EXPLORING THE ASSOCIATION BETWEEN SEXUAL BEHAVIOR, SOCIO-DEMOGRAPHIC AND BIOLOGICAL FACTORS WITH HIV INFECTION USING DATA FROM THE 2011 UGANDA AIDS INDICATOR SURVEY (UAIS).

By

CHARLES LUSWATA, BSTAT

APPROVED:

JOSÉ-MIGUEL YAMAL, PHD

ALAN G. NYITRAY, PHD

MINJAE LEE, PHD

DEDICATION

To my parents, Charles and Rose

In memory.

To Yen, Caleb, and Aaron

Without your love and support, this would not have been possible.

EXPLORING THE ASSOCIATION BETWEEN SEXUAL BEHAVIOR, SOCIO-DEMOGRAPHIC AND BIOLOGICAL FACTORS WITH HIV INFECTION USING DATA FROM THE 2011 UGANDA AIDS INDICATOR SURVEY (UAIS).

by

CHARLES LUSWATA, BSTAT

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH
Houston, Texas
December, 2015

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to the UTSPH faculty – notably Dr. José-Miguel Yamal, Dr. Alan G. Nyitray, Dr. MinJae Lee and Dr. Adriana Pérez for your guidance, patience, support, and willingness to help at all times. I would also want to thank my classmates and friends for the encouragement and support throughout the years. Last but not the least, I would wish to thank my wife, kids, siblings and Mama Nong for the wonderful support throughout this whole process.

EXPLORING THE ASSOCIATION BETWEEN SEXUAL BEHAVIOR, SOCIO-DEMOGRAPHIC AND BIOLOGICAL FACTORS WITH HIV INFECTION USING DATA FROM THE 2011 UGANDA AIDS INDICATOR SURVEY (UAIS).

Charles Luswata, BSTAT, MS
The University of Texas
School of Public Health, 2015

Thesis Chair: José-Miguel Yamal, PHD

New annual HIV infections have been increasing steadily in Uganda despite a decreasing HIV prevalence. This thesis aimed to identify sexual behavior, socio-demographic, biological and other factors associated with HIV infection.

Data extracted from the 2011 UAIS, a nationally representative two-staged stratified sample, for sexually active respondents aged 15-59 who provided a blood sample for HIV and syphilis testing were used for modeling (n=18,395). The first model, a design-based logistic regression, was used to examine the association between gender and self-perceived risk of getting HIV interaction effect with HIV infection. The second model, a classification tree, was used to identify important complex higher level interaction effects.

Overall, the prevalence of HIV in this study population was 8.15%. Results from the unadjusted analyses show males with a low compared to those with a high self-perceived risk of getting HIV were more likely to engage in high-risk sexual behaviors. After adjusting for other factors, self-perceived risk of getting HIV was significantly associated with HIV infection for uncircumcised male respondents only (p-value = 0.0018). The odds of HIV infection among uncircumcised males with a high self-perceived risk of getting HIV were 65% higher compared to uncircumcised male respondents with a low self-perceived risk of getting HIV (adjusted (adj.) OR = 1.65, 95% CI: 1.21 – 2.26). Also, circumcised males with a high self-perceived risk of contracting HIV had 49% lower odds of HIV infection when compared with uncircumcised males with a high self-perceived risk of contracting HIV (adj. OR = 0.51, 95% CI: 0.28 –0.92).

Other factors that were significantly associated with HIV infection included: being married (adj. OR = 1.55, 95% CI: 1.08 – 2.24) or divorced/separated (adj. OR =2.10, 95% CI: 1.35 – 3.26) or widowed (adj. OR =3.74, 95% CI: 2.24 – 6.23) compared to never married, alcohol consumption during sex (adj. OR =1.28, 95% CI: 1.03 – 1.58), engaging in commercial or exchange sex (adj. OR =2.07, 95% CI: 1.36 – 3.17), having an STI (adj. OR =1.60, 95% CI: 1.33 – 1.93), having 2-3 (adj. OR =1.56, 95% CI: 1.17 – 2.08) or 4 or more (adj. OR =1.98, 95% CI: 1.47 –2.68) total lifetime number of sex partners compared to having one, and having 1-2 (adj. OR =1.42, 95% CI: 1.15 – 1.77) or 3-4 (adj. OR =1.32, 95% CI: 1.01 – 1.73) biological children away from home compared to not having a child away from home.

The classification tree identified four potential interaction patterns. However, only *having an STI* × *total lifetime number of sex partners* was found to be statistically significant after evaluation in the design-based logistic regression model.

Overall, the findings of this study provide an insight into Uganda's HIV epidemic. A number of factors, and interactions amongst them, were found to be associated with HIV infection in 2011. Outstandingly, this data provided evidence that self-perceived risk of getting HIV is significantly associated with HIV infection for uncircumcised males only. Confirmation of these findings in prospective studies may be needed.

TABLE OF CONTENTS

LIST OF TABLES i
LIST OF FIGURES ii
LIST OF APPENDICES iii
BACKGROUND 1
HIV/AIDS definition, history and transmission 1
The Global and Sub-Saharan HIV/AIDS Epidemic 2
HIV/AIDS in Uganda 3
Sexual Behaviors and HIV infection 6
Male circumcision and HIV infection 7
Perceived risk of HIV infection 8
Rationale 9
Public Health Significance 11
Specific Aims 11
METHODS 12
Study Design 12
The 2011 Uganda AIDS Indicator Survey 12
Data collection 12
Sample design 13
Selection of data sets 15
Inclusion Criteria 16
Outcome Variable 16
Explanatory Variables 16
Statistical Analysis 23
Logistic regression model 24
Interaction (effect-measure modification) 27
Classification and Regression Trees (CART) analysis 28
Goodness-of-fit and regression diagnostics 30
Missing data 32
Statistical Power 35
Software 35
Human Subjects, Animal Subjects, or Safety Considerations 35
RESULTS 36
Exploratory Data analysis 36
HIV prevalence 38
Univariate analysis 42
Multivariable analysis 46
Missing data analysis 52
CART analysis 57
DISCUSSION 63
CONCLUSION 65
APPENDICES 66
REFERENCES 73

LIST OF TABLES

Table 1: survey variable description 18

Table 2: Background characteristics of the study population 37

Table 3: Subject characteristics by HIV results 39

Table 4: Selected factors against gender and self-perceived risk of HIV infection 41

Table 5: Factors associated with HIV infection based on Multivariable weighted logistic regression model (Model 1)... 49

Table 6: Proportion of missing data (weighted) 53

Table 7: Comparison of the final model with missing data imputed (model 3) and the model with missing data excluded (model 1) 54

Table 8: Multivariable weighted logistic regression models for CART analysis 60

LIST OF FIGURES

Figure 1: Map of Uganda.. 4

Figure 2: Estimates of HIV new infections and prevalence over time in Uganda. 5

Figure 3: Multilevel structure of the data... 14

Figure 4: Recursive binary splitting with corresponding decision tree. 28

Figure 5: Scatter plot matrix to assess collinearity 44

Figure 6: Weighted empirical logit plots for assessing linearity for age and age at first sexual intercourse in years 45

