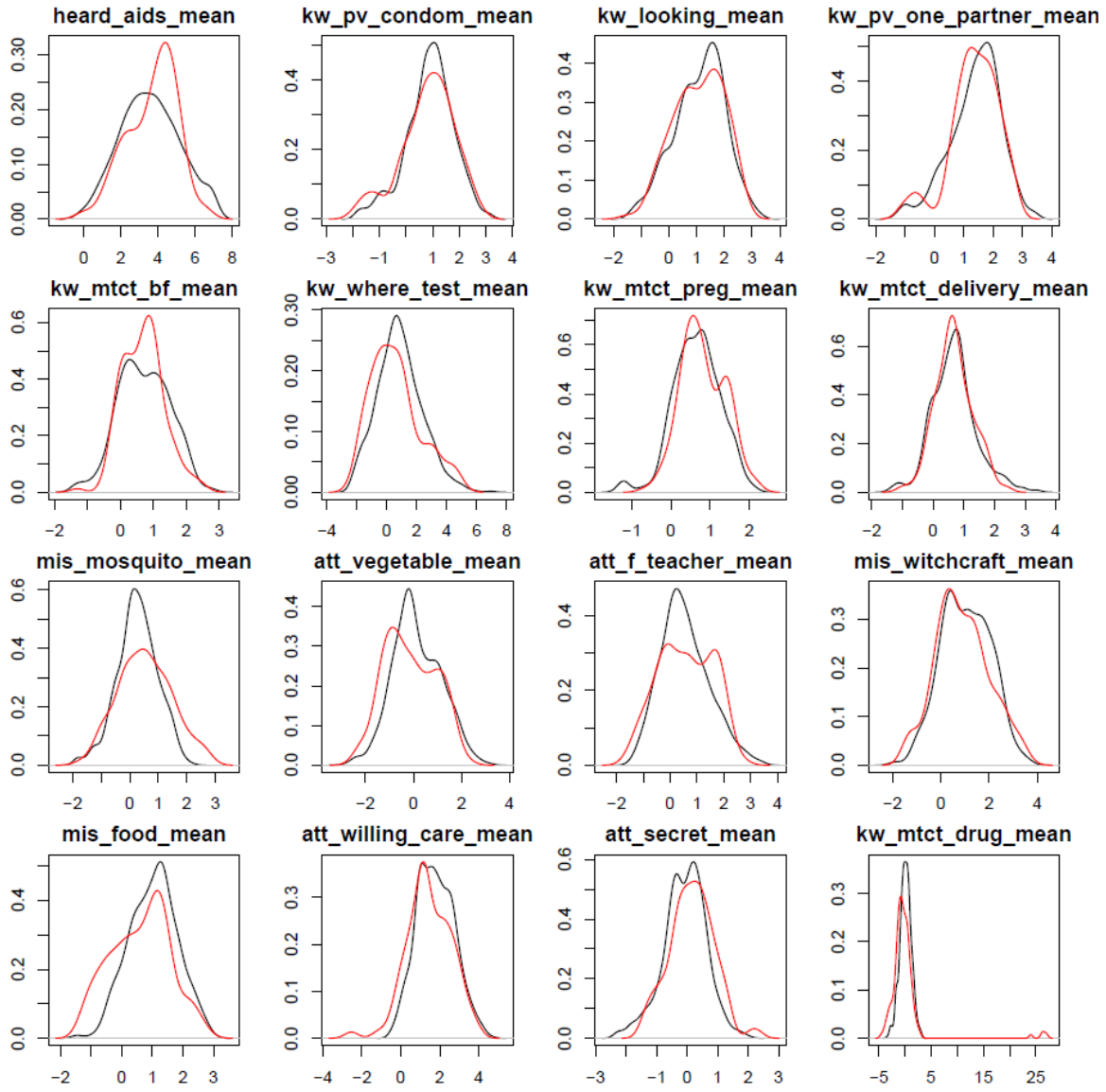


Figure 3-10 The density plots of observed (in black) and imputed (in red) values for each indicator using pan.200 method

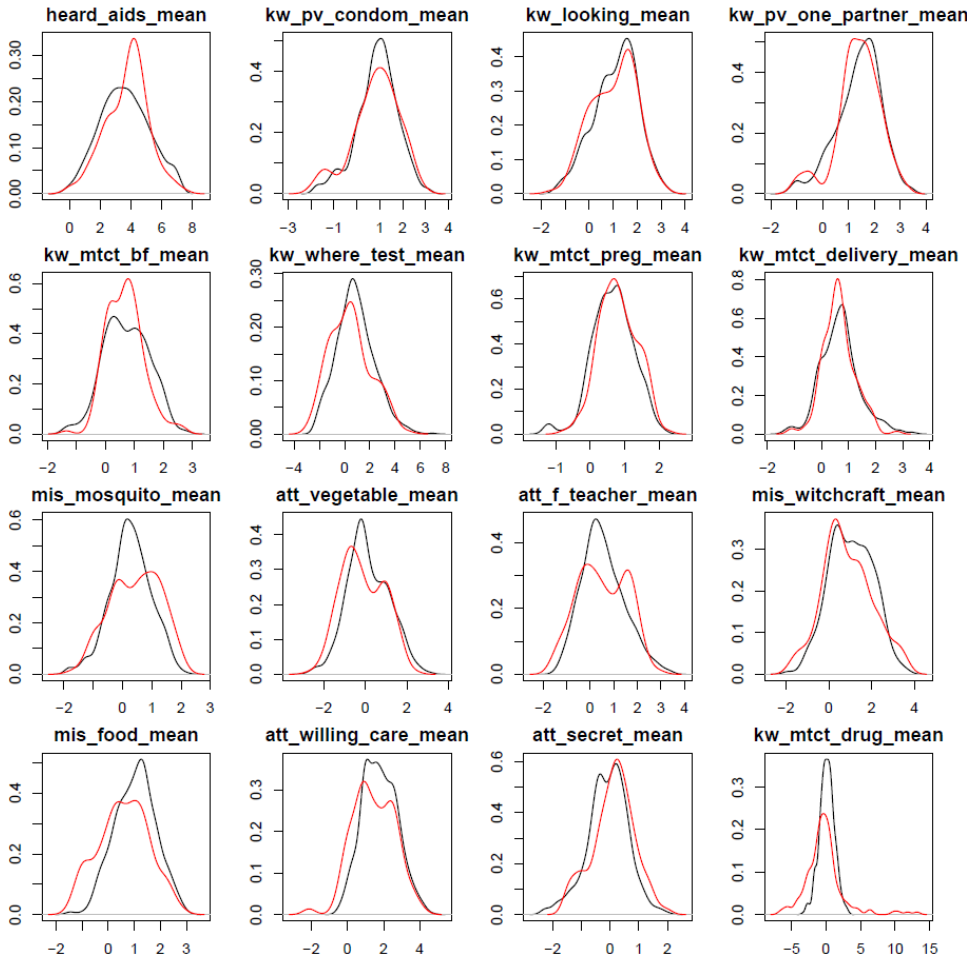


Figure 3-11 The density plots of observed (in black) and imputed (in red) values for each indicator using jomo.200 method

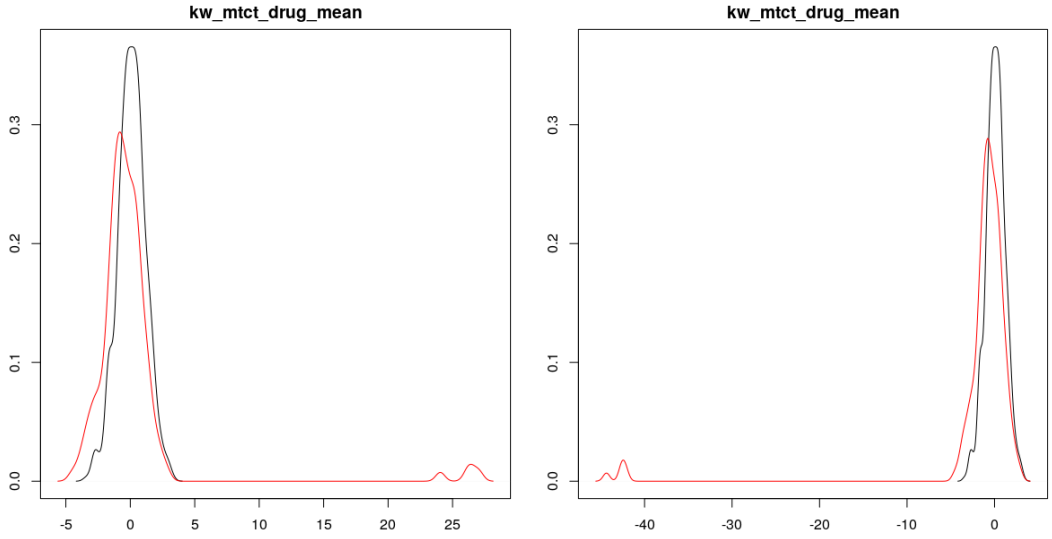


Figure 3-12 The density plots of observed (in black) and imputed (in red) values for kw_mtct_drug using pan.200 (left panel) and pan.5k (right panel)

4.3.3 diagnostics of MICE methods

Since *mice.2l.pan* is the only MICE method having good $RMSE$ and CR_{95}, we only provide diagnostics for *mice.2l.pan*. ***Figure 3-13*** shows the density plots of observed and imputed values for each indicator for *mice.2l.pan*. The plots are very similar to those of pan.200 except that the distribution of imputed values of *kw_mtct_drug* does not have long tail and is similar to the distribution of observed values. Compared with the PAN method, the mice.2l.pan method seems to perform better.

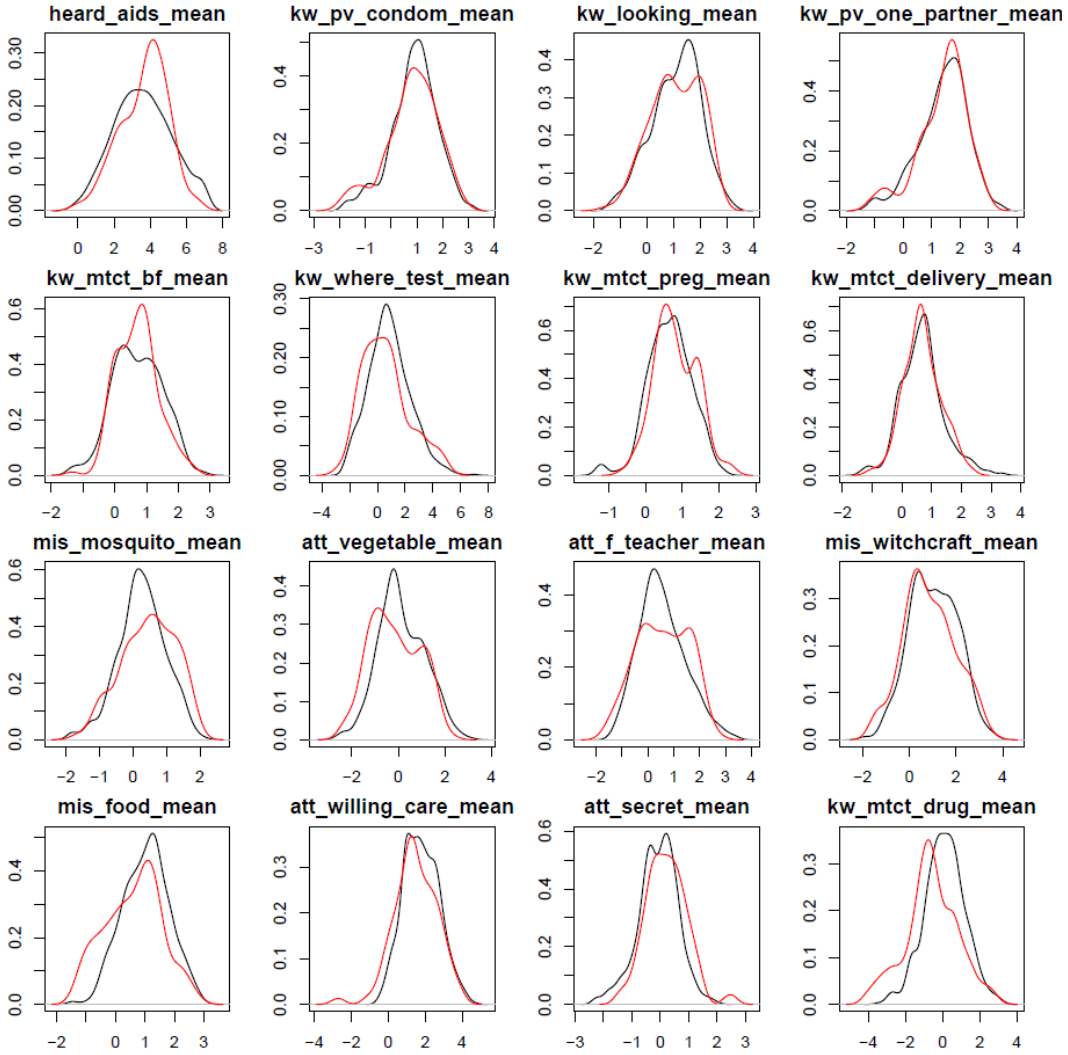
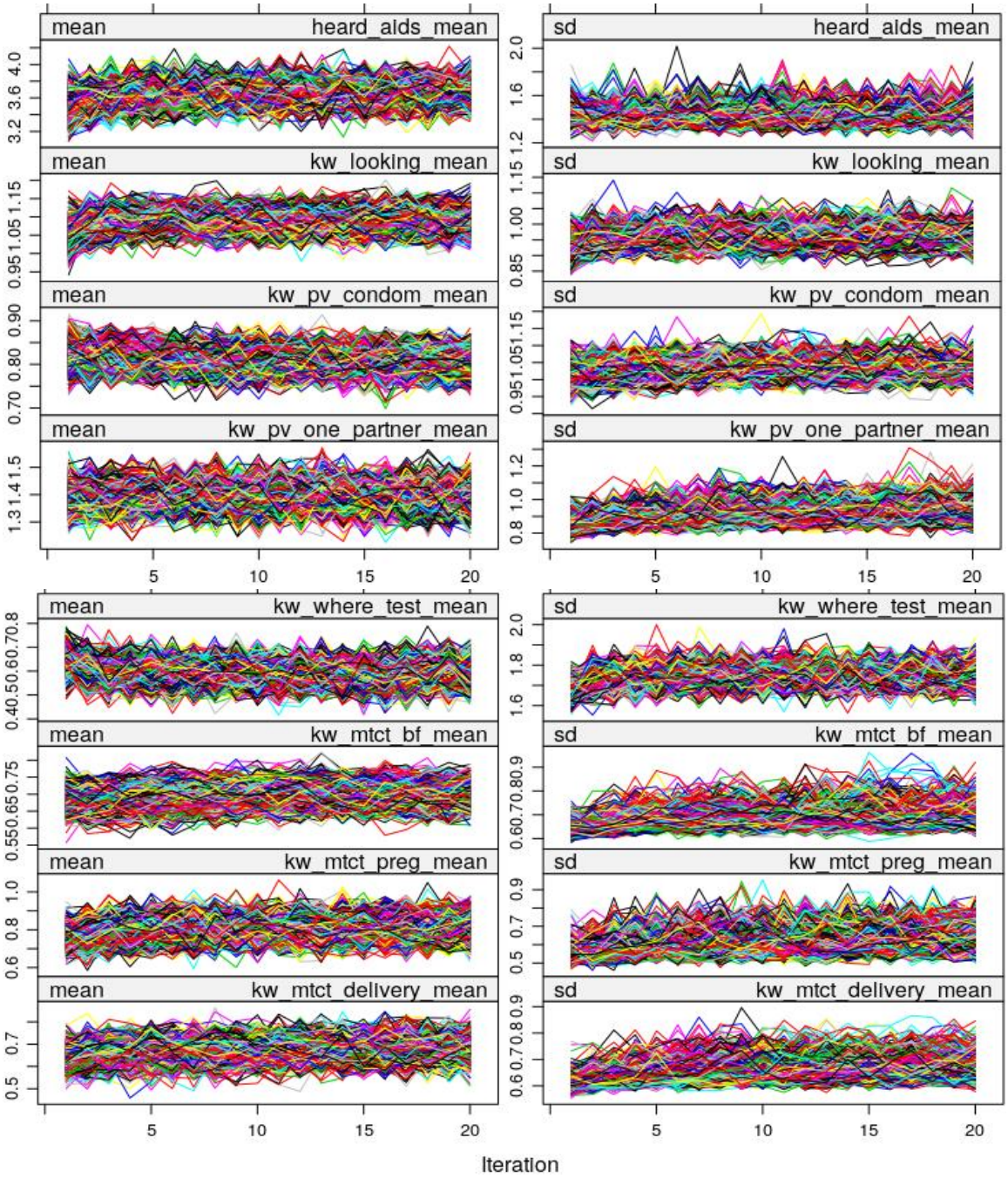