Figure 7: ROC curve from the weighted multivariable logistic regression model 50

Figure 8: Diagnostic plots for the weighted multivariable logistic regression model 46

Figure 9: Weighted CART model for HIV data 58

Figure 10: ROC for the CART model 59

LIST OF APPENDICES

Appendix A: The Institutional Review Board (IRB) Approval Letter 66

Appendix B: Plots of estimated probability of HIV infections 67

Appendix C: Checking for complete separation 69

Appendix D: Patterns of missing data... 71

BACKGROUND

HIV/AIDS definition, history and transmission

The Human Immunodeficiency Virus (HIV) is a lentivirus that causes Acquired Immunodeficiency Syndrome (AIDS). HIV readily mutates resulting into different strains of the virus that are classified into types, groups and subtypes. There are 2 types of the virus; HIV-1 (classified into four groups: M, N, O and P) and HIV-2 (with groups A-H) with HIV-1 group M being the principal cause of the AIDS pandemic (Sharp et al., 2001; Sharp & Hahn, 2011). HIV-1 group M viruses are further divided into 9 subtypes known as clades (A–D, F–H, J, and K). Clade B is prevalent in the developed world while C and recombinant strain A/G are common in the heavily infected regions of southern Africa and West Africa, respectively. The recombinant strain B/C is prevalent in China (M. S. Cohen, Hellmann, Levy, DeCock, & Lange, 2008).

Many theories surround the origin of HIV but the closest viral resemblance is the Simian Immunodeficiency Virus (SIV) found in primates living in south-eastern Cameroon Western Africa (Sharp et al., 2001; Sharp & Hahn, 2011). The SIVs might have found their way into humans during the butchering of primates including chimpanzees and mangabeys that are routinely hunted by men. This cross-species transmission of SIV resulted in viruses that spread among humans. The timing of when the virus might have been transmitted to humans is estimated to be near the beginning of the twentieth century, sometime between 1884 and 1924 for HIV-1 and 1924-1959 for HIV-2 (Sharp & Hahn, 2011; Wertheim & Worobey, 2009).

The first cases of AIDS in the United States were reported by the Centers for Disease Control and Prevention (CDC) in 1981 in gay men. A number of gay men in California and New York had developed Pneumocystis carinii pneumonia (PCP), Kaposi's sarcoma (KS) and other opportunistic infections all related to a common exposure (CDC, 1981; Janier et al., 1984). Meanwhile in 1982, people were dying of an unknown disease in south-western Uganda which was first described as the "slim disease" due to the extreme weight loss and diarrhea associated with it. The disease occurred predominantly in individuals that frequently practiced casual sex with different partners and it was later found to be HIV/AIDS (Serwadda et al., 1985).

HIV-1 can be transmitted from an infected person to another through a number of sources, including contaminated blood and blood products (e.g. medical injections, blood transfusions, and injection drug use), an infected mother transmitting the virus to her baby (before, during, or after birth and through breast milk), and most frequently through either vaginal or anal unprotected sexual intercourse (M. S. Cohen et al., 2008). Once in the human body, HIV attacks the human immune system cells, CD4 T cells specifically, replicates itself and causes the depletion of the T cells thereby making the human body susceptible to opportunistic infections like tuberculosis and various cancers. If it is not diagnosed early or not treated with antiretroviral therapy (ART) to keep the level of HIV viral load in the body low, HIV infected individuals can develop AIDS. There is still no vaccine or cure for the virus (CDC[Internet], 2014).

The Global and Sub-Saharan HIV/AIDS Epidemic

By the end of 2010, approximately 34 million people globally were living with HIV/AIDS, 97% of whom reside in low- and middle-income countries (WHO, UNICEF, & UNAIDS, 2011). While the global adult HIV prevalence was estimated at 0.8% in 2010, the HIV prevalence in Sub-Saharan Africa (SSA) was estimated to be 5.2% (Manda, Lombard, & Mosala, 2012). SSA, home to 12% of the world's population, continues to be disproportionately affected by the HIV/AIDS epidemic with 1 in every 20 adults living with HIV/AIDS. SSA accounts for 68% of all people living with HIV worldwide (WHO et al., 2011). Furthermore, the HIV prevalence in SSA varies by region. Countries in Southern Africa have the highest HIV prevalence varying between 15 - 30% while countries in West and Central Africa have a prevalence between 1 - 5% and those in East Africa 3 – 7% (WHO et al., 2011; Manda et al., 2012).

The global HIV incidence has declined since 1997. There were 2.7 million new HIV infections in 2010, which was 15% less than in 2001 and 21% below the number of new infections at the peak of the epidemic in 1997. SSA accounted for 70% of the new HIV infections in 2010 although the total number of new infections in this region has dropped by more than 26%, from 2.6 million in 1997 to 1.9 million in 2010 (WHO et al., 2011).

Between 1981 and 2010, more than 25 million people died of HIV/AIDS globally (UNAIDS, 2012). Nonetheless, the deaths from AIDS and related illnesses have also decreased from a peak of 2.2 million in 2005 to an estimated 1.8 million in 2010. This decrease was mainly due to increased access to ART, the drugs that suppresses HIV in the body thereby reducing progression and transmission of the disease. In total, 2.5 million deaths have been averted in low- and middle-income countries since 1995 (WHO et al., 2011).

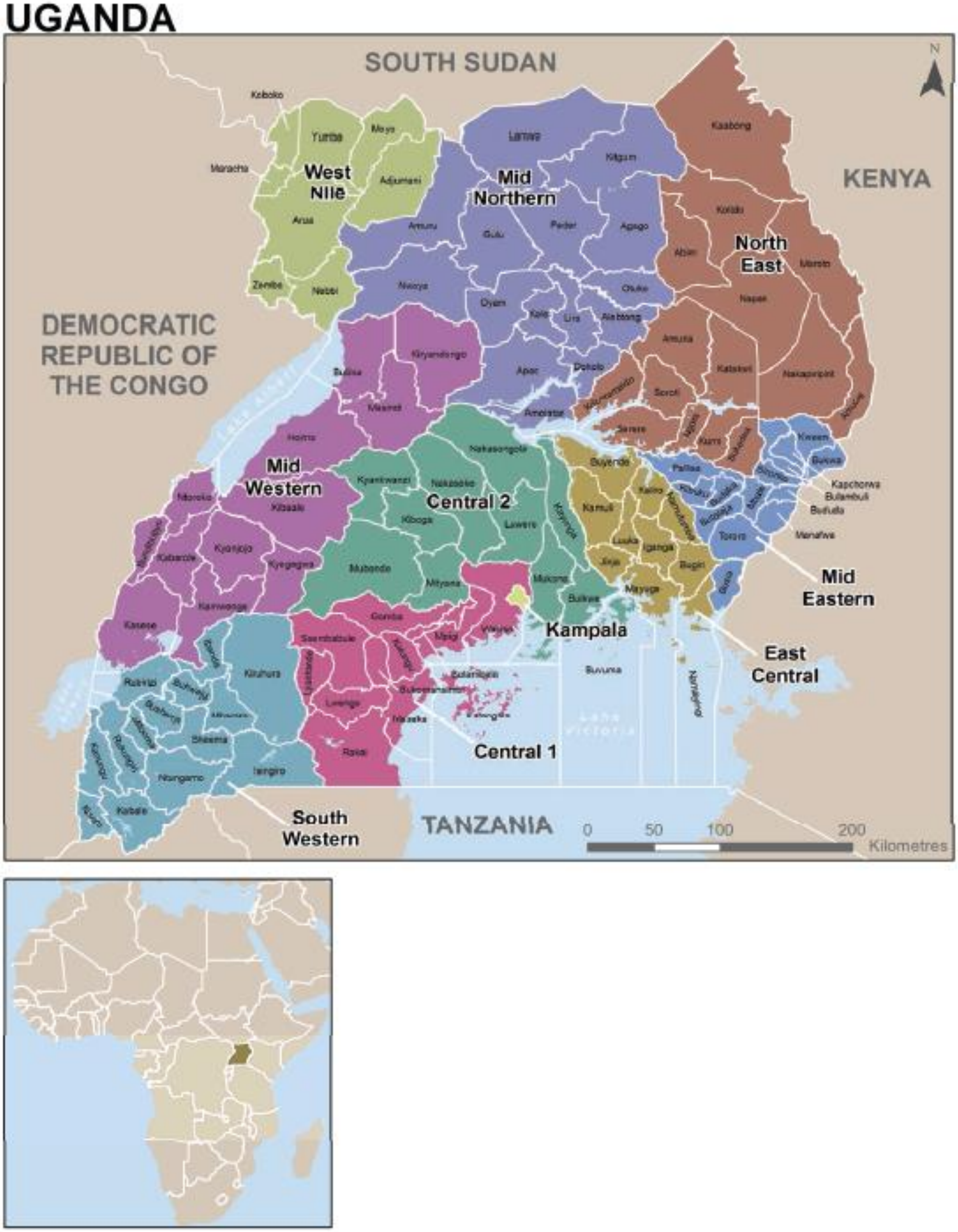
In the developed world, HIV infections are highest among men who have sex with men (MSM) (CDC[Internet], 2014). According to the CDC, approximately 50% of persons diagnosed with an HIV infection in the US by the end of 2010 were MSM yet this group comprised only 2% of the US population (CDC[Internet], 2014). In the developing world, the HIV/AIDS epidemic is generalized and spread mainly through vaginal intercourse. However, the few studies carried out among MSM in Africa indicated a higher HIV prevalence among this population than those in heterosexual relationships (Baral et al., 2009; Hladik et al., 2012).

HIV/AIDS in Uganda

The republic of Uganda is an East African country lying across the equator, about 800 kilometers inland from the Indian Ocean. It lies between 1^{0} 29’ South and 4^{0} 12’ North latitude, 29^{0} 34’ East and 35^{0} 0’ East longitude. The country is landlocked, bordered by Kenya in the East; South Sudan in the North; Democratic Republic of Congo in the West; Tanzania in the South; and Rwanda in South West. The country has an area of 241,550 square kilometers of which the land area covers 199,807 square kilometers (Uganda Bureau of Statistics (UBOS) [Internet], 2013).

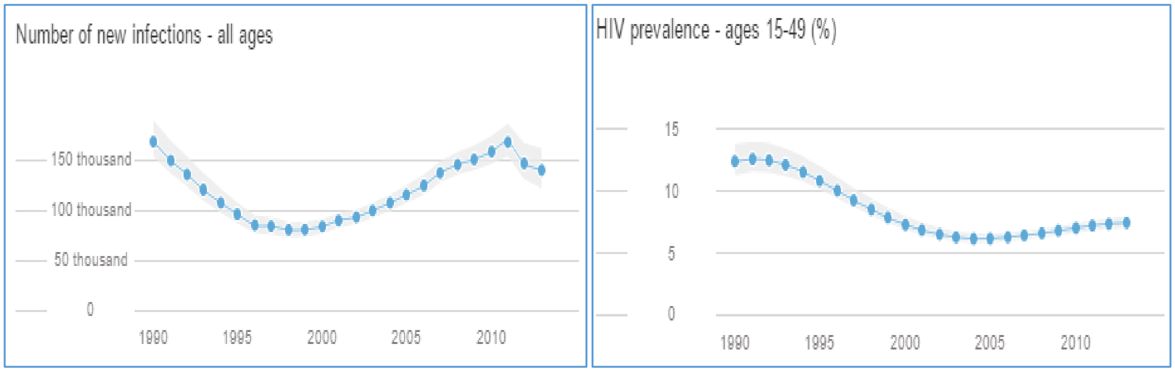
According to the Uganda Bureau of Statistics (UBOS), the population of Uganda increased from 24.2 million people in 2002 to an estimated 34.1 million people in 2012. The total fertility rate, defined as the number of children a woman would have by the end of her childbearing years (15-49), was at 6.2 in 2012 (Uganda Bureau of Statistics (UBOS) and Macro International Inc., 2012).

Figure 1: Map of Uganda.



Uganda has endured a severe HIV/AIDS epidemic for over a quarter of a century. However, Uganda is one of the countries that has reduced its HIV prevalence from double digits to single digits. In 1992, the national average adult HIV prevalence peaked at 18.5% but declined to 5% in 2000 mainly due to strong political leadership, a comprehensive approach (approached as more than a health issue) to combating the epidemic, and a strong multi-sectoral, decentralized, and community response (WHO et al., 2011). A number of HIV prevention, care, treatment, and support interventions were implemented. These included promotion of "zero grazing" (faithfulness and partner reduction) and the "ABC behaviors" for abstinence, be faithful (fidelity) and condom use (MOH, ICF International, CDC, USAID, & WHO, 2012; Murphy, Greene, Mihailovic, & Olupot-Olupot, 2006). Between 2001 and 2011, the HIV prevalence stabilized at an average 6.4%. However, the 2011 Uganda AIDS Indicator Survey (UAIS) reported an increase in prevalence to 6.7% with more women (7.7%) than men (5.6%) living with HIV/AIDS. Also, HIV prevalence was highest among individuals in the age group 35-39 years (10.3%) and those living in urban areas with high population densities (8.7%) (MOH et al., 2012).

Figure 2: Estimates of HIV new infections and prevalence over time in Uganda.



Source: UNAID, Global Report 2012: AIDSinfo (UNAIDS, 2012).

Approximately 80% of HIV infections were attributable to having unprotected sexual intercourse with an infected partner whereas mother-to-child transmission accounts for approximately 20% and blood borne and other infections account for less than 1% (WHO et al., 2011). For this study, sexual behavior, social-demographic and biological factors that influence the spread of HIV/AIDS in Uganda were explored.