Figure 3-13 The density plots of observed (in black) and imputed (in red) values for each indicator using mice.2l.pan method

Since MICE implements an iterative MCMC algorithm, we need to examine the convergence of the model to make sure the imputed data are valid. ***Figure 3-14*** shows the trace plots of the mean and standard deviation of imputed values at each iteration for the incomplete variables. We can see that the traces of all variables are intermingle and free from any trend in the end, suggesting that the model has converged.[167]



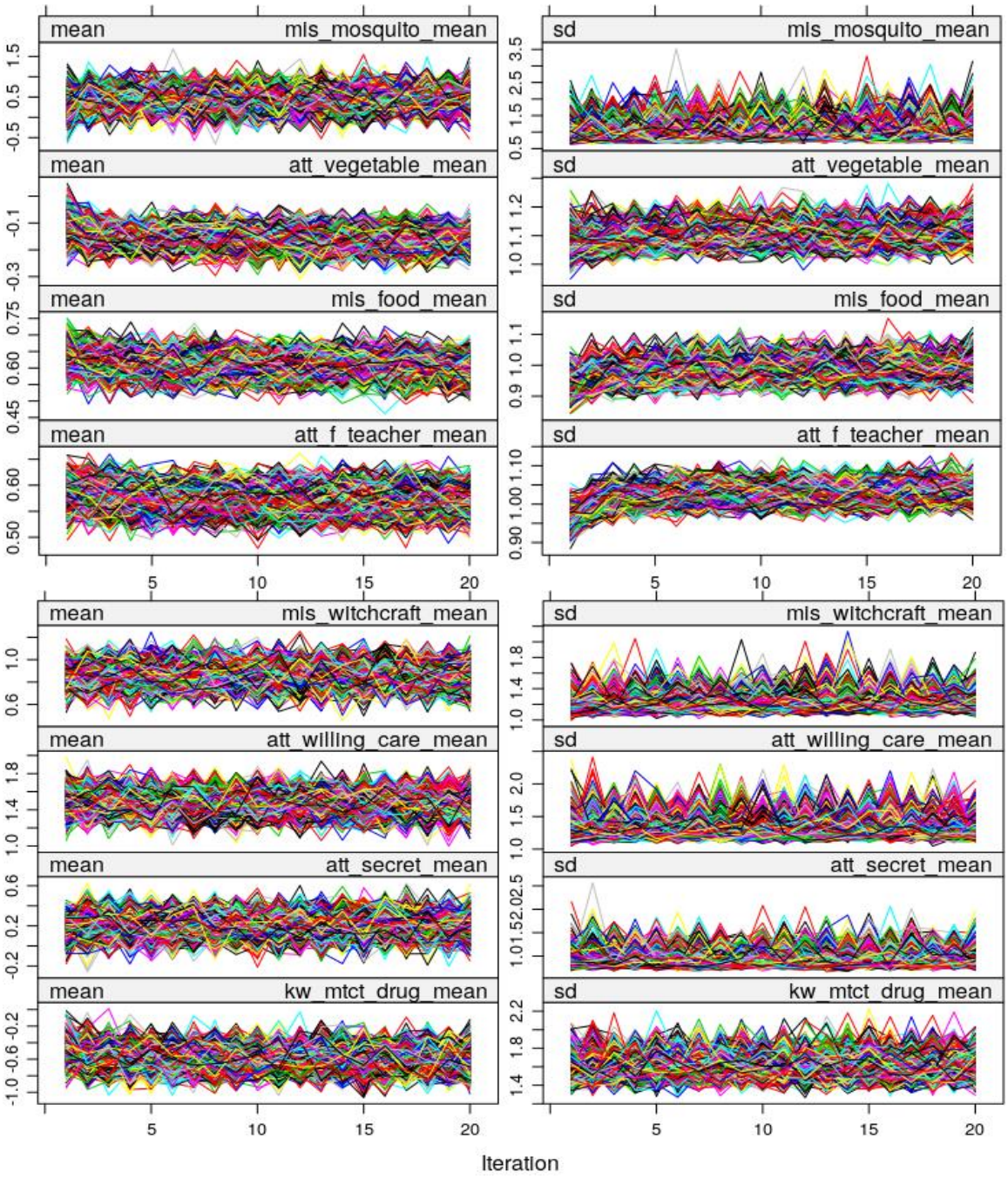


Figure 3-14 Trace plots of the mean and standard deviation of imputed values at each iteration for each indicator for mice.2l.pan method

4.4 The Impact of Including Cluster Means

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

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

Table 3-10 compares $RMSE$, CR_{95}, PS and PS^t of the three MI methods using the primary imputation model versus the imputation model including the cluster means. From **Table 3-10** we can see that including cluster means in the imputation model has little impact on the $RMSE$ or the CR_{95} but does significantly increase the running time.

Table 3-10 Comparison of $\boldsymbol{RMSE}$, $\boldsymbol{CR_{95}}$, $\boldsymbol{PS}$ and $\boldsymbol{PS^t}$ of the three MI methods using primary imputation model and imputation model including cluster means

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

4.5 The Impact of Including Incomplete Auxiliary Variables

To examine the impact of including incomplete auxiliary variables with different missing rate, we include other indicators of HIV/AIDS knowledge and attitudes with different missing rate by steps. Building upon the primary model, we include, in sequence, the indicators with missing rate less than 60%, 70%, 80%, 90% and lastly include all the indicators. For each model, we impute the missing data 1000 times using Amelia and PAN with the same configurations as before. **Table 3-11** summarizes the additional incomplete indicators of HIV/AIDS knowledge and attitudes and **Table 3-12** summarizes the $RMSE$, CR_{95} and PS scores of models including indicators with different missing rate.

Table 3-11 additional HIV/AIDS knowledge and attitudes indicators with various missing rates

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

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

From **Table 3-12** we can see that including more auxiliary variables, regardless of the missing rate of the variables, always decrease the $RMSE$.

Table 3-12 $\boldsymbol{RMSE}$, $\boldsymbol{CR_{95}}$ and $\boldsymbol{PS}$ scores of models including indicators with different missing rates

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

4.6 The Impact of Including Random Effects between Incomplete Variables

When imputing the missing proportions of key indicators using a 2-level imputation model, it is likely that the effects between the incomplete indicators are random across clusters. For example, people's knowledge on drug to prevent mother-to-child transmission (MTCT) of HIV may have different effects on their knowledge on ways of MTCT transmission across different countries. However, including the random effects between incomplete variables is not straightforward for MIJM methods.[32] PAN and Amelia cannot account for random effects between incomplete variables whereas JOMO has such flexibility by allowing the variance-covariance matrix of the level-1 error to randomly vary across the clusters to mimic the underlying random effect.[32,95] However, the computation of JOMO is very complex and it takes very long time to run. For instance, the jomo.200 method takes over 9 days to finish one-fold imputation. Moreover, the

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

Compared with MIJM methods, MICE methods account for random effects between incomplete variables more easily because MICE methods model each incomplete variable separately. To examine the impact of including random effects between incomplete variables for MICE, we build upon *mice.2l.pan* and add random effects between the incomplete variables on the basis of the primary imputation model, i.e., *mice.2l.pan.re*. **Table 3-13** and **Table 3-14** compare the overall and indicator-specific $RMSE$ and CR_{95} for pan.200, jomo.200, mice.2l.pan and mice.2l.pan.re, respectively.

Based on **Table 3-13** and **Table 3-14**, including random effects between incomplete variables significantly reduces $RMSE$ but increases CR_{95} especially for *mice.2l.pan.re*. In addition, the results of PAN and JOMO methods are very similar to the results of *mice.2l.pan* and *mice.2l.pan.re* respectively.

Table 3-13 $RMSE$, CR_{95} and **PS** scores of pan.200, jomo.200, mice.2l.pan and mice.2l.pan.re

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

Table 3-14 indicators specific **$RMSE$** and **CR_{95}** of pan.200, jomo.200, mice.2l.pan and mice.2l.pan.re

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

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

5. Discussion

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

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

evidence for Resche-Rigon and White's recent simulation study.[145] However, as Buuren[99] pointed out, the results may be dataset and/or model specific. For a different dataset or model, the cluster means may have huge effects on the missingness and thus on the imputations.

It is a consensus that including complete auxiliary variables improves the imputations.[161] However, incomplete auxiliary variables with high missing rate are usually not recommended to be included in the imputation model.[99] However, the results of our empirical study suggest that including more auxiliary variables is always beneficial regardless of the missing rate of the auxiliary variables. Of course, there is always a practical limit on how many auxiliary variables to include due to computational capacity and practical running time. This finding provides new insight on the inclusive strategy of auxiliary variables. However, since this is an empirical study, the finding can be dataset or model dependent. For instance, the auxiliary variables tested in this study are highly correlated with the targeted indicators. Therefore, future simulation evidence is still needed to fully understand the impact of different missing rates of auxiliary variables on the imputations.

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

Previous studies examining the performance of MI methods often use parameters of the analytical model as the evaluative targets.[92,161] However, our study, along with Mandel[156] and Ahmat Zainuri et al.'s study,[155] examines the performance of MI methods by assessing the accuracy of the imputed values. In other words, we examine the predictive rather than estimation accuracy of the imputation methods.[149] Another reason why we examine the predictive accuracy of imputations is that the imputed values are of interest and will be used to estimate the trends of the key indicators in a following study. To account for uncertainty of the imputations, we calculate the variance of the 1000 imputations for each missing value and will incorporate this variance in estimating trends of the key indicators. Therefore, although Rubin[151] and Buuren[99] think that the objective of MI is not to produce accurate imputations, we believe that the predictive accuracy of MI methods is justified and preferred in our study.

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

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

Similar to Castellacci et al.'s[26] and He et al.'s[102] studies, our study is an example of the "outside" application of MI, in which MI is used to produce multiply imputed datasets which can be used for many different analyses. A goal of this study is to make the multiply imputed datasets on HIV/AIDS knowledge and attitudes in the 47 SSA countries public available and researchers can use them to conduct different analyses. Therefore, the predictive accuracy of MI is highly preferred in this study.

Although carefully conducted, this study is not without limitations. First, according to Gelman et.al.[25] and Rendall et.al.,[168] missingness due to survey design is more likely to be MAR but the simulated missingness in this study are MCAR, which may affect the performance of MI. However, MI can also be used to handle MCAR and since all MI methods are evaluated using the same dataset, the performance of different MI methods is still comparable. Second, this study uses a real world dataset to evaluate the performance of MI methods. There may be uncontrolled and complex factors specific to this dataset that affect the performance of MI methods differentially. However, since all methods are evaluated using the same dataset and in the same way, we believe that our study provides useful empirical evidence on performance of different MI methods when imputing missingness in TSCS data. Lastly, since we do not extract all variables in the surveys, the multiply imputed datasets produced in this study can definitely be improved further by including more auxiliary variables from the surveys. However, based on the out-of-sample $RMSE$ and CR_{95}, we believe that our imputed datasets are good enough for a wide range of future analyses.

6. Conclusion

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

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

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

Chapter 4 : Calculating the Composite Score of Knowledge and Attitudes about HIV/AIDS and Estimating the Trends of the Composite Scores in 47 Sub-Saharan Africa Countries from 1998 to 2017

Abstract

Background

HIV/AIDS has been a leading cause of death in sub-Saharan Africa (SSA) for decades. People's knowledge and attitudes about HIV/AIDS are potentially important determinants of their behaviors, which in turn are major contributors to both the spread of HIV/AIDS as well as the diffusion of effective treatment and prevention interventions. In this study, we estimate trends of composite scores of knowledge and attitudes about HIV/AIDS in 47 sub-Saharan African countries from 1998 to 2017.

Methods

We systematically searched for nationally representative data on HIV/AIDS knowledge and attitudes for the 47 SSA countries in the Global Health Data Exchange (GHDx), a comprehensive health data catalog established by the Institute for Health Metrics and Evaluation (IHME) at the University of Washington. All survey data were extracted in a systematic and consistent way so that national estimates of 16 key HIV/AIDS knowledge and attitudes indicators are comparable across surveys. We then synthesized the point estimates of each key indicator from different surveys into trend estimates using spatial-temporal Gaussian process regression (ST-GPR), an innovative technique that gains strength over time and space to produce complete time series of an indicator as well as a 95% confidence interval. Lastly, we estimated the composite scores of knowledge and attitudes about HIV/AIDS and their 95% confidence intervals the estimates of 12 knowledge indicators and 4 attitudes indicator respectively.

Results

Among 3,002 surveys of the 47 SSA countries in GHDx from 1998 to 2017, we identified 220 surveys that have at least one of 16 key HIV/AIDS knowledge and attitudes indicators. Although the trends vary greatly from country to country, the composite scores of knowledge and attitudes about HIV/AIDS demonstrate an increasing pattern across the board. Overall, the composite score of knowledge about HIV/AIDS (*kw_score*) in SSA has increased from 0.516 to 0.762 with an annualized growth rate of 2.07% and the composite score of attitudes about HIV/AIDS (*att_score*) has increased from 0.509 to 0.647 with an annualized growth rate of 1.27%. Among 4 subregions of SSA, central SSA have the lowest level but highest growth rate of knowledge and attitudes about HIV/AIDS while southern SSA have the highest level but lowest growth rate. Men in SSA generally have better knowledge and attitudes about HIV than women do although the gap has been narrowing. Older people (25-49) in SSA tend to have better knowledge and attitudes about HIV/AIDS than younger people (15-24) and the gap has been widening.

Conclusion

Although there is great heterogeneity regarding levels and rates of change across countries, in sub-Saharan Africa, people's knowledge of HIV/AIDS and their attitudes toward people living with HIV/AIDS have, in general, improved over the past two decades. However, the inequalities in HIV/AIDS knowledge and attitudes by sex and age are concerning given the demographic expansion of younger populations and the large fraction of new HIV cases among adolescents and young adult, particularly young women ,in SSA.

1. Introduction

After four decades since its discovery in the 1980s, Human Immunodeficiency Virus (HIV), the cause of Acquired Immune Deficiency Syndrome (AIDS), continues to be a leading infectious disease in the world with 1.8 million people becoming newly infected in 2017.[1] Around the world, sub-Saharan Africa (SSA) shares a disproportionate burden of HIV/AIDS with 70% of people living with HIV (PLWH), 65% of new infections and more than 70% of AIDS-related deaths happening in SSA countries in 2017.[2] Most SSA countries have generalized HIV/AIDS epidemic, meaning that the HIV prevalence rate is greater than 1% countrywide.[3]

It has been well recognized that successful prevention and treatment of HIV/AIDS relies not only on the provision of condoms or antiretroviral therapy (ART), but, more importantly, on changes in individual behaviors,[5–12] which are in turn influenced by their knowledge and attitudes on HIV/AIDS and on its prevention and treatment.[13–22] Since the late 1990s, the rapid scale-up of ART in SSA countries has contributed to the historic decline of new HIV infections and HIV-related deaths in this region.[169,170] ART not only extends life expectancy but also reduces the transmissibility of PLWHs who adhere to the treatment. These effects of ART may affect people's knowledge and attitudes about HIV/AIDS, which in turn affects the scale-up of ART and PLWH's adherence to ART. Therefore, knowing the level and the change of people's knowledge and attitudes about HIV/AIDS is important to the success of HIV/AIDS interventions in this ART era.

Since the late 1980s, donor-funded national surveys such as Demographic Health Survey (DHS), Multiple Indicator Cluster Survey (MICS) and AIDS Indicators Survey (AIS) have collected data on people's knowledge and attitudes about HIV/AIDS. Besides these "Big three" surveys, some SSA countries also conducted their own national surveys, some of which collected HIV/AIDS

information. However, these data have not been compiled to understand the trends of people's knowledge and attitudes about HIV/AIDS in the SSA countries over time. In addition, there are two major barriers should one want to estimate the trends of HIV/AIDS knowledge and attitudes over time for all SSA countries. Firstly, the data are very scarce and unbalanced. The national surveys are usually conducted every few years and there are more surveys for some SSA countries (e.g. Senegal) than for others (e.g. Swaziland). Secondly, the indicators of HIV/AIDS knowledge and attitudes vary across surveys. Although there are some standard questions asking people's knowledge and attitudes about HIV/AIDS across surveys, not all surveys ask all the questions. Thus, there are more data for some indicators than for others.

To fill in the knowledge gap and to overcome the barriers in the data, we impute country-level proportion in the surveys using multiple imputation and estimate national trends of knowledge and attitudes about HIV/AIDS from 1998 to 2017 in 47 SSA countries using spatiotemporal Gaussian process regression (ST-GPR) to borrow information over time and space. In addition, to make the results more interpretable and more easily used, we calculate the composite scores of knowledge and attitudes about HIV/AIDS and estimate the trends of the composite scores from 1998 to 2017. By systematically searching for and utilizing all data available, this study provides crucial evidence on people's knowledge and attitudes on HIV/AIDS in SSA countries, which will make comparison between countries possible and facilitate future research on HIV/AIDS knowledge and attitudes in this region.

2. Methods

2.1 Countries of interest

In this study we continue to focus on the 47 SSA countries (Chapter two, **Table 1** and **Figure1**) as in Chapter two and three. We estimate the national trends of key indicators as well as the composite scores of people's knowledge and attitudes on HIV/AIDS in these 47 countries.

2.2 Data searching and data extraction

The data searching and extraction strategies are the same with those in Chapter two except that we select survey data from 1998 to 2017 instead of from 1980 to 2017. We choose a shorter period in this study because there are very few data available before 1998 and thus the estimated trends before 1998 heavily rely on model extrapolations which is sensitive to model selection. When searching the survey data, we include all national surveys having data on one of the 16 key indicators of knowledge and attitudes about HIV/AIDS. The 16 key indicators are the same with those in the previous chapters.

2.3 Country-level estimates of the key indicators

After extracting the individual level data, we obtain the country-level estimates of the key indicators by calculating the weighted mean of individual responses. The weights used are the sampling weights of the survey. The weighted mean of each indicator is calculated by gender and by age groups, including 15-49, 15-24 and 25-49. Since the individual data are all binary, the weighted means are proportions between 0 and 1. We use the *survey* package in R to obtain the weighted means and their corresponding standard errors while accounting for survey design elements such as stratification and clustering.

2.4 Country-level covariates

To estimate the trends of knowledge and attitudes about HIV/AIDS, important covariates are used throughout multiple imputation, bias adjustment and data synthesis process to improve the estimates. The country-level covariates used in this study include mean years of education per capita (*education*), GDP per capita based on 2010 international dollars (*GDP*), age-specific fertility rate (*ASFR*), modern contraception prevalence in women (*contra_prev*), healthcare access and quality index (*HAQI*)[37], proportion of population living in urban area (*prop_urban*), indicator of Muslim country where more than 50% population are Muslim (*muslim*), health system access, i.e., a composite score of immunization, hospital beds, in-facility delivery and skilled birth attendance (*HSA*), and proportion of pregnant women receiving 4 or more antenatal care from a skilled provider (*ANC4*). All covariates are estimated by IHME for GBD study 2017. Estimates of the covariates are extracted from IHME database for the period from 1998 to 2017.

2.5 Imputing the country-level missing proportions using multiple imputation

Among the 220 surveys, all of them have data for women but only 160 (72.7%) have data for men. Given the scarcity of data in many countries, the imbalance of data between men and women is concerning because the estimated trends for men and for women can be very different simply because the data are missing for men but not for women in certain year. Moreover, it is reasonable to assume that an indicator for women is informative of the same indicator for men and vice versa. Similarly, the data are imbalanced between age groups because some reports do not provide estimates for all age groups of interest.

In addition to data imbalance between gender and between age groups, the data are also imbalanced across indicators. Although there are 16 key indicators of knowledge and attitudes about HIV/AIDS, most surveys only have data on some of the 16 indicators because not all

relevant questions are asked in all the surveys. However, levels of the observed variables are likely to be informative of levels of the missing variables. Given the data scarcity, utilizing the observed variables to impute the missing variables can help improve the estimated trends.

In Chapter two, we addressed the data imbalance by using a linear mixed model to impute country-level missing proportions for men using the same observed proportions for women, and we did this "crosswalking" for each indicator separately. In this study, however, we addressed the data imbalance between gender, between age groups, and across indicators using multiple imputation (MI). Based on the results in Chapter three, we choose the MI method "*Amelia*"[143] to impute the country-level missing proportions of key indicators and we used R 3.5.0 with *amelia* package to implement the imputation.

2.6 Data bias adjustment

Since the county-level estimates may differ systematically between different types of surveys, we conduct data bias adjustment before synthesizing the data to produce trends. The approach is adapted from Wang et al.'s study.[38] Specifically, we model the logit-transformed country-level estimates using a linear mixed effect model, which includes a fixed effect for data source type across all locations and a random effect for data source type nested within each country (country-source random effect), controlling for important covariates such as year of data, mean years of education per capita and log-transformed GDP per capita (base 2010 international dollar).

$$logit(I_{cys}) = \beta_1 * education_{cy} + \beta_2 * GDP_{cy} + \gamma_c + \boldsymbol{\gamma_{cs}} + \boldsymbol{\alpha_s} + \varepsilon_{cys} \quad (1)$$

where, I is indicator, c is country, y is year, s is source type, γ_c is country random effect, γ_{cs} is country-source random effect, α_s is source type fixed effect across countries and ε_{cys} is the

residual term. There are three source types in total, i.e. DHS (including AIS), MICS and others, among which DHS is believed to be the least biased and thus is used as the reference source.

Based on equation (1), each data source has an associated random effect ($\boldsymbol{\gamma_{cs}}$) and a source type fixed effect ($\boldsymbol{\alpha_s}$). The values of these random and fixed effects for the reference source (DHS) are deemed to be the true deviation from the unbiased estimate. Therefore, we adjust the non-reference source types by replacing the estimated random and fixed effect values for these non-reference source types with the values for the reference type, as shown below.

$$adjusted\ logit(I_{cys}) = \beta_1 * education_{cy} + \beta_2 * GDP_{cy} + \gamma_c + \boldsymbol{\gamma_{c,ref}} + \boldsymbol{\alpha_{ref}} + \varepsilon_{cys} \quad (2)$$

Where $\boldsymbol{\gamma_{c,ref}}$ and $\boldsymbol{\alpha_{ref}}$ are the random and fixed effects estimated for DHS survey respectively. Therefore, national estimates from DHS surveys are unchanged but the estimates from other sources are adjusted accordingly. Using equation (2), we correct the bias in the data due to source type for male and female separately.

2.7 Data synthesis using ST-GPR

Once we obtain the fully imputed and bias-adjusted country-level estimates of the key indicators, we use spatiotemporal Gaussian process regression (ST-GPR) to synthesize the data and to estimate trends and the uncertainty intervals of the indicators in all 47 countries. The estimation is done by gender and by age group. Full details on the ST-GPR method have been published previously [38–41,44] and have been described in Chapter two as well. The ST-GPR method in this chapter is the same with that in Chapter two except that we model *year* linearly in the first-stage prediction instead of modeling *year* using natural splines. We change the modeling of year because there are abundant data for most indicators from 1998 to 2017 and the observed data of the indicators clearly demonstrate a linear trend over time.

Specifically, in the first-stage of ST-GPR, we fit the fully imputed and bias-adjusted data to a linear mixed model with random slopes and random intercepts and then predict the time series of the indicator with the fitted model. The model is as follow:

$$\begin{aligned} logit(I_{cy}) = {} & (\beta_0 + \gamma_{0,c}) + (\beta_1 + \gamma_{1,c})year + \beta_2 education_{cy} + \beta_3 log(GDP_{cy}) \\ & + \beta_4 logit(contra\_prev_{cy}) + \beta_5 ASFR_{cy} + \beta_6 HAQI_{cy} + \beta_7 Muslim_c \\ & + \beta_8 logit(prop\_urban_{cy}) + \beta_9 HSA_{cy} + \beta_{10} logit(ANC4_{cy}) + \varepsilon_{cy} \quad (3) \end{aligned}$$

where c is country, y is year, I_{cy} is indicator of country c in year y, $\gamma_{0,c}$ is country random intercept, $\gamma_{1,c}$ is country random slop of $year$. The second and the third stage of ST-GPR is the same with those in Chapter two.

2.8 Identification and removal of outliers

To ensure the quality of the estimates, we carefully review the raw data and the estimates to identify and remove outliers. Firstly, we remove data sources with quality concern such as Nigeria DHS 1999 and Malawi Global Fund Household Health Coverage Survey 2007-2008. Secondly, we remove certain indicator(s) from a survey due to quality concerns such as having too many missing values for no reason or applying wrong skip logic for some indicators. Lastly, we utilize statistical methods and expert opinion to identify extreme or unusual county-level estimates. Specifically, once we have obtained the residuals between the first-stage predictions and the bias-adjusted country-level estimates, we calculate the MAD of residuals within each country and identify data points with residual three MAD away from the median.[62] We then consult experts to determine whether these extreme data points are actual outliers.

In short, outliers are removed with great caution and they were carefully documented for reference. The detailed descriptions of problematic surveys, indicators, and outliers are included as appendix.

2.9 Calculation of the composite scores of knowledge and attitudes about HIV/AIDS

In addition to the 16 key indicators, we also calculate the composite scores of knowledge and attitudes about HIV/AIDS in this chapter. In the literature, we find that the composite scores of knowledge and attitudes about HIV/AIDS for individual are often the sum of correct answers to a bunch of binary questions on knowledge and attitudes about HIV/AIDS.[171–174] Therefore, when calculating the country-level composite scores of knowledge and attitudes about HIV/AIDS, we take the mean of the country-level estimates of 12 knowledge indicators and 4 attitudes indicators respectively. Specifically, after obtaining 1,000 draws for each indicator in each country and in each year from 1998 to 2017, we calculate the mean of the indicators on draw level for 1,000 times. Therefore, we end up with 1,000 calculated composite scores of knowledge and of attitudes for each country in each year. We then use mean of the 1,000 composite scores as the final estimate and use the 2.5^{th} and 97.5^{th} quantiles as the uncertainty interval of final estimate of the composite scores.

Besides the composite score of people's overall knowledge about HIV/AIDS, we also calculate composite scores of knowledge about prevention (*kw_pv_one_partner, kw_pv_condom, kw_looking, kw_where_test*), MTCT (*kw_mtct_preg, kw_mtct_delivery, kw_mtct_bf, kw_mtct_drug*) and misconceptions (*mis_mosquito, mis_food, mis_witchcraft*) of HIV/AIDS.

2.10 Data visualization

After identifying the country-sex-age-group specific time series of composite scores of knowledge and attitudes, we synthesize the time series of each composite score to obtain the mean trends of the indicators by sex, by region, and by age group using the GAM smooth function with the formula

$$\bar{I}_{cys} = cs(year) + \varepsilon_{cys}$$

where c, y, s, ε are country, year, sex, and the error term. The $cs(year)$ is the shrinkage version of penalized cubic splines of year with knots spread evenly through the covariate values. [63–65]

In addition to mean trends of the composite scores, we also map the scores over time to show changes of the scores over time and heterogeneity of the scores across SSA countries.

3. Results

By April 1st, 2019, there were a total of 3,002 surveys in the 47 SSA countries from 1998 to 2017 in GHDx, among which 220 are accepted for having at least one of the 16 key indicators of knowledge and attitudes about HIV/AIDS. Among the 220 surveys used in the study, there are 116 DHS, 72 MICS, and 32 other surveys, accounting for 52.7%, 32.7%, and 14.5% of the surveys respectively. The DHS and the MICS are the majority sources, jointly accounting for more than 85% of all surveys.