Sexual Behaviors and HIV infection

Sexual behaviors are defined as high-risk if they increase the likelihood of acquiring or transmitting HIV. In Uganda, previously documented high risk sexual behaviors include: early coital debut (age at first sexual intercourse less than 18 years), multiple and concurrent sexual partnerships, total lifetime number of sex partners greater than four, inconsistent condom use, transactional sex, as well as alcohol and drug use before and during sex (Bolton, 1992; Chimoyi & Musenge, 2014; Kibira, Nansubuga, Tumwesigye, Atuyambe, & Makumbi, 2014; Konde-Lule et al., 1997; Lema, Katapa, & Musa, 2008; Serwadda et al., 1992; Tumwesigye et al., 2012).

The association between high-risk sexual behaviors and HIV infection has been widely documented but gaps still exist in the epidemiology (Chimoyi & Musenge, 2014; Hrdy, 1987; Kibira et al., 2014; Konde-Lule et al., 1997; Tumwesigye et al., 2012; Wawer et al., 1991). The HIV epidemic tends to concentrate in key groups of people including commercial sex workers, uniformed services, the fishing communities, truck drivers and men having sex with men who practice high-risk sexual behaviors (Kiwanuka et al., 2014). Based on research, intervention programs have targeted high-risk groups of people over the years although HIV infections have slightly increased in the recent past.

Multiple sexual partnerships, particularly concurrent partnerships, either pre- or post-marriage, are highly associated with HIV infection (Bolton, 1992; Hrdy, 1987). Concurrent sexual relationships occur when an individual has overlapping sexual relationships with more than one person over a period of time. Unlike serial monogamy (one partnership at a time), long-term concurrent or simultaneous partnerships might amplify the spread of HIV because they link people into a web of sexual relationships that can extend across huge regions (Epstein, 2008). If any one of the individuals involved in a long-term simultaneous sexual partnership acquires HIV, the rest of the individuals (even the monogamous partners) may be placed at increased risk of infection especially when condoms are not used consistently. However, having multiple sexual partners (polygamy) is accepted and increasingly practiced in some traditional African cultures and religious groups especially the Muslims (Hrdy, 1987; Smolak, 2010). Several researchers have published about polygamy being a protective factor against HIV infection among Muslims in the middle-East and SSA (P. B. Gray, 2004; Kamarulzaman, 2013; Smolak, 2010). The lower HIV prevalence might be due to factors other than polygamy. These include male circumcision, adequate economic resources on the part of the man, and adherence to Islamic beliefs forbidding alcohol and drug use, sex before marriage, extramarital sexual relations, and homosexuality (P. B. Gray, 2004; Kamarulzaman, 2013; Smolak, 2010). However, annual new HIV infections increased by more than 35% between 2001 and 2010 (43000 to 59000) in the Middle-East and North Africa. These new infections were mostly concentrated among high-risk populations like injectable drug users, female sex workers and MSM (Kamarulzaman, 2013; WHO et al., 2011).

Condom use is so far the most effective measure for preventing the transmission of HIV and some other sexually transmitted diseases (Lema et al., 2008). Inconsistent and incorrect use of condoms is associated with HIV/AIDS and other sexually transmitted diseases. According to previous research in Uganda, inconsistent use of male condoms has been found among people of older ages, the less educated, people residing in rural areas with limited access to health centers, and couples with one of the partners older by more than 5 years (Bankole, Ahmed, Neema, Ouedraogo, & Konyani, 2007; Matovu & Ssebadduka, 2013). According to Mermin et al., 2008, married individuals with more than one partner outside of marriage who never used condoms had high odds of HIV infection. There was no risk at all among married individuals with more than one partner outside of marriage who consistently used condoms. Condom use is further impeded by factors including cultural beliefs, religious beliefs, pregnancy status, fear of rejection, as well as violence from the partner (Smolak, 2010).

Alcohol consumption before sexual intercourse is associated with engaging in transactional or exchange sex, having multiple partners (or a partner with multiple partners), and inconsistent condom use (Tumwesigye et al., 2012; Walusaga, Kyohangirwe, & Wagner, 2012). Given that alcohol consumption impairs a person's sexual decision-making skills, it increases his/her susceptibility to HIV infection.

Male circumcision and HIV infection

Male circumcision (MC), the surgical removal of the intact foreskin from the human penis, has been shown to reduce HIV acquisition by 60% among young adult men in three randomized clinical trials in SSA (Auvert et al., 2005; Bailey et al., 2007; Gray et al., 2007). Furthermore, among males practicing high-risk sexual behaviors in Uganda (e.g. total lifetime number of sex partners (4+), non-use of condoms with non-marital sexual partners, and having non-marital sex), circumcision was found to be protective against HIV infection (Kibira et al., 2014).

Other benefits of MC that specifically benefit women include reduced risk of sexually transmitted infections such as female genital ulcerations, bacterial vaginosis, trichomoniasis, human papillomavirus (HPV) and chlamydia which would otherwise increase the likelihood of contracting HIV (Kibira et al., 2014). However, there is little evidence that MC directly reduces the risk of HIV in women. It only provides a long-term indirect protection to women by reducing the risk of HIV among heterosexual men (Weiss, Hankins, & Dickson, 2009).

Perceived risk of HIV infection

There is a reported correlation between self-perceived risk of getting HIV and HIV infection (Koh & Yong, 2014). Risk perception is mostly influenced by sexual behaviors of an individual or their partner. Individuals who practice high-risk sexual behaviors, e.g., having multiple partners, inconsistent condom use, and transactional sex as well as those individuals who consume alcohol before or during sex are more likely to perceive themselves as high-risk compared to individuals practicing preventive behaviors. Besides sexual behaviors, new HIV prevention and treatment methods including male circumcision, ART use, and use of microbicides and pre-exposure prophylaxis (PrEP) that substantially reduce the risk of HIV infection have been found to be associated with lower perceived risk of HIV infection (Cassell, Halperin, Shelton, & Stanton, 2006; Maughan-Brown & Venkataramani, 2012; Westercamp, Agot, Jaoko, & Bailey, 2014). The low perceived risk may subsequently result in an increase in risky sexual behaviors, also referred to as risk compensation. This risk compensation may lead to new HIV infections (Cassell, Halperin, Shelton, & Stanton, 2006).

While examining whether knowledge of the HIV-protective benefits of MC led to risk compensating behavior among men and women in Cape Town, South Africa, Maughan et al., 2012, found evidence of risk compensation among women but not men. Women informed about the HIV-protective benefits of MC perceived lower HIV risk and were less likely to use condoms even with partners of positive or unknown serostatus whereas informed men perceived slightly higher risk of contracting HIV and were more likely to use condoms at last sex (Maughan-Brown & Venkataramani, 2012). Furthermore, Cederbaum et al., 2014, reported a higher rate of condom use among adolescents with a higher perceived risk of HIV in SSA compared to adolescents with a lower perceived risk. There is therefore a need for assessing and raising awareness of risk in order to encourage behavioral change based on accurate self-risk assessment.

Rationale

An increase in HIV infections in recent years is an indication of changing sexual behaviors. According to Kibira et al., 2014, individuals tend to adjust their behavior in response to the perceived level of risk, usually behaving more cautiously in situations where they feel that the risk level is high and less cautiously where they feel protected. For instance, some individuals perceive MC as a natural condom or a vaccine against HIV thereby engaging in high-risk sexual behaviors. However, the association between self-perceived risk of getting HIV and other risk factors for HIV infection, like male circumcision, and with HIV infection has not been studied extensively. Yet risk reduction goals may be impeded by risk-compensation behaviors (Cederbaum, Gilreath, & Barman-Adhikari, 2014; Eaton & Kalichman, 2007; Maughan-Brown & Venkataramani, 2012). Furthermore, the Roman Catholic Church's opposition to contraceptives (including condoms) use and the government of Uganda's promotion of abstinence-only programs among non-married individuals and monogamy among married individuals may lead to less safe sex. This is a result of censoring and distortion of information regarding condom use. Also, complacency (normalization of HIV/AIDS) and increased access to ARTs reduces people's fears of the virus resulting in increased likelihood of engaging in risky sexual behaviors (J. Cohen & Tate, 2006).

A few studies have explored factors associated with HIV infection using the 2011 UAIS data. According to Kibira et al., 2014, among men practicing risky sexual behaviors in 2011, circumcised men had lower HIV prevalence than uncircumcised men. Chimoyi et al., 2014 reported alcohol use before sexual intercourse, presence of STI and multiple sexual partners as highly associated with HIV infection among young people aged 15-24 years. Both studies, however, analyzed a subsample of the 2011 UAIS data. Kibira et al., 2014, using only a male subsample, studied differences in risky sexual behaviors and HIV prevalence between circumcised and uncircumcised men. Chimoyi et al., 2014 used data on youths aged 15-24 years when researching spatial factors associated with HIV infection since the youths account for 50% of new HIV infections. However, HIV prevalence rates in 2011 were highest among women and individuals in their 30s and 40s of age (MOH et al., 2012). In addition, these studies used the weighted logistic regression model but there was no mention of utilizing all the complex sampling design features (e.g. sampling weights, cluster and strata), evaluation of interaction effects, assessing linearity for continuous variables or assessing final model fit in their papers. Ignoring design effects of complex survey data, evaluation of interactions and not checking all the necessary assumptions of a logistic regression model might result in biased findings.

Furthermore, risk factors for HIV infection in Uganda have been changing over the years. High HIV incidences have shifted from high-risk groups such as commercial sex workers, bar maids, and long-distance truck drivers to those in long term stable-relationships. Highest incidences in 2011 were reported among the widowed followed by the divorced, married/cohabiting and lastly among the never married (MOH et al., 2012; Ntozi, Najjumba, Ahimbisibwe, Ayiga, & Odwee, 2003). Therefore, this study seeks to establish sexual behavior, socio-demographic and biological factors as well as complex interactions among the sexual behavior, socio-demographic and biological factors that were associated with or protective against HIV infection in Uganda in 2011.

The association between any of the sexual behavior, social-demographic or biological factors with HIV infection could be confounded or modified by other factors like complacency, self-perceived risk of HIV, cultural beliefs, contraceptive use, unemployment, access to health centers and HIV prevention and transmission information, gender inequality, and access to main roads. Evaluating the relative importance of each explanatory variable during model building as well as examining the effects of a combination of variables is pertinent to identifying potential confounders and statistical interactions respectively. In order to detect complex higher level (more than 2 variables) interaction effects that are hard to identify with standard regression methods, advanced statistical techniques like classification and regression tree (CART) models can be used.

Public Health Significance

The epidemiology of HIV in Uganda has changed over time. New infections are mostly attributed to risky sexual behaviors and a number of studies have identified factors associated with HIV infection. However, most of these studies have been carried out to evaluate new preventive methods like circumcision or are targeting specific populations like fishermen and young adults aged 15-24 years suspected to contribute most to new HIV infections.

This study therefore sought to identify sexual behavior, socio-demographic, and biological factors as well as association patterns among the identified factors using a nationally representative sample, while taking into account the complex design features and bias due to missing information. Identification of significant factors or combination of factors associated with HIV infection can lead to improved disease control and treatment mechanisms.

Specific Aims

Aim 1: Fit design-based logistic regression models to assess the association between HIV infection and sexual behavior, socio-demographic, and biological factors using data from the 2011 Uganda AIDS Indicator Survey (UAIS).

Hypothesis

i. Circumcised males with a high self-perceived risk of getting HIV are more likely to be HIV positive than uncircumcised males with a low self-perceived risk of getting HIV.

Aim 2: Conduct an exploratory analysis of high-order interactions among identified sexual behavior, socio-demographic, and biological factors associated with HIV infection using weighted classification trees.

METHODS

Study Design

This study is a secondary data analysis of HIV data from the 2011 Uganda AIDS Indicator Survey (UAIS).

The 2011 Uganda AIDS Indicator Survey

The Ministry of Health (MOH), with technical assistance from the Demographic and Health Surveys (DHS) program, conducted the UAIS from February to September 2011. This was a population-based cross-sectional survey designed to obtain national and sub-national HIV prevalence data among adults age 15-59 years and children less than five years. Furthermore, syphilis infection among adults and information about indicators for effective monitoring of national HIV/AIDS programs were obtained (MOH et al., 2012). The study covered over 11,000 households within 79 urban areas and 391 rural areas. Almost 22,000 participants were interviewed. Interviewed women and men were asked to voluntarily give a blood specimen for testing. Upon obtaining consent from parents, a blood specimen was obtained from children under 5 years. The DHS program provides the data for public use by researchers through their website: http://www.dhsprogram.com/. In addition to the survey data, blood test results data are available to researchers by submitting additional requests after agreeing to the terms and conditions to protect respondents' confidentiality. The blood test results include HIV tested with home-based rapid tests (Determine, Statpak and Unigold) and laboratory-based confirmatory HIV polymerase chain reaction (PCR) test for children < 18 months and an enzyme immunoassay (EIA) (Murex and Vironostika Uniform II+O) test for individuals 15-59 years; syphilis tested by Bioline rapid test, and laboratory based RPR and EIA; and CD4 cell counts for HIV-positive adults tested by BD TruCount (MOH et al., 2012).

Data collection

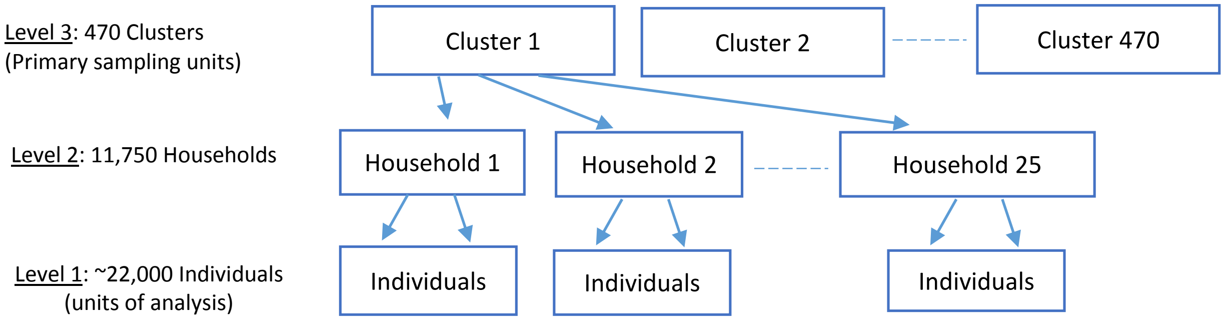
The survey used two questionnaires for data collection: the Household Questionnaire and the Individual Questionnaire. The two questionnaires were translated to six local languages and loaded onto personal digital assistants (PDAs) that were used to conduct the interviews in the field. Furthermore, a specimen collection and field test results form was used by laboratory technicians for capturing the consenting process for blood specimen collection, storage and testing. Collected specimens and the results of the home-based testing during the survey were recorded on the same specimen collection and field test results form (MOH et al., 2012; The DHS Program, 2012). All data were verified for accuracy before entry and uploading on to the DHS program website: http://www.dhsprogram.com/.