After we populate the original dataset to make sure that each survey has the same number of rows with each row representing a specific age-sex group, the overall missing rate of the 16 key indicators of HIV/AIDS knowledge and attitudes in the populated dataset is 31.8%. Among the 16 key indicators, *heard_aids* and *kw_mtct_drug* have the smallest and the largest missing rate of 20.0% and 55.3% respectively. The overall out-of-sample RMSE between the imputed and the observed data using the *Amelia* method is 0.028, suggesting that the imputed data are close to the truly observed data on average. The trend plots of the imputed and the observed data also show that the imputed data are reasonable. ***Figure 4-1*** shows some examples of these trend plots. The imputed data are those with the uncertainty interval due to imputation. Men's data are blue and women's data are red. The solid lines in the plots are the Lowess curve of the data by sex.

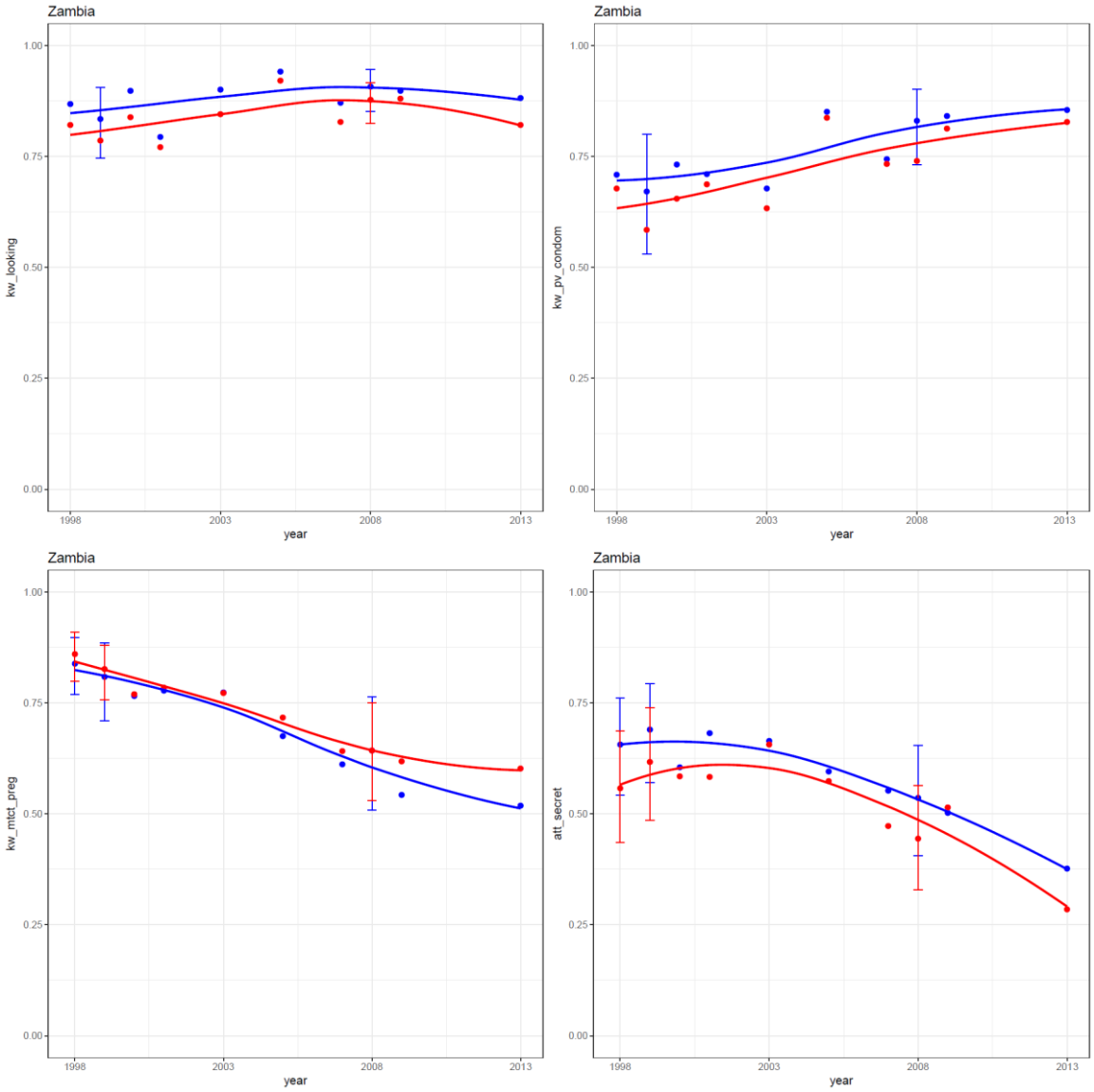


Figure 4-1 The trend plots of the imputed and the observed indicators for Zambia

In sub-Saharan Africa, from 1998 to 2017, the composite score of knowledge about HIV/AIDS (*kw_score*) has increased from 0.516 to 0.762 with an annualized growth rate of 2.07%. The composite score of attitudes about HIV/AIDS (*att_score*) has increased from 0.509 to 0.647 with an annualized growth rate of 1.27%. Changes of the composite scores of knowledge and attitudes about HIV/AIDS over time by region, by sex and by age group are summarized in **Table 4-1**.

Table 4-1 Changes of composite scores of knowledge and attitudes about HIV/AIDS

	kw_score			att_score		
	1998	**2017**	**annualized rate (%)**	**1998**	**2017**	**annualized rate (%)**
Overall	0.516	0.762	2.07	0.509	0.647	1.27
By region						
Central	0.496	0.758	2.26	0.525	0.663	1.23
Eastern/Northern	0.524	0.775	2.09	0.533	0.673	1.23
Southern	0.671	0.847	1.23	0.646	0.789	1.06
Western	0.472	0.725	2.28	0.445	0.579	1.39
By sex						
Male	0.554	0.768	1.73	0.546	0.664	1.04
Female	0.481	0.756	2.41	0.476	0.632	1.51
By age groups						
15-24	0.504	0.741	2.05	0.500	0.631	1.23
25-49	0.523	0.775	2.09	0.516	0.659	1.29

Table 4-2 Changes of composite scores of knowledge about prevention, MTCT and misconceptions about HIV/AIDS

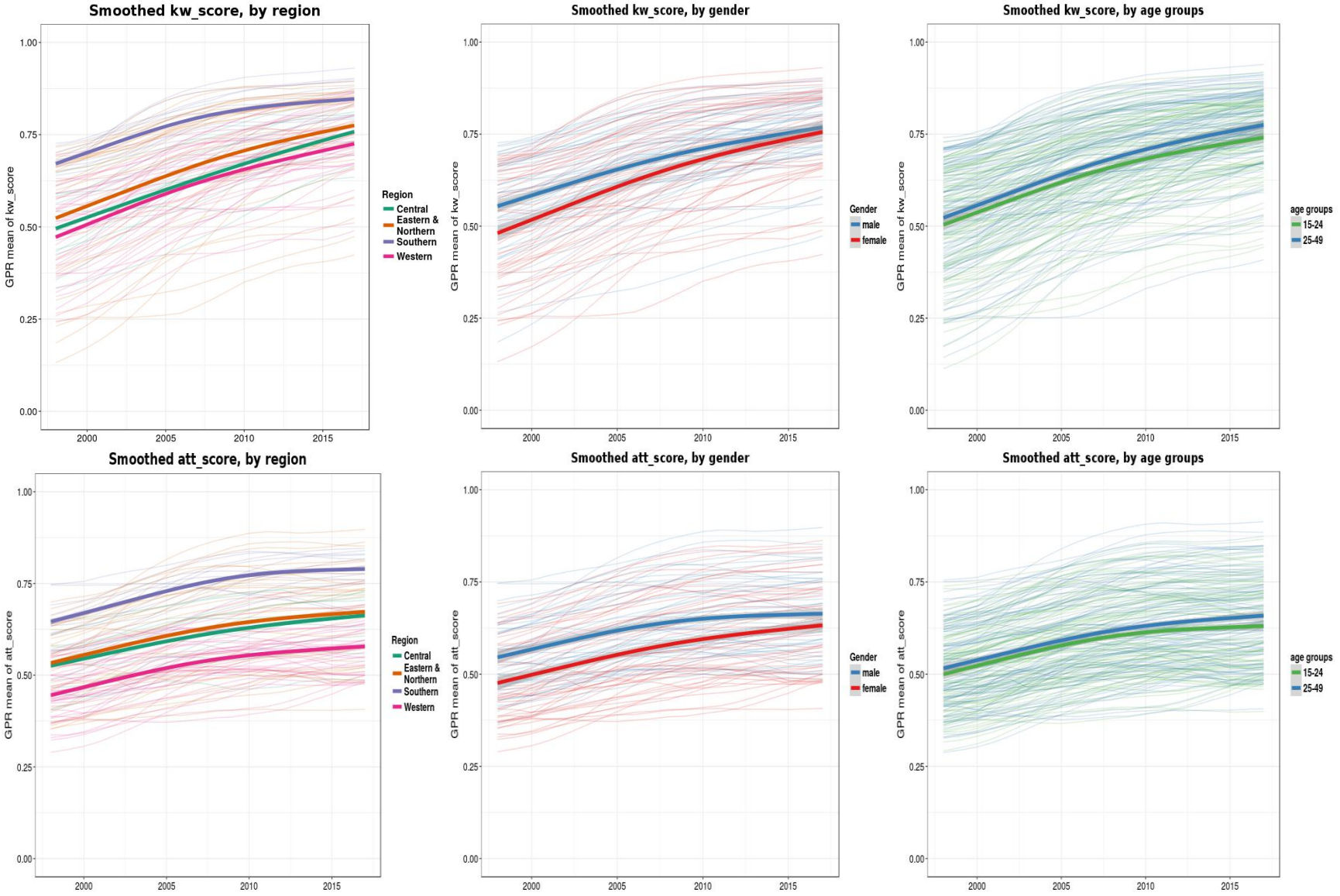
	kw_pv_score			mtct_score			mis_score		
	1998	2017	annualized rate (%)	1998	2017	annualized rate (%)	1998	2017	annualized rate (%)
Overall	0.534	0.800	2.15	0.438	0.691	2.42	0.466	0.739	2.46
By region									
Central	0.528	0.818	2.33	0.426	0.705	2.69	0.407	0.681	2.74
Eastern/ Northern	0.538	0.795	2.07	0.430	0.703	2.62	0.509	0.785	2.31
Southern	0.719	0.892	1.14	0.596	0.780	1.42	0.613	0.833	1.63
Western	0.483	0.773	2.50	0.404	0.648	2.52	0.410	0.689	2.77
By sex									
Male	0.593	0.823	1.74	0.451	0.669	2.10	0.515	0.762	2.08
Female	0.480	0.779	2.58	0.429	0.714	2.71	0.421	0.717	2.84
By age groups									
15-24	0.517	0.772	2.13	0.414	0.660	2.48	0.475	0.738	2.34
25-49	0.547	0.819	2.15	0.456	0.712	2.37	0.451	0.738	2.62

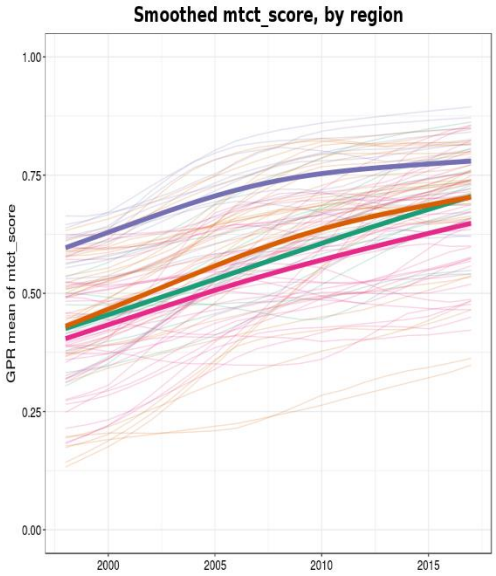
People's knowledge about HIV/AIDS prevention (*kw_pv_score*), mother-to-child transmission of HIV/AIDS (*mtct_score*), and misconceptions of HIV/AIDS (mis_score) have also improved significantly over the past two decades. In SSA, from 1998 to 2017, kw*_pv_score*, *mtct_score*,

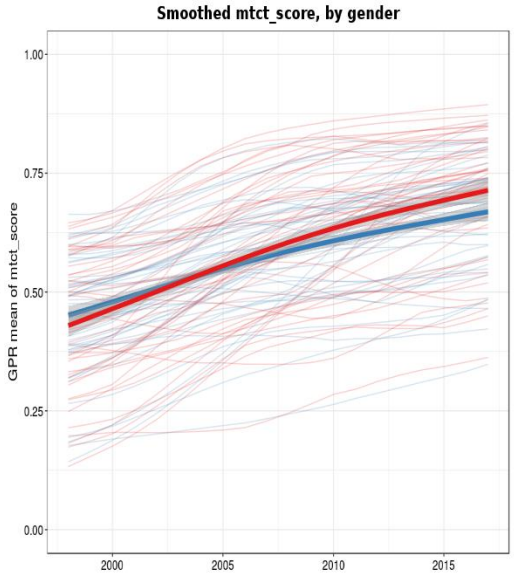
and *mis_score* have increased from 0.534, 0.438, and 0.466 to 0.800, 0.691, and 0.739 respectively, with annualized growth rates of 2.15%, 2.42% and 2.46%. Changes of these composite scores by region, by sex, and by age group are summarized in **Table 4-2**.

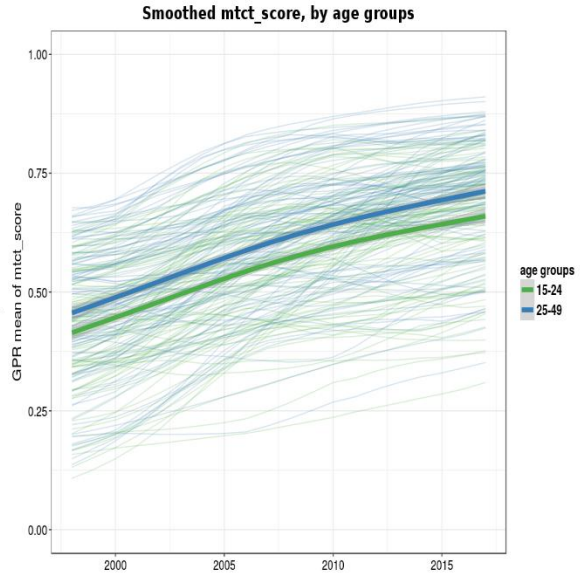
Figure 4-2 visualizes the mean trends of the composite scores by region, sex, and age group. It also shows the country-sex-age specific trends. The solid thick lines in the plots are the mean trends and the transparent thin lines are country-sex-age specific trends.

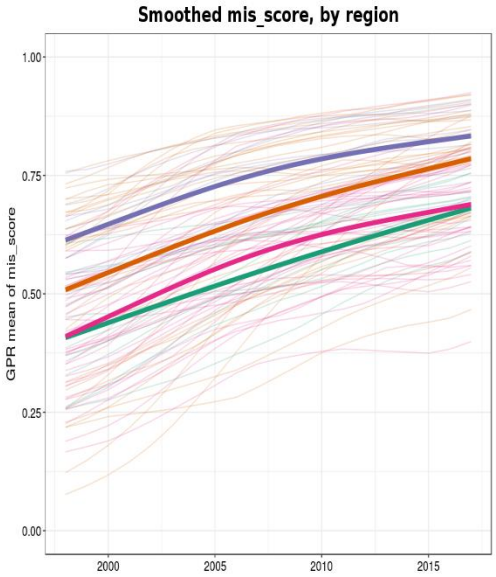
Figure 4-3 shows the maps of the composite scores in 1998, 2007, and 2017 by gender. It helps visualize not only changes of the composite scores over time but also heterogeneity of the composite scores between sex and across countries.