Sample design

The survey used a two-stage stratified sample design. Stratification was achieved by separating each region into rural and urban areas. The first stage involved equally selecting a total of 47 clusters from each of the 10 created geographical regions across the country (total of 470 clusters) from a list of enumeration areas (EAs) covered in the 2002 Uganda Population Census. The clusters were selected with probability proportional to their size. Of the 470 selected clusters/EAs, 79 were in urban areas and 391 in rural areas.

The second stage involved systematically selecting a fixed number of 25 households per cluster/EA from a list of EAs provided by the Uganda Bureau of Statistics (UBOS). All eligible women and men age 15-59 years residing at or had visited the household the night before the survey were interviewed and asked to voluntarily give a blood specimen for testing. Since the sample was not allocated in proportion to the size of each region, the UAIS sample was not self-weighting at the national level hence weighting factors were applied to the data to produce nationally representative estimates (Chimoyi & Musenge, 2014; MOH et al., 2012). From the sample design used, Individuals (level 1) were nested within households (level 2) and households were nested in clusters (level 3) that were selected from all regions across the country. The nesting is illustrated in the figure below;

Figure 3: Multilevel structure of the data.



Since the 2011 UAIS sample was a result of a multi-stage stratified design, any statistical methods used for analysis of such data should account for the complex sampling design in order to appropriately estimate standard errors associated with the model parameters. Most standard statistical analysis programs assume observations to be independently and identically distributed. Simple random sample (SRS) designs produce samples that most closely approximate this assumption. In a SRS design, a sample is selected from a population with each individual chosen into the sample entirely by chance. Each member of the population has an equal chance of being included in the sample (Heeringa, West, & Berglund, 2010). In order to increase statistical precision and administrative efficiency, complex sampling designs are used. Complex sampling designs have features including stratification, clustering and weighting and all have an effect (referred to as the design effect) on standard errors of estimates from complex survey data (Heeringa et al., 2010).

Stratification creates non-overlapping, homogenous groupings of population elements or clusters of elements in order to oversample specific subpopulations (e.g., rural areas) to ensure sufficient sample sizes for analysis. Forming strata that are *homogeneous within* and *heterogeneous between* increases sample precision thereby reducing the sampling variance (Heeringa et al., 2010). On the other hand, clustering yields estimates with larger standard errors than those from a SRS of equal size. Characteristics measured on individuals within the same cluster tend to be correlated (nonindependent). For example, individuals living in the same neighborhood may access the same health center, share information about HIV prevention, be of the same socioeconomic status or have the same political attitudes. The amount of statistical information in such clustered samples is less than in an independently selected SRS of the same size resulting in inflated variances (Heeringa et al., 2010). In complex samples combining stratification and clustering in sample selection, there is a net loss of precision (Heeringa et al., 2010).

Furthermore, when probability sampling is used in surveys, inclusion probabilities vary across samples hence weighting is required in order to map the sample back to an unbiased representation of the survey population (Heeringa et al., 2010). In the UAIS, the probability of selection and interview of participants was unequal due to the two-stage stratified sample design used (MOH et al., 2012). Because of the disproportional allocation of samples according to the size of each region and any possible differences in response rate, sampling weights were applied to the data to produce nationally-representative estimates (MOH et al., 2012). In the survey, two main sampling weights were calculated: the household sampling weights and individual sampling weights. Sampling weights were derived as the inverse of the product of the inclusion probabilities, for household and individual respectively, at each stage of sampling. The difference between household and individual sampling weights is due to individual non-response (MOH et al., 2012). Additional sampling weights were created for sample subsets, such as the individual HIV and syphilis results weights for individuals who volunteered to give a blood specimen. Additional weights were only created if there was a differential probability in selecting the subsamples (MOH et al., 2012; The DHS Program, 2012). In our study, the individual HIV and syphilis results weights were used since our analysis is restricted to individuals with HIV results.

Without taking into account sampling weights, clustering and stratification when analyzing data from complex surveys, estimates obtained will be misleading and their variances will be under-estimated. However, a number of statistical methodologies have been modified to account for complex survey designs.

Selection of data sets

Data used in the analysis were obtained with permission from the DHS program website: http://www.dhsprogram.com/. Two data sets were used: 1) the individual dataset which contained all social-demographic, biological and sexual behavior information pertaining to individual participants and 2) the HIV dataset which contained HIV and syphilis laboratory results as well as other results including CD4 cell counts and viral loads for effective monitoring of HIV.

Both data sets were merged using case id as the unique identifier and a subsample created based on inclusion criteria described in the following section.

Inclusion Criteria

Our analysis was restricted to respondents age 15-59 years who: (1) reported ever having sexual intercourse; (2) volunteered to provide a blood specimen for HIV and syphilis testing; and (3) had either an HIV sero-positive or sero-negative result in the HIV dataset.

Outcome Variable

The main outcome variable was HIV infection. This variable was coded as "1" if an individual was HIV sero-positive and "0" if HIV sero-negative.

Explanatory Variables

The explanatory or independent variables used in this study were sexual behavior, social-demographic, and biological variables. The explanatory variables were directly used as shown in Table 1 or they were derived.

Sexual behavior factors included in the analysis were condom use, multiple sex partners, cumulative or point concurrent sexual relationships, age at first sexual intercourse, total lifetime number of sex partners, and transactional/commercial sex. Socio-demographic factors included sex of respondent, age, religion, marital status, education, type of residence, region, away from home for more than a month and wealth index. The biological factors included having an STI and circumcision status. Other variables explored included alcohol consumption during sex, coerced sex, self-perceived risk of getting HIV/AIDS, gender of household head, history of HIV testing, total number of children, number of biological children away from home or died, knowledge of whether male circumcision helps to prevent diseases, and respondents' opinions on whether children 12-14 years should be taught to wait until marriage for sex or taught about condoms.

Several variables from the survey were recategorized or combined. Variables s*ex of the respondent* and *circumcision status* were combined to create a new variable with categories: female, circumcised male and uncircumcised male. For the rest of this thesis, this variable will be referred to as g*ender*. The effect of interest for the first model was created by combining categories of gender with self-perceived risk of getting HIV. The new variable, known as gender and self-perceived risk of getting HIV interaction effect has categories: female-high-risk, female-low-risk, circumcised male-high-risk, circumcised male-low-risk, uncircumcised male-high-risk, and uncircumcised male-low-risk.