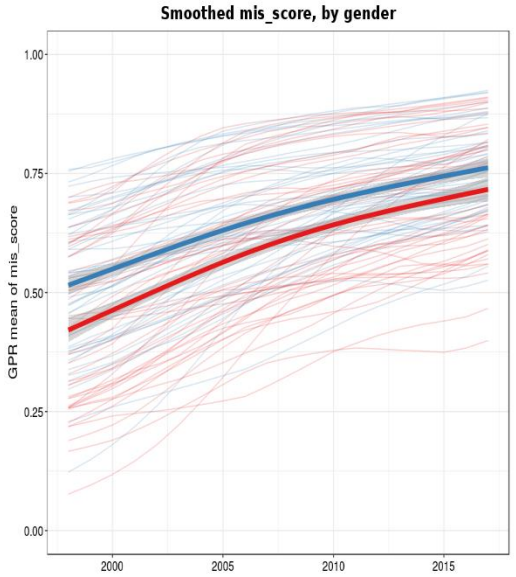

Smoothed kw_score, by region
GPR mean of kw_score
1.00
0.75
0.50
0.25
0.00
2000
2005
2010
2015
Region
Central
Eastern & Northern
Southern
Western
Smoothed kw_score, by gender
Gender
male
female
Smoothed kw_score, by age groups
age groups
15-24
25-49
Smoothed att_score, by region
GPR mean of att_score
Smoothed att_score, by gender
Smoothed att_score, by age groups

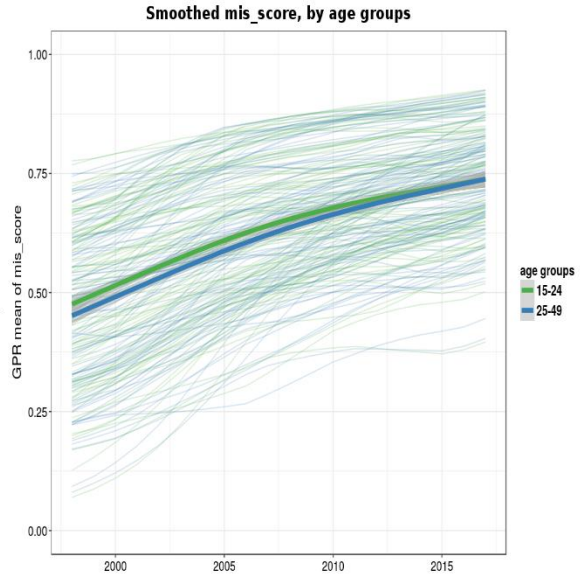

Smoothed mtct_score, by region
GPR mean of mtct_score
1.00
0.75
0.50
0.25
0.00
2000
2005
2010
2015
Region
Central
Eastern & Northern
Southern
Western


Smoothed mtct_score, by gender
GPR mean of mtct_score
1.00
0.75
0.50
0.25
0.00
2000
2005
2010
2015
Gender
male
female


Smoothed mtct_score, by age groups
GPR mean of mtct_score
1.00
0.75
0.50
0.25
0.00
2000
2005
2010
2015
age groups
15-24
25-49


Smoothed mis_score, by region
GPR mean of mis_score
1.00
0.75
0.50
0.25
0.00
2000
2005
2010
2015
Region
Central
Eastern & Northern
Southern
Western


Smoothed mis_score, by gender
GPR mean of mis_score
1.00
0.75
0.50
0.25
0.00
2000
2005
2010
2015
Gender
male
female


Smoothed mis_score, by age groups
GPR mean of mis_score
1.00
0.75
0.50
0.25
0.00
2000
2005
2010
2015
age groups
15-24
25-49

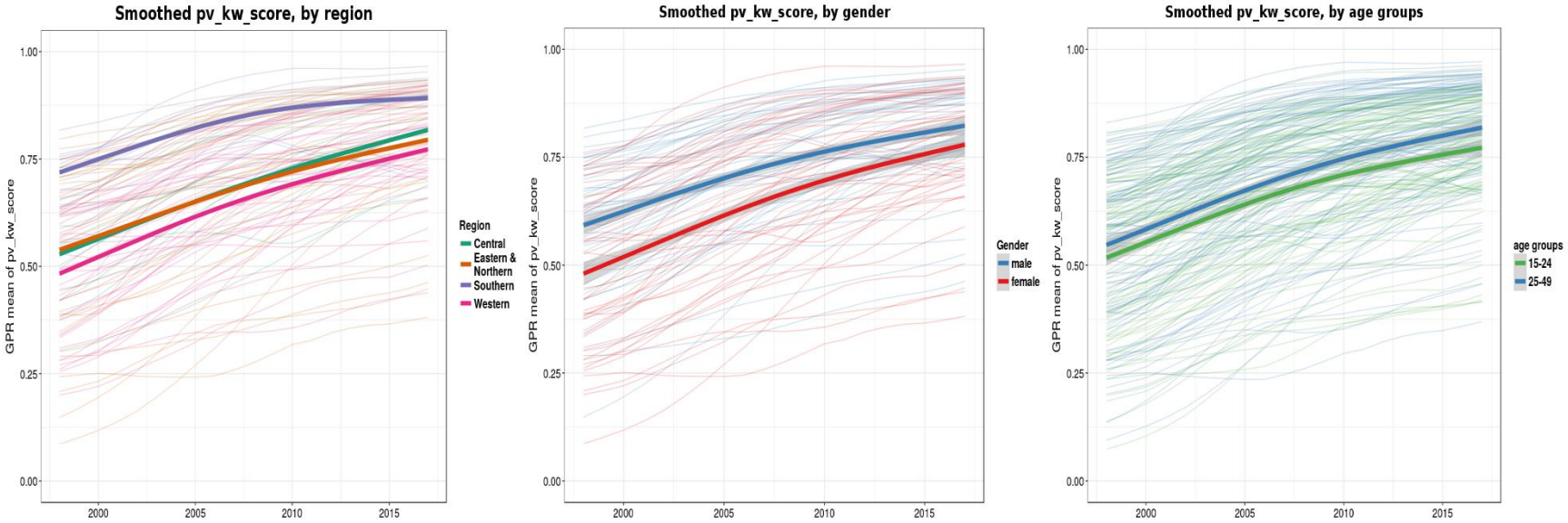
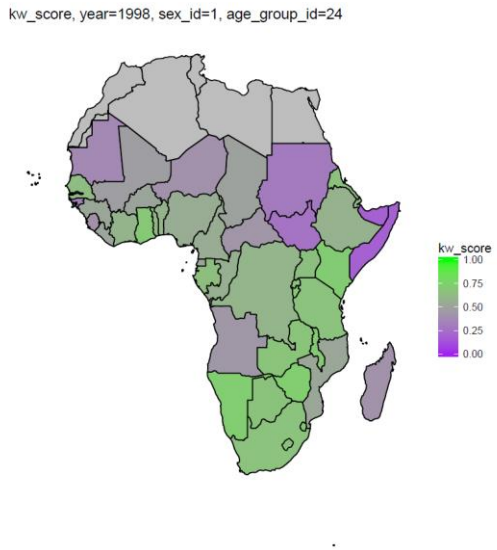
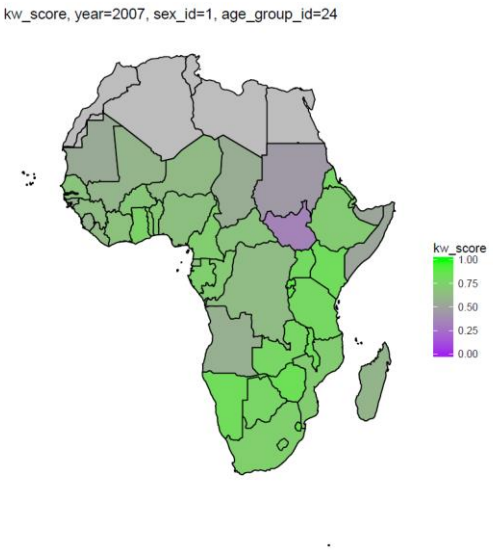
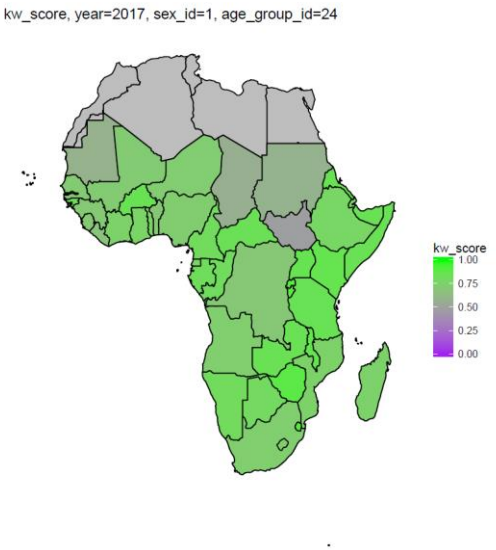


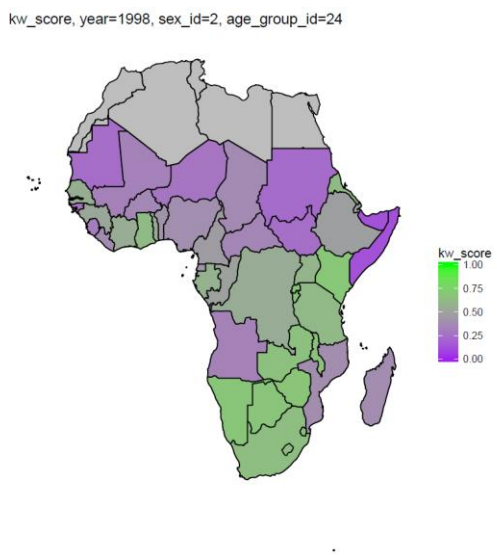
Figure 4-2 Mean trends of the composite scores by gender, subregion, and age group

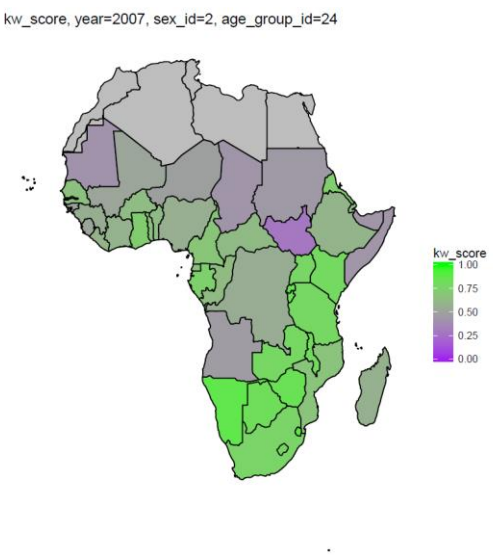
Map of the composite score of overall knowledge about HIV/AIDS (*kw_score*)

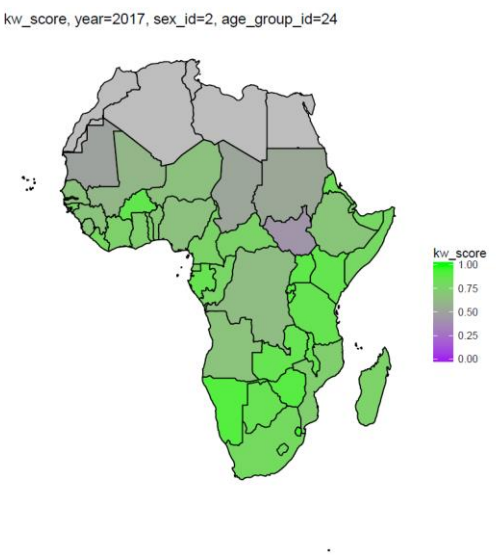




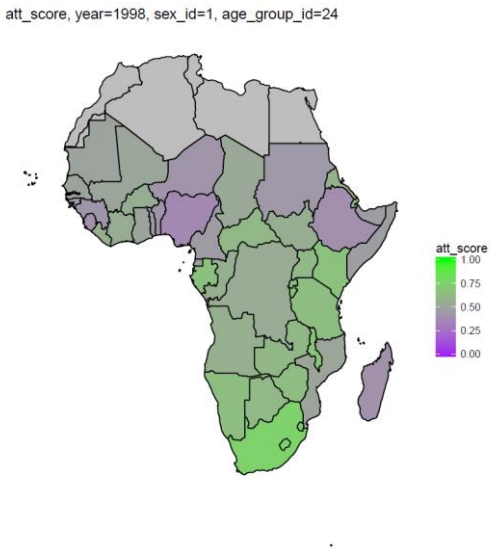


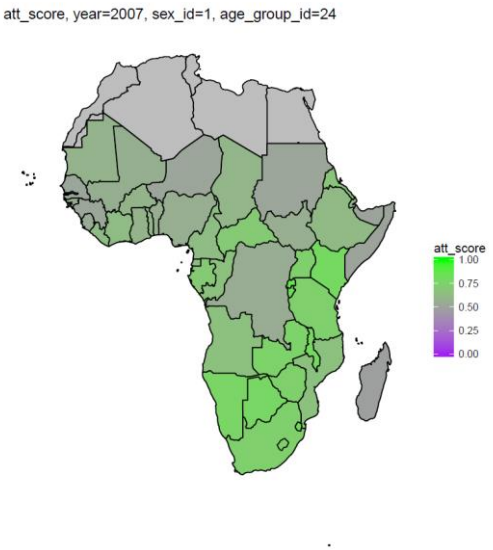


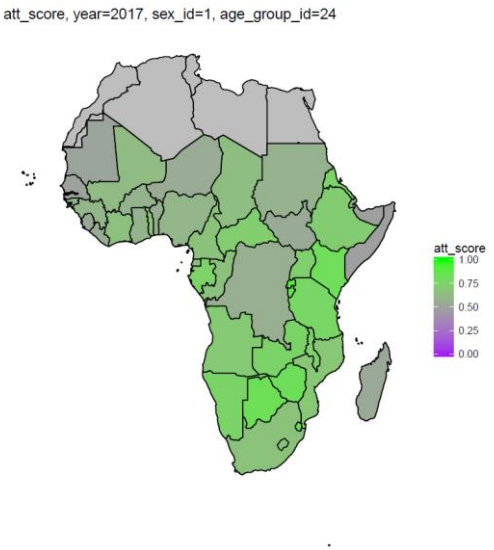


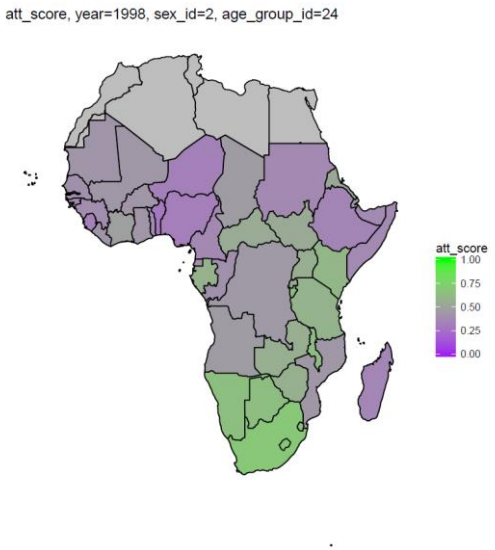


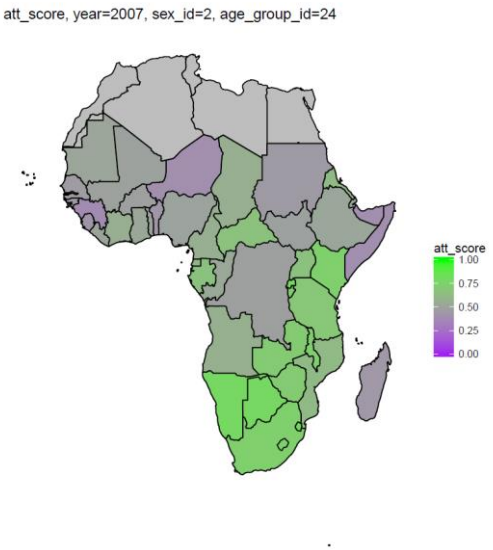
Map of the composite score of overall attitudes about HIV/AIDS (*att_score*)

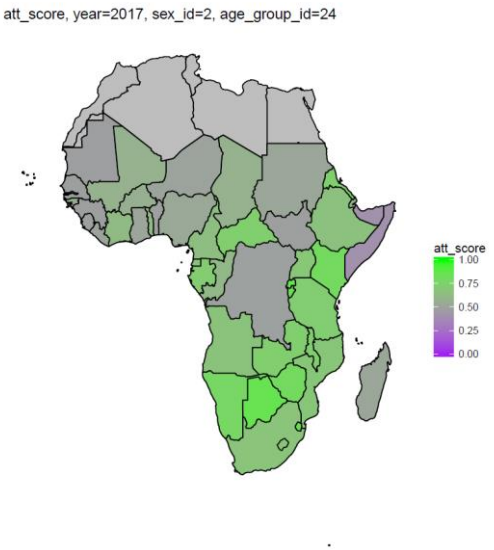




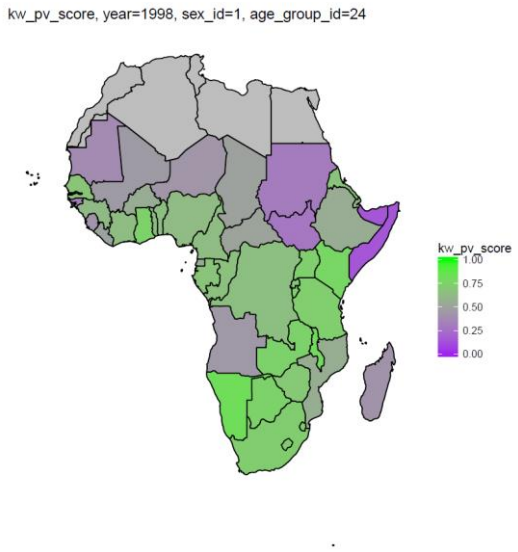


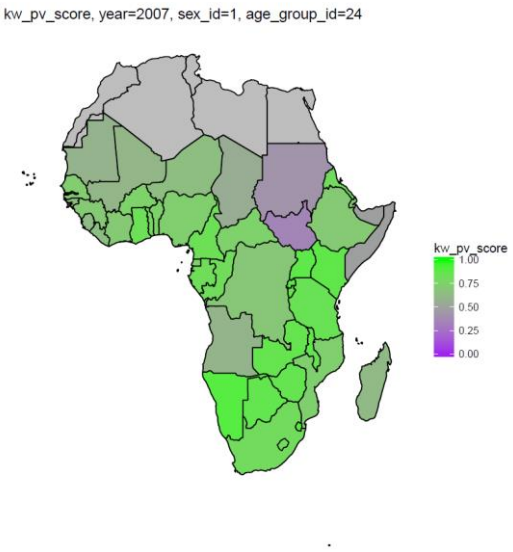


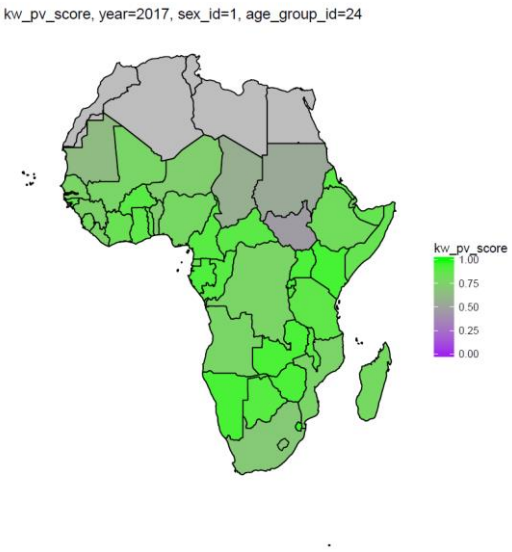


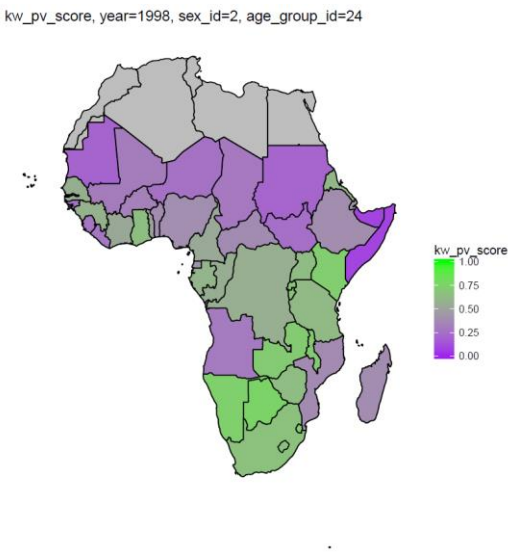


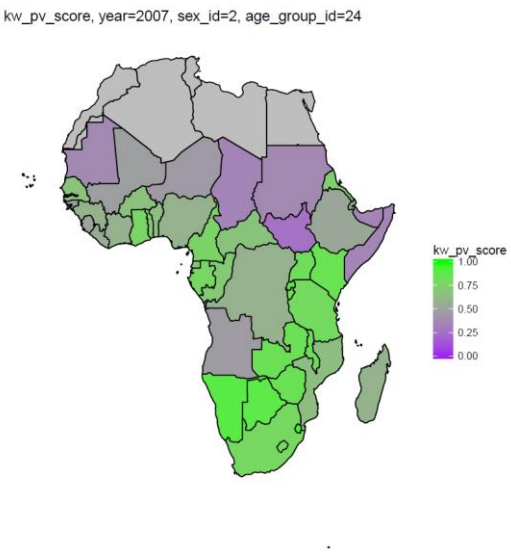
Map of the composite score of prevention knowledge about HIV/AIDS (*kw_pv_score*)

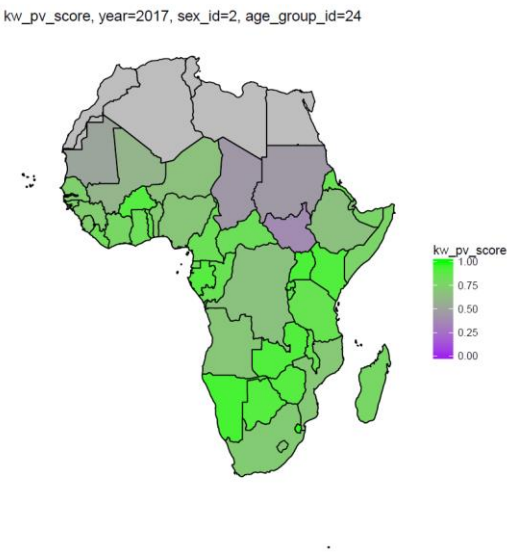




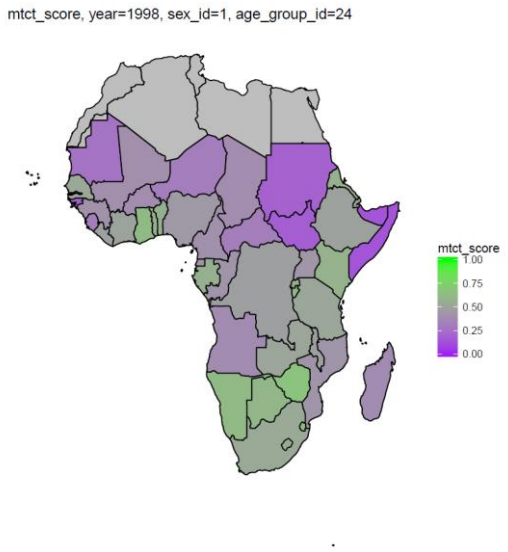


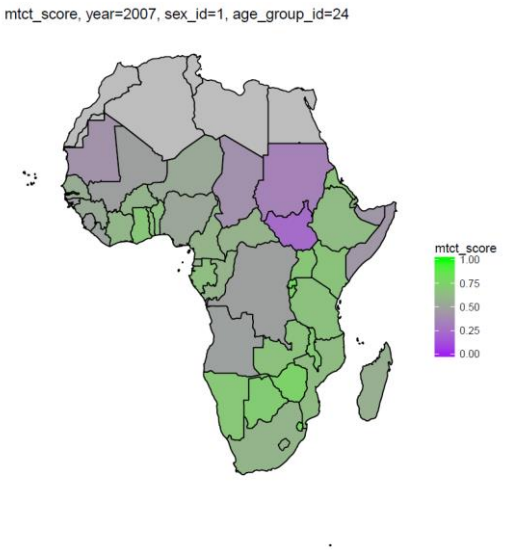


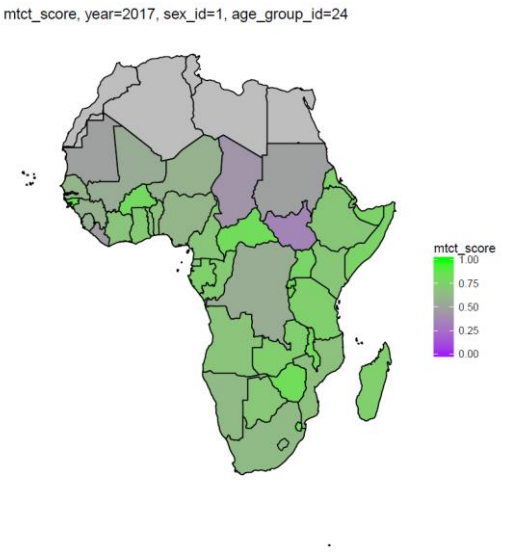


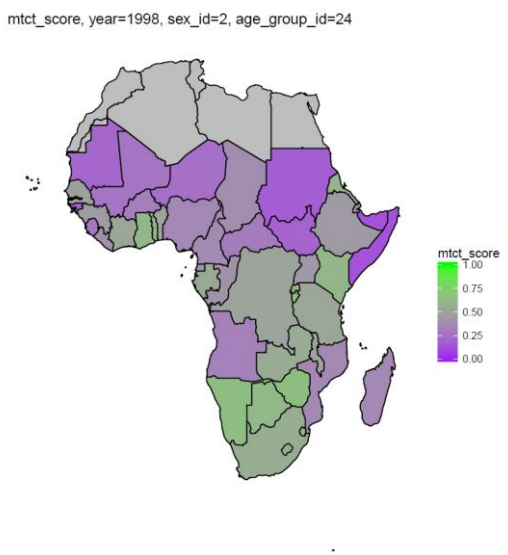


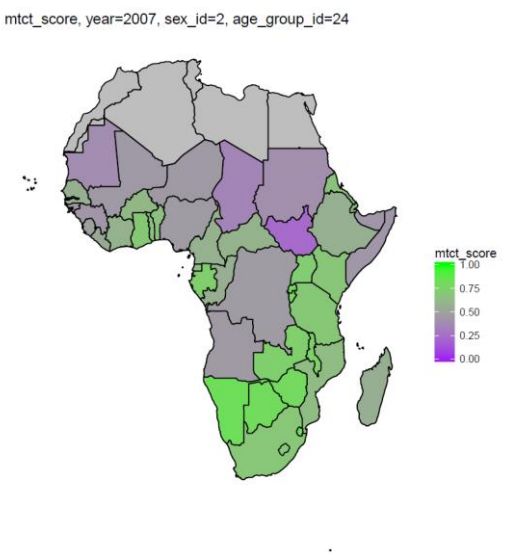
Map of the composite score of knowledge about MTCT of HIV/AIDS (*mtct_score*)

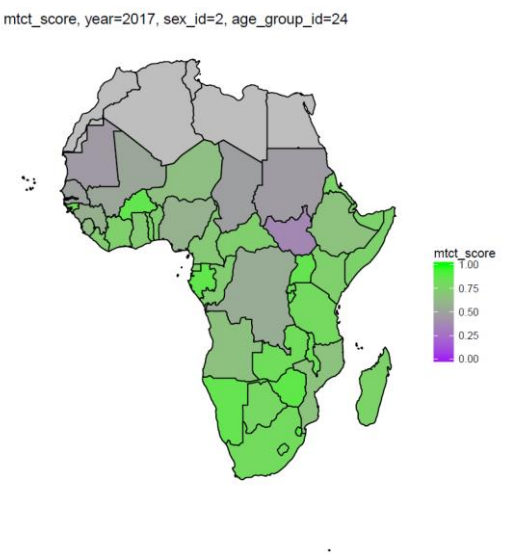












Map of the composite score of knowledge about misconceptions of HIV/AIDS (*mtct_score*)