Religion was collapsed as follows: Catholic, Anglican/Protestant, Pentecostal, Muslim, other Christians (collapsing: other Christian, SDA, Orthodox, Baptist and Jehovah), and other religions (collapsing: Bahai, Traditional, Hindu, Bisaka group, African religion, Mungu mwama, United faith, Anytime message, and Others). Region of the country was collapsed as follows: Kampala, Central (collapsing: Central 1 and Central 2), Eastern (collapsing: East-Central and Mid-Eastern), Northern (collapsing: North-East, West Nile and Mid-Northern) and Western (collapsing: South-Western and Mid-Western). Marital status was collapsed as follows: never married (never in union), married/cohabiting (collapsing: married and living with partner), widowed, divorced/separated (collapsing: divorced and no longer living together/separated). Age was collapsed as 15-21, 22-39, and 40-59.

Non-consistent condom use in last 12 months preceding the survey referred to when the respondent reported having had sex with a non-marital or non-cohabiting partner or commercial sex worker within the last 12 months and a condom was not used during any of the sexual encounters. Condom use will therefore be coded as ‘always’ if a respondent reported using a condom during all such encounters, ‘not always’ if the condom was inconsistently used and ‘never’ if the condom was not used at all during such sexual encounters. Multiple sexual partners in last 12 months was coded as ‘yes’ when a respondent reported having an on-going sexual relationship with one or more partners other than the spouse or cohabiting partner. Otherwise it was coded as ‘no’. If the timing for such on-going sexual relationships overlapped, a new variable for concurrency sexual relationship was generated and coded as ‘yes’ else as ‘no’. Total lifetime number of sex partners was collapsed as 1, 2-3, and greater than or equal to 4. Biological children away from home was collapsed as 0, 1-2, 3-4 and greater or equal to 5. Alcohol consumption during sex in last 12 months was coded as ‘yes’ (collapsing: respondent drunk only, partner drunk only, both drunk, and neither drunk but consumed alcohol) else ‘no’ for 0 response. Commercial sex in last 12 months was coded ‘yes’ if the individual paid (or was paid) for sex or exchanged any gift items for sex in last 12 months prior to the survey otherwise it was coded as ‘no’.

Having had an STI was coded as ‘yes’ if the participant had a syphilis sero-positive test result or if he/she reported having genital sores or genital discharge in the 12 months prior to the survey. Otherwise it was coded as ‘no’.

Total lifetime number of sex partners and marital status were further collapsed based on the CART model cut-points. Marital status was collapsed as ‘never/married/cohabiting’ and ‘divorced/separated/widowed’ while total lifetime number of sex partners was collapsed as < 3 and 3+.

During data processing, any categories marked as 3 - Refused to answer or 8/98/998 – Don’t know were coded as missing.

Table 1: survey variable description

No.	description	Code
AIDSEX	Sex of respondent	1 – Male; 2 – Female
V012	Respondent's current age	15 – 59 years
V013	Age in 5-year groups	1 = "15-19"; 2 = "20-24"; 3 = "25-29"; 4 = "30-34"; 5 = "35-39"; 6 = "40-44" 7 = "45-49"; 8 = "50-54"; 9 = "55-59"
V152	age of household head	0 - 96 97 = "97+" 98 = "Don't know"
V130	Religion	1-Catholic; 2-Anglican/Protestant; 3-SDA; 4-Orthodox; 5-Pentecostal; 6-Other Christian; 7-Muslim; 8-Bahai; 9-Traditional; 10-Hindu; 11-None; 12-Baptist; 13 - Bisaka group; 14 - African religion; 15 - Jehovah; 16 - Mungu mwama; 17 - United faith; 18 - Anytime message; 96 -Others
V501	Current marital status	0 = "Never in union"; 1 = "Married" 2 = "Living with partner"; 3 = "Widowed" 4 = "Divorced"; 5 = "No longer living together/separated"

V504	currently residing with husband/partner	1 = "Living with her" 2 = "Staying elsewhere"
V505	number of other wives	0 = "No other wives" 1 - 97 98 = "Don't know"
V106	Highest educational level	0 = "No education"; 1 = "Primary" 2 = "Secondary"; 3 = "Higher"
V025	Type of place of residence	1 – Rural; 2 - Urban
V024	Region	1- Central 1; 2- Central 2; 3- Kampala; 4- East Central; 5- Mid Eastern; 6- North East; 7- West Nile; 8- Mid Northern; 9- South Western; 10- Mid Western
V168	Away for more than one month in last 12 months	0 – No; 1 – Yes
V167	Number of trips in last 12 months	0 - 95
V190	Wealth index	1 - Poorest; 2 - Poorer; 3 - Middle; 4 - Richer; 5 -Richest
V201	Total children ever born	0 - 95
V202	Sons at home	0 - 95
V203	Daughters at home	0 - 95
V204	Sons elsewhere	0 - 95
V205	Daughters elsewhere	0 - 95
V525	Age at first sex	0 = "Not had sex" 1-95 years; 96 = "At first union" 97 = "Inconsistent" 98 = "Don't know"
V531	age at first sex (imputed)	0 = "Not had sex" 1-96 years; 97 = "Inconsistent" 98 = "Don't know"
V820	Condom used at first sex	0 – No; 1 – Yes; 8 – Don’t know
V536	recent sexual activity	0 = "Never had sex" 1 = "Active in last 4 weeks" 2 = "Not active in last 4 weeks - postpartum abstinence" 3 = "Not active in last 4 weeks - not postpartum abstinence"

V537	Months of abstinence	60 = "60+" 96 = "Before last birth" 97 = "Inconsistent" 98 = "Don't know"
S441B	In total, with how many different people have you had sexual intercourse in the last 3 months?	0-95; 98-Don’t know
S442A	With how many of this people is your sexual relationship continuing?	0-95; 98-Don’t know
V766B	Number of sex partners, including spouse, in last 12 months	0-95; 98-Don’t know
V852A	How long ago first had sex with most recent partner	101 = "Days: 1" 199 = "Days: number missing" 201 = "Weeks: 1" 299 = "Weeks: number missing" 301 = "Months: 1" 399 = "Months: number missing" 401 = "Years: 1" 499 = "Years: number missing"
V852B	How long ago first had sex with 2nd most recent partner	
V852C	How long ago first had sex with 3rd most recent partner	
V853A	Times in last 12 months had sex with most recent partner	0-95 95 = "95+" 98 = "Don't know"
V853B	Times in last 12 months had sex with 2nd most recent partner	
V853C	Times in last 12 months had sex with 3rd most recent partner	
V854A	Point concurrent sexual partners	0 = "No"; 1 = "Yes"
V854B	Cumulative concurrent sexual partners	0 = "No"; 1 = "Yes"
V767A	Relationship with most recent sex partner	1 = "Spouse" 2 = "Boyfriend not living with respondent" 3 = "Other friend" 4 = "Casual acquaintance" 5 = "Relative" 6 = "Commercial sex worker" 7 = "Live-in partner" 96 = "Other"
V767B	Relationship with 2nd to most recent sex partner	
V767C	Relationship with 3rd to most recent sex partner	
S434BA	Is your sexual relationship with this person ongoing? (last partner)	0 = "No" 1 = "Yes"
S434BB	Is your sexual relationship with this person ongoing? (second to last partner)	
S434BC	Is your sexual relationship with this person ongoing? (third to last partner)	