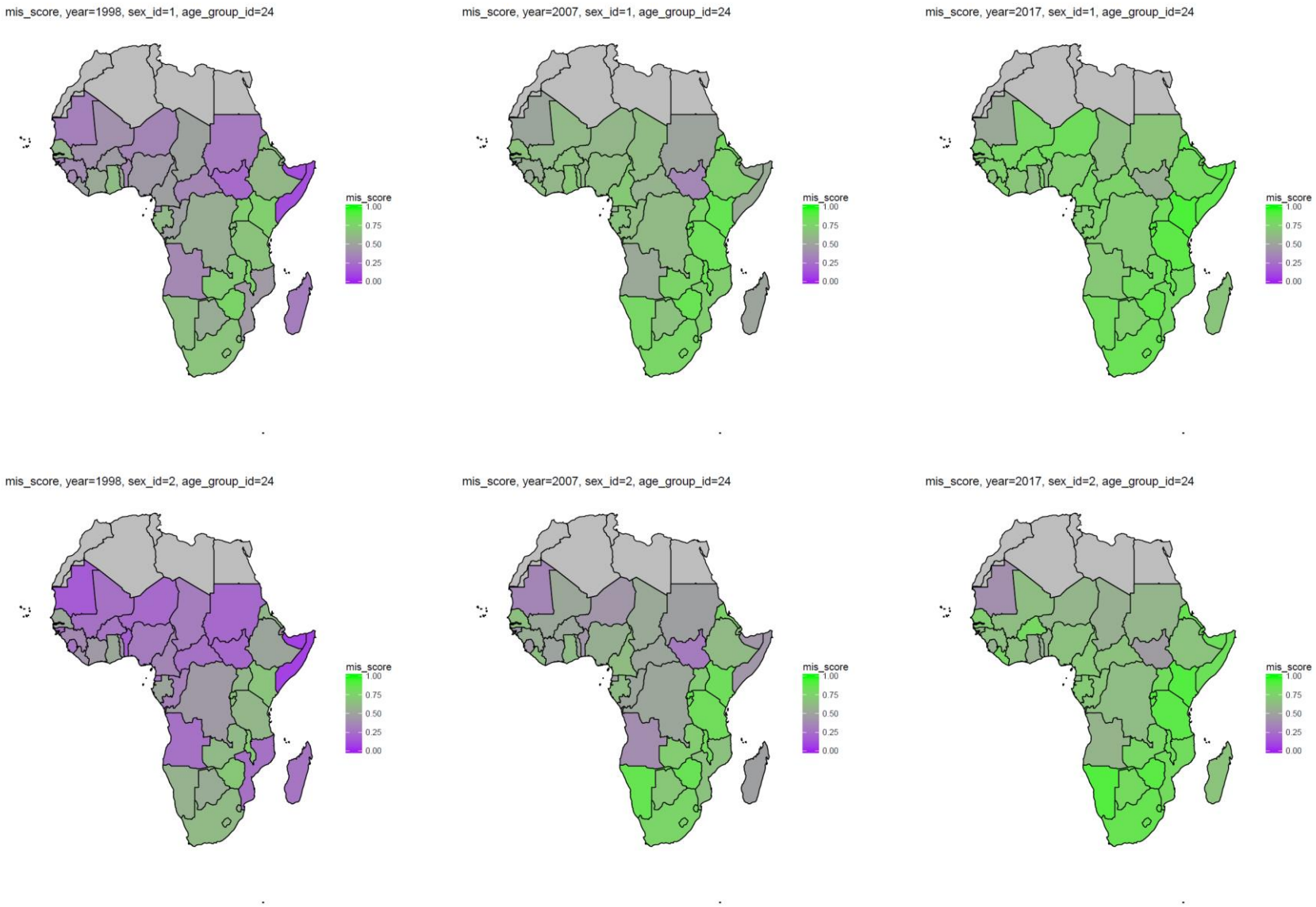


Figure 4-3 Maps of the composite scores in 1998, 2007, and 2017 by gender

4. Discussion

Our results show that people's overall knowledge and attitudes about HIV/AIDS have improved significantly in SSA over the past two decades, with their knowledge improving faster than their attitudes. However, there are still marked variations across SSA countries. People in Western SSA countries on average have the lowest knowledge and attitudes about HIV/AIDS but they have the highest annualized increase rate of the knowledge and attitudes about HIV/AIDS. On the contrary, people in Southern SSA countries have the highest knowledge and attitudes about HIV/AIDS but they have the lowest annualized increase rate. Broadly, places with very high HIV prevalence have greatest knowledge and best attitudes while those of lowest prevalence have poorest knowledge and worst attitudes, but the difference of HIV/AIDS knowledge and attitudes between countries are narrowing in general.

Men in SSA generally have better knowledge of and attitudes about HIV/AIDS than women. However, we are pleased to see that the gap between men and women is significantly narrowing over time, especially for knowledge on MTCT, in which women have surpassed men in the early 2000s. The contributors to this encouraging achievement of women are multifaceted and future studies are needed to fully understand them. However, this significant achievement of women coincides with the advent and upscale of PMTCT-ART programs in many SSA countries.[60,66–68] We believe that the PMTCT programs, which provide voluntary testing and counseling (VTC) services for HIV/AIDS and were started and scaled up in many SSA countries during the 2000s[60,66–68], have played an important role in improving knowledge and attitudes about HIV/AIDS for women. Being linked to the antenatal clinic, the PMTCT programs often serve as a key point of entry for women and children for HIV/AIDS education and services.[69] If future studies prove that PMTCT programs indeed greatly contribute to the increase of women's

knowledge on MTCT of HIV, there is an opportunity for us to further improve women's knowledge and attitudes about other HIV/AIDS indicators by incorporating a more comprehensive HIV/AIDS education in the PMTCT programs. Therefore, this would further narrow the gap between men and women.

Demographically, sub-Saharan Africa has the youngest population in the world, which is also growing fast. In 2012, more than 70% of the population in SSA were young people under 30 years old.[70] Young people, especially young women, are the key population for HIV/AIDS prevention because they are disproportionately affected by HIV/AIDS. The AIDS-related mortality among adolescents tripled between 2000 and 2015, the only age group for which ADIS-related mortality had increased during the period.[72] In the UNAIDS 2016 Prevention Gap Report, lack of comprehensive knowledge about HIV/AIDS is identified as one of the major barriers for HIV prevention among young people in SSA. In this study, although we find that young people's knowledge and attitudes about HIV/AIDS have improved over the past decades, compared with older people, young people generally have lower knowledge and attitudes about HIV/AIDS. It is even more concerning that the gap between older and young people have been widening in in recent years, with young people having lower annualized growth rate than older people.

Africa is in the middle of big demographic shift, with a huge adolescent bubble coming. There is also a large fraction of new HIV cases in adolescents and young adults.[74] However, our results suggest that the rising generation has not really known the illness. Meanwhile, the global financing of HIV/AIDS has shifted from primary prevention to treatment of HIV/AIDS,[75] which may have contributed to the widening knowledge gap between the young and the old. No matter what is the underlying cause of the widening knowledge gap between the young and the old, our

results clearly suggest that we should at least continue, if not increase, the efforts to improve young people's knowledge of HIV/AIDS in SSA.

Regarding methodology, this study uses MI to impute country-level missing proportions of key indicators across indicators and between age and sex groups. Compared with the traditional crosswalking using linear model, the current crosswalking method using MI has many benefits. First, MI solves data imbalance between gender, between age groups, and across indicators altogether and the imputation is done in one shot instead of by sex, by age group and by indicator. Second, MI can utilize *all* available information to impute the country-level missing proportions of key indicators. Besides the 16 key indicators, we can include other indicators of HIV/AIDS knowledge and attitudes as well as country-level covariates in the imputation model to further improve the imputation. Third, MI naturally accounts for the uncertainty of imputation by imputing the country-level missing proportions multiple times. In this chapter, we impute each country-level missing proportions 1000 times to capture uncertainty of the imputations. In addition to the three general benefits of using MI, as a special MI method for TSCS data, *Amelia* is fast and produces smoothed imputations over time, which makes the method particularly useful for this study.

Although we conducted this study carefully, it is not without limitations. The survey data used in this study were collected by different people, using different questionnaires and under different settings. There can be huge heterogeneity in the quality of data across surveys due to these discrepancies. However, we cannot verify the quality of each survey because we do not know what exactly happened during the data collection procedure. For report data, the data quality is even more concerning because it is not uncommon for people to make mistakes when calculating the national estimates. To mitigate this limitation, we have added NSV to the data to account for

the heterogeneity due to random errors and to adjust the bias due to survey type. In addition, we examine each survey carefully and remove extremely unusual data points especially if they are report data.

5. Conclusion

By including 220 national surveys of 47 SSA countries from 1998 to 2017, this study produces robust evidence on people's knowledge and attitudes about HIV/AIDS in SSA countries over the past decades. According to our findings, although there have been substantial improvements in people's knowledge and attitudes about HIV/AIDS in SSA, there are still efforts to be made. Specifically, our findings suggest that women and the young tend to have lower knowledge and poorer attitudes about HIV/AIDS than men and the old. Therefore, we urge that more resources should be used to improve knowledge and attitudes about HIV/AIDS among women and the young in the SSA countries. Especially, given the booming young population and the large fraction of new HIV cases among adolescents and young adults in SSA, policy makers as well as researchers should pay more attention to changes in young people's knowledge and attitudes about HIV/AIDS in this region.

Chapter 5 : Conclusion

Conclusion

This dissertation is a compilation of three publishable manuscripts that present findings from an analysis of national survey data. The aim of the dissertation is to provide solid evidence on changes in people's knowledge and attitudes about HIV/AIDS in sub-Saharan African countries over the past decades and to inform decision making and planning for HIV/AIDS prevention and treatment programs conducted in sub-Saharan Africa. Additionally, other researchers conducting HIV/AIDS research in sub-Saharan African can also benefit from the results.

Important findings have emerged from the set of three studies conducted in this research. In Chapter two, we identify 248 national surveys from 47 sub-Saharan African countries which contain information on 16 key indicators of people's knowledge and attitudes about HIV/AIDS. Using the 248 national surveys, we estimate the trends of the 16 key indicators in the 47 SSA countries from 1988 to 2017. Although there is great heterogeneity regarding levels and rates of change across countries, in sub-Saharan Africa, people's knowledge of HIV/AIDS and their attitudes toward people living with HIV/AIDS have, in general, improved over the past 30 years. However, one attitudes indicator, namely, people's attitudes toward disclosing family members' HIV/AIDS status shows a significant deteriorating trend across the countries. The indicator reflects people's perceived discrimination as family members of PLWHs and the deteriorating trends of the indicator highlights more HIV/AIDS stigma and its internalization in SSA countries. Among the 47 SSA countries, those in Somalia, Sudan, and South Sudan demonstrate lower levels of HIV/AIDS knowledge and accepting attitudes compared with people in other countries. Men in SSA generally have better knowledge and attitudes about HIV/AIDS than women do. However, the gap between men and women has been narrowing. In recent years, women's knowledge on mother-to-child transmission of HIV has even surpassed men's, which

may be due to the wide-spread PMTCT programs in these SSA countries. Lastly, older people in SSA tend to have better knowledge and attitudes about HIV/AIDS than younger people and the gap between the young and the old is widening, which is alarming given the booming young population and the large fraction of new HIV cases among adolescents and young adults in SSA.

In Chapter three, we evaluate the performance of seven multiple imputation methods in imputing missingness of country-level estimates of the knowledge and attitudes indicators in the survey due to questions not asked. Based on the results of the study, *Amelia* and *mice.2l.pan* methods are the best among the seven MI methods when imputing missingness in TSCS continuous data due to questions not asked in the survey. Both methods converge fast, produce reasonable and stable imputations and have small out-of-sample RMSE less than 0.05 and CR_95 very close to 95%. Amelia and MICE can also be implemented parallelly, which greatly reduces running time and makes the two methods more practical. In addition, we find that including incomplete auxiliary variables that are correlated with targeted incomplete variables improves the imputation performance regardless of the missing rate of the auxiliary variables. However, including cluster means of variables in the imputation model has little impact on the imputations.

In Chapter four, we use data from 220 national surveys to estimate trends of the composite scores of people's knowledge and attitudes about HIV/AIDS from 1998 to 2017. In this study, we use *Amelia* multiple imputation method to impute the country-level missing proportions of key indicators and to crosswalk data between sex and age groups at the same time. We then calculate the knowledge and attitudes composite scores using the estimates of the 16 key indicators of knowledge and attitudes about HIV/AIDS. Estimated trends of the composite scores show that people's overall knowledge and attitudes about HIV/AIDS have improved significantly in SSA over the past two decades, with their knowledge improving faster than their

attitudes. However, there are still marked variations across SSA countries. In general, places with very high HIV prevalence have greatest knowledge and best attitudes while those of lowest prevalence have poorest knowledge and worst attitudes, but the gap of HIV/AIDS knowledge and attitudes between countries has been narrowing in general. Our findings also suggest that women and the young tend to have lower knowledge and poorer attitudes about HIV/AIDS than men and the old. It is especially concerning that the gap of knowledge and attitudes scores between the young and the old has been widening. Given the booming young population and the large fraction of new HIV cases among adolescents and young adults in SSA, we urge that policy makers as well as researchers should pay more attention to changes in young people's knowledge and attitudes about HIV/AIDS in this region.

In summary, this dissertation takes all available national surveys across 16 key indicators of knowledge and attitudes about HIV/AIDS, and condenses them in a principle way to create new variables of knowledge and attitudes about HIV/AIDS. The new metrics of HIV/AIDS knowledge and attitudes open doors to future studies on HIV/AIDS. For example, we can now use the new metrics of HIV/AIDS knowledge and attitudes to explore contribution of knowledge and attitudes to the evolution of HIV/AIDS epidemic, which we could not do before using all the separated surveys. The new metrics can also be used to improve estimates of prevalence or incidence of HIV/AIDS or of other diseases. In terms of policy implications, the results of the dissertation suggest that PMTCT programs may have played an important role in improving women's knowledge and attitudes about HIV/AIDS, particularly about MTCT of HIV, and thus may be an effective platform to reach out to women and to educate them about HIV/AIDS or about other important health issues. Therefore, interventions targeted women may be more effective if they can be incorporated into the PMTCT programs. The results of the dissertation

also suggest that we need to pay more attentions to the younger generations' knowledge and attitudes about HIV/AIDS and to the internalized discrimination HIV patients and their family members perceive especially given the booming young population and the large fraction of new HIV cases among adolescents and young adults in sub-Saharan Africa.

Reference

1. World Health Organization. WHO | HIV/AIDS. *WHO* http://www.who.int/gho/hiv/en/ (2018).

2. UNAIDS. UNAIDS HIV/AIDS Fact Sheet. http://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf (2018).

3. World Health Organization. WHO | Definition of key terms. *WHO* https://www.who.int/hiv/pub/guidelines/arv2013/intro/keyterms/en/ (2018).

4. Institute for Health Metrics and Evaluation (IHME). *Financing Global Health 2016: Development Assistance, Public and Private Health Spending for the Pursuit of Universal Health Coverage.* http://www.healthdata.org/sites/default/files/files/policy_report/FGH/2017/IHME_FGH2016_Technical-Report.pdf (2017).

5. Albarracín, D. *et al.* A Test of Major Assumptions About Behavior Change: A Comprehensive Look at the Effects of Passive and Active HIV-Prevention Interventions Since the Beginning of the Epidemic. *Psychol. Bull.* **131**, 856–897 (2005).

6. GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet Lond. Engl.* **385**, 117–171 (2015).

7. Kaufman, M. R., Cornish, F., Zimmerman, R. S. & Johnson, B. T. Health behavior change models for HIV prevention and AIDS care: practical recommendations for a multi-level approach. *J. Acquir. Immune Defic. Syndr. 1999* **66 Suppl 3**, S250-258 (2014).

8. Catania, J. A., Kegeles, S. M. & Coates, T. J. Towards an understanding of risk behavior: An AIDS risk reduction model (ARRM). *Health Educ. Q.* **17**, 53–72 (1990).

9. Campbell, C. & Cornish, F. Towards a 'fourth generation' of approaches to HIV/AIDS management: creating contexts for effective community mobilisation. *AIDS Care* **22 Suppl 2**, 1569–1579 (2010).

10. Gupta, G. R., Parkhurst, J. O., Ogden, J. A., Aggleton, P. & Mahal, A. Structural approaches to HIV prevention. *The Lancet* **372**, 764–775 (2008).

11. Zeglin, R. J. & Stein, J. P. Social determinants of health predict state incidence of HIV and AIDS: a short report. *AIDS Care* **27**, 255–259 (2015).

12. Coates, T. J., Stall, R. D., Catania, J. A. & Kegeles, S. M. Behavioral factors in the spread of HIV infection. *AIDS Lond. Engl.* **2 Suppl 1**, S239-246 (1988).

13. Bettinghaus, E. P. Health promotion and the knowledge-attitudes-behavior continuum. *Prev. Med.* **15**, 475–491 (1986).

14. Fisher, J. D., Fisher, W. A., Misovich, S. J., Kimble, D. L. & Malloy, T. E. Changing AIDS risk behavior: effects of an intervention emphasizing AIDS risk reduction information, motivation, and behavioral skills in a college student population. *Health Psychol. Off. J. Div. Health Psychol. Am. Psychol. Assoc.* **15**, 114–123 (1996).

15. Bazargan, M., Kelly, E. M., Stein, J. A., Husaini, B. A. & Bazargan, S. H. Correlates of HIV risk-taking behaviors among African-American college students: the effect of HIV knowledge, motivation, and behavioral skills. *J. Natl. Med. Assoc.* **92**, 391–404 (2000).

16. Fisher, J. D., Fisher, W. A., Williams, S. S. & Malloy, T. E. Empirical tests of an information-motivation-behavioral skills model of AIDS-preventive behavior with gay men and heterosexual university students. *Health Psychol. Off. J. Div. Health Psychol. Am. Psychol. Assoc.* **13**, 238–250 (1994).

17. Chang, S. J., Choi, S., Kim, S.-A. & Song, M. Intervention Strategies Based on Information-Motivation-Behavioral Skills Model for Health Behavior Change: A Systematic Review. *Asian Nurs. Res.* **8**, 172–181 (2014).

18. Durojaiye, O. C. Knowledge, attitudes and practice of HIV/AIDS: Behavior change among tertiary education students in Lagos, Nigeria. *Ann. Trop. Med. Public Health* **4**, 18 (2011).

19. Thanavanh, B., Harun-Or-Rashid, Md., Kasuya, H. & Sakamoto, J. Knowledge, attitudes and practices regarding HIV/AIDS among male high school students in Lao People's Democratic Republic. *J. Int. AIDS Soc.* **16**, (2013).

20. Sohn, A. & Cho, B. Knowledge, Attitudes, and Sexual Behaviors in HIV/AIDS and Predictors Affecting Condom Use among Men Who Have Sex with Men in South Korea. *Osong Public Health Res. Perspect.* **3**, 156–164 (2012).

21. Fabrigar, L. R., Petty, R. E., Smith, S. M. & Crites, S. L. Understanding knowledge effects on attitudes-behavior consistency: The role of relevance, complexity, and amount of knowledge. *J. Pers. Soc. Psychol.* **90**, 556–577 (2006).

22. Miller, T. E., Booraem, C., Flowers, J. V. & Iversen, A. E. Changes in knowledge, attitudes, and behavior as a result of a community-based AIDS prevention program. *AIDS Educ. Prev. Off. Publ. Int. Soc. AIDS Educ.* **2**, 12–23 (1990).

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

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

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

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

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

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

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

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

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

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

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

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

35. Institute for Health Metrics and Evaluation (IHME). Global Health Data Exchange | GHDx. http://ghdx.healthdata.org/ (2019).

36. Demographic and Health Survey. DHS Model Questionnaire -Phase 7. (2018).

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

38. Wang, H. *et al.* Global, regional, and national under-5 mortality, adult mortality, age-specific mortality, and life expectancy, 1970–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet* **390**, 1084–1150 (2017).

39. James, S. L. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet* **392**, 1789–1858 (2018).

40. Wang, H. *et al.* Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980â 2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet* **388**, 1459–1544 (2016).

41. Reitsma, M. B. *et al.* Smoking prevalence and attributable disease burden in 195 countries and territories, 1990–2015: a systematic analysis from the Global Burden of Disease Study 2015. *The Lancet* **389**, 1885–1906 (2017).

42. Knowles, J. & Frederick, C. Prediction Intervals from merMod Objects. https://cran.rstudio.com/web/packages/merTools/vignettes/Using_predictInterval.html (2018).

43. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using **lme4**. *J. Stat. Softw.* **67**, (2015).

44. Foreman, K. J., Lozano, R., Lopez, A. D. & Murray, C. J. Modeling causes of death: an integrated approach using CODEm. *Popul. Health Metr.* **10**, 1 (2012).

45. Denœux, T. Lecture 7: Splines and Generalized Additive Models - Computational Statistics. (2016).

46. Maindonald, J. Smoothing Terms in GAM Models. 15 (2010).

47. Eubank, R. L. *Nonparametric Regression and Spline Smoothing, Second Edition*. (CRC Press, 1999).

48. Rosenberg, P. S. Hazard Function Estimation Using B-Splines. *Biometrics* **51**, 874–887 (1995).

49. Shepherd, B. E. & Rebeiro, P. F. Assessing and interpreting the association between continuous covariates and outcomes in observational studies of HIV using splines. *J. Acquir. Immune Defic. Syndr. 1999* **74**, e60–e63 (2017).

50. Chan, B. T. & Tsai, A. C. HIV knowledge trends during an era of rapid antiretroviral therapy scale-up: an analysis of 33 sub-Saharan African countries. *J. Int. AIDS Soc.* **21**, e25169 (2018).

51. Balogun, A. S. Islamic perspectives on HIV/AIDS and antiretroviral treatment: the case of Nigeria. *Afr. J. AIDS Res. AJAR* **9**, 459–466 (2010).

52. Hasnain, M. Cultural Approach to HIV/AIDS Harm Reduction in Muslim Countries. *Harm. Reduct. J.* **2**, 23 (2005).

53. Ghaly, M. COLLECTIVE RELIGIO-SCIENTIFIC DISCUSSIONS ON ISLAM AND HIV/AIDS: I. BIOMEDICAL SCIENTISTS: with Mohammed Ghaly, "Islamic Bioethics in the Twenty-first Century"; Henk ten Have, "Global Bioethics: Transnational Experiences and Islamic Bioethic. *Zygon*® **48**, 671–708 (2013).

54. Oster, A. M. *et al.* Prevalence of HIV, sexually transmitted infections, and viral hepatitis by Urbanicity, among men who have sex with men, injection drug users, and heterosexuals in the United States. *Sex. Transm. Dis.* **41**, 272–279 (2014).

55. Williams, P. B. & Sallar, A. M. HIV/AIDS and African American men: urban-rural differentials in sexual behavior, HIV knowledge, and attitudes towards condoms use. *J. Natl. Med. Assoc.* **102**, 1139–1149 (2010).

56. Yehadji, D. Urban-rural disparities in HIV related knowledge, behavior and attitudes in Burkina Faso: Evidence from Burkina Faso Demographic and Health Survey 2010. 40 (2015).

57. Ayodele, O. & Ayodele, O. M. Urban-Rural Differentials in HIV/AIDS Knowledge of Nigerian Senior Secondary School Students. *Int. J. Health Sci.* **4**, (2016).

58. Gunn, J. K. L. *et al.* Antenatal care and uptake of HIV testing among pregnant women in sub-Saharan Africa: a cross-sectional study. *J. Int. AIDS Soc.* **19**, (2016).

59. Teshome, R. & Youjie, W. Comparison and Association of Comprehensive HIV/AIDS Knowledge and Attitudes towards people Living with HIV/AIDS among Women Aged 15-49 in Three East African Countries: Burundi, Ethiopia and Kenya. *J. AIDS Clin. Res.* **07**, (2016).

60. WHO. Prevention of Mother-To-Child Transmission (PMTCT) Briefing Note. (2007).

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

62. Rosenmai, P. Using the Median Absolute Deviation to Find Outliers. https://eurekastatistics.com/using-the-median-absolute-deviation-to-find-outliers/ (2013).

63. Wood, S. N. *Generalized Additive Models: An Introduction with R*. (Chapman and Hall/CRC, 2006).

64. Hadley Wickham *et al.* Create Elegant Data Visualisations Using the Grammar of Graphics, package ggplot2. (2018).

65. Simon Wood. R: Smooth terms in GAM. https://stat.ethz.ch/R-manual/R-devel/library/mgcv/html/smooth.terms.html.

66. Barron, P. *et al.* Eliminating mother-to-child HIV transmission in South Africa. *Bull. World Health Organ.* **91**, 70–74 (2013).

67. Buchanan, A. M. *et al.* Progress in the Prevention of Mother to Child Transmission of HIV in Three Regions of Tanzania: A Retrospective Analysis. *PLOS ONE* **9**, e88679 (2014).

68. WETTSTEIN, C. *et al.* Missed Opportunities to Prevent Mother-to-Child-Transmission in sub-Saharan Africa: Systematic Review and Meta-Analysis. *AIDS Lond. Engl.* **26**, 2361–2373 (2012).

69. The United States President's Emergency Plan for AIDS Relief. III. Strengthening PMTCT Programs: Prevention First. https://www.pepfar.gov/reports/progress/76906.htm.

70. UNESCO. Statistics on Youth | United Nations Educational, Scientific and Cultural Organization. http://www.unesco.org/new/en/unesco/events/prizes-and-

celebrations/celebrations/international-days/world-radio-day-2013/statistics-on-youth/ (2013).

71. United Nation. Youth population trends and sustainable development. (2015).

72. UNICEF. Adolescent deaths from AIDS tripled since 2000. *UNICEF* https://www.unicef.org/media/media_86384.html (2015).

73. UNAIDS. Prevention gap report. http://www.unaids.org/en/resources/documents/2016/prevention-gap (2016).

74. UNAIDS. Turning the tide against AIDS will require more concentrated focus on adolescents and young people. *Adolescent HIV prevention* https://data.unicef.org/topic/hivaids/adolescents-young-people/ (2019).

75. Haakenstad, A. *et al.* Potential for additional government spending on HIV/AIDS in 137 low-income and middle-income countries: an economic modelling study. *Lancet HIV* **6**, e382–e395 (2019).

76. Fedor, T. Changes in HIV/AIDS Knowledge, Attitudes and Behaviors in Malawi. (2014).

77. Faust, L., Ekholuenetale, M. & Yaya, S. HIV-related knowledge in Nigeria: a 2003-2013 trend analysis. *Arch. Public Health* **76**, (2018).

78. Umeh, C. N., Essien, E. J., Ezedinachi, E. N. & Ross, M. W. Knowledge, Beliefs and Attitudes about HIV/AIDS related issues, and the Sources of Knowledge among Health Care Professionals in Southern Nigeria. *J. R. Soc. Promot. Health* **128**, 233–239 (2008).

79. Delobelle, P. *et al.* HIV/AIDS knowledge, attitudes, practices and perceptions of rural nurses in South Africa. *J. Adv. Nurs.* **65**, 1061–1073 (2009).

80. Nachega, J. B. *et al.* HIV/AIDS and Antiretroviral Treatment Knowledge, Attitudes, Beliefs, and Practices in HIV-Infected Adults in Soweto, South Africa. *JAIDS J. Acquir. Immune Defic. Syndr.* **38**, 196 (2005).

81. Reddy, P. & Frantz, J. HIV/AIDS knowledge, behaviour and beliefs among South African university students. *SAHARA J J. Soc. Asp. HIVAIDS Res. Alliance* **8**, 166–170 (2011).

82. Ojieabu, W. A., Femi-Oyewo, M. N. & Eze, U. I. HIV/AIDS Knowledge, Attitudes and Risk Perception among Pregnant Women in a Teaching Hospital, Southwestern Nigeria. *J. Basic Clin. Pharm.* **2**, 185–198 (2011).

83. Burgoyne, A. D. & Drummond, P. D. Knowledge of HIV and AIDS in women in sub-Saharan Africa. **12**, 18 (2008).

84. Mishra, V., Agrawal, P., Alva, S., Gu, Y. & Wang, S. Changes in HIV-related knowledge and behaviors in sub-Saharan Africa. (2009).

85. IHME. Social Determinants of Health Visualization | IHME Viz Hub. http://vizhub.healthdata.org/sdh.

86. Institute for Health Metrics and Evaluation (IHME). *A Hand Up: Global Progress Toward Universal Education*. http://www.healthdata.org/sites/default/files/files/policy_report/2015/PolicyReport_IHME_EducationalAttainment_2015.pdf (2015).

87. Kagaayi, J. & Serwadda, D. The History of the HIV/AIDS Epidemic in Africa. *Curr. HIV/AIDS Rep.* **13**, 187–193 (2016).

88. AIDS.gov. A Timeline of HIV and AIDS. *HIV.gov* https://www.hiv.gov/hiv-basics/overview/history/hiv-and-aids-timeline (2016).

89. Avert. History of HIV and AIDS overview. *AVERT* https://www.avert.org/professionals/history-hiv-aids/overview (2015).

90. Center for Disease Control and Prevention. HIV and AIDS Timeline. *CDC* https://npin.cdc.gov/pages/hiv-and-aids-timeline.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

169. Kharsany, A. B. M. & Karim, Q. A. HIV Infection and AIDS in Sub-Saharan Africa: Current Status, Challenges and Opportunities. *Open AIDS J.* **10**, 34–48 (2016).

170. Wang, H. *et al.* Estimates of global, regional, and national incidence, prevalence, and mortality of HIV, 1980–2015: the Global Burden of Disease Study 2015. *Lancet HIV* **3**, e361–e387 (2016).

171. Yaya, S., Bishwajit, G., Danhoundo, G. & Seydou, I. Extent of Knowledge about HIV and Its Determinants among Men in Bangladesh. *Front. Public Health* **4**, (2016).

172. Nubed, C. K. & Akoachere, J.-F. T. K. Knowledge, attitudes and practices regarding HIV/AIDS among senior secondary school students in Fako Division, South West Region, Cameroon. *BMC Public Health* **16**, (2016).

173. Shokoohi, M. *et al.* HIV Knowledge, Attitudes, and Practices of Young People in Iran: Findings of a National Population-Based Survey in 2013. *PLoS ONE* **11**, (2016).

174. Farotimi, A. A., Nwozichi, C. U. & Ojediran, T. D. Knowledge, attitudes, and practice of HIV/AIDS-related stigma and discrimination reduction among nursing students in southwest Nigeria. *Iran. J. Nurs. Midwifery Res.* **20**, 705–711 (2015).