V832B	Time since last sex with 2nd to most recent partner	100 = "<1 day ago" 101 = "Days: 1" 199 = "Days: number missing" 200 = "Weeks: 0" 201 = "Weeks: 1" 299 = "Weeks: number missing" 300 = "Months: 0" 301 = "Months: 1" 399 = "Months: number missing" 401 = "Years: 1" 995 = "Within last 4 weeks" 996 = "Before last birth" 997 = "Inconsistent" 998 = "Don't know"
V832C	Time since last sex with 3rd to most recent partner	
V769	can get a condom	0 – No; 1 – Yes; 8 – Don't know
V761	Condom used during last sex with most recent partner	0 = "No" 1 = "Yes" 8 = "Don't know"
V761B	Condom used during last sex with 2nd to most recent partner	
V761C	Condom used during last sex with 3rd to most recent partner	
V833A	Used condom every time had sex with most recent partner in last 12 months	0 = "No" 1 = "Yes"
V833B	Used condom every time had sex with 2nd to most recent partner in last 12 months	
V833C	Used condom every time had sex with 3rd to most recent partner in last 12 months	
V834A	Age of most recent partner	1-95 years; 95 = "95+" 98 = "Don't know"
V834B	Age of 2nd to most recent partner	
V834C	Age of 3rd to most recent partner	
V835A	Alcohol consumption at last sex with most recent partner	0 = "No" 1 = "Respondent drunk only" 2 = "Partner drunk only" 3 = "Both drunk" 4 = "Neither drunk but consumed alcohol"
V835B	Alcohol consumption at last sex with 2nd to most recent partner	
V835C	Alcohol consumption at last sex with 3rd to most recent partner	
V836	Total lifetime number of sex partners	0-95; 95 = "95+"; 98 = "Don't know"
SV793	Paid for sex in last 12 months	0 – No; 1 – Yes;
SV793A	Condom used last time paid for sex in last 12 months	0 – No; 1 – Yes;
SV793B	Condom used every time paid for sex in last 12 months	0 – No; 1 – Yes; 8 – Don't know
S448B	Did you ever give sex in exchange for goods or services?	0 – No; 1 – Yes; 3- Refused to answer; 8 – Don't know
S448C	Did this happen in the last 12 months?	0 – No; 1 – Yes; 3- Refused to answer;
		8 – Don't know

S448D	The last time this happened, was a condom used?	0 – No; 1 – Yes; 3- Refused to answer; 8 – Don’t know
S448E	Ever give sex in exchange of money?	0 – No; 1 – Yes; 3- Refused to answer; 8 – Don’t know
S448F	Did this happen in the last 12 months?	0 – No; 1 – Yes; 3- Refused to answer; 8 – Don’t know
S448G	The last time this happened, was a condom used?	0 – No; 1 – Yes; 3- Refused to answer; 8 – Don’t know
S453	Have you ever run short of condom?	0 = "No"; 1 = "Yes" 3 = "Never used condom" 8 = "Don't know, unsure"
V763A	Had any sexually transmitted infection (STI) in last 12 months?	0 – No; 1 – Yes; 8 – Don’t know
V763B	Had genital sore/ulcer in last 12 months?	0 – No; 1 – Yes; 8 – Don’t know
	Had genital discharge in last 12 months	0 – No; 1 – Yes; 8 – Don’t know
SLSYPH	Syphilis laboratory results	0 – Negative/Nonreactive; 1 – Positive/Reactive
S608	Respondent circumcised	0 – No; 1 - Yes
S609	Age at circumcision	0 – 95 years; 98 – Don’t know; Record 00 if less than 1 year.
S508B	Does medical male circumcision help to prevent HIV infection?	0 = No; 1 = Yes; 8 = Don't know
S454	Ever forced to have sex against your will?	0 – No; 1 – Yes; 3- Refused to answer; 8 – Don’t know
S455	Did this happen in the last 12 months?	0 – No; 1 – Yes; 3- Refused to answer; 8 – Don’t know
S456	Ever coerced to have sex without the use of physical force?	0 – No; 1 – Yes; 3- Refused to answer; 8 – Don’t know
S457	Did this happen in the last 12 months?	0 – No; 1 – Yes; 3- Refused to answer; 8 – Don’t know
S556	Are you equally careful about avoiding HIV?	1- More careful; 2 - Less careful; 3 - Equally careful; 8 - Don't Know
S557	In your opinion, are the chances that you can get HIV high or low?	1- High; 2 – Low; 8 - Don't Know
S558	If you would get HIV from whom would you most likely get it?	1 - Spouse; 2 - Boy/Girlfriend; 3 – Stranger; 4 - Commercial sex partner; 8 - Don't know
S508B	Does medical male circumcision help to prevent HIV infection?	0 = No; 1 = Yes; 8 = Don't know
V780	Children should be taught about condoms to avoid AIDS	0 = No; 1 = Yes; 8 = Don't know"
V849	Should children 12-14 be taught to wait for sex until marriage?	0 = No; 1 = Yes; 8 = Don't know/depends"

Complex sample design Variables

No.	Complex design variables description	Code
V022	sample strata for sampling errors	1 - Central 1 (urban); 2 - Central 1 (rural) 3 - Central 2 (urban); 4 - Central 2 (rural) 5 - Kampala (urban); 6 - Kampala (rural) 7 - East Central (urban); 8 - East Central (rural) 9 - Mid Eastern (urban); 10 - Mid Eastern (rural) 11 - North East (urban); 12 - North East (rural) 13 - West Nile (urban); 14 - West Nile (rural) 15 - Mid Northern (urban); 16 - Mid Northern (rural) 17 - South Western (urban); 18 - South Western (rural) 19 - Mid Western (urban); 20 - Mid Western (rural)
HIVCLUST	Cluster	
HIV05	Sample weight	

Statistical Analysis

Univariate analysis of each independent variable was carried out using descriptive statistics. Weighted counts and percentages were estimated for categorical variables using the SAS SURVEYFREQ procedure while weighted means and standard error of the means were estimated for continuous variables using the SAS SURVEYMEANS procedures taking into account all the design features of complex surveys.

For aim 1, fitting design-based logistic regression models to assess the association of HIV infection with sexual behavior, socio-demographic, and biological factors, we hypothesized (null hypothesis) that there is no association between gender and self-perceived risk of getting HIV interaction effect with HIV infection. A weighted multivariable logistic regression model was built following steps for purposeful variable selection suggested by Hosmer and Lemeshow to select the final best fitting yet biologically plausible model (Hosmer & Lemeshow, 2013). For aim 2, we sought to assess the association between complex higher level interaction effects as identified by classification trees with HIV infection using weighted multivariable logistic regression models.

Logistic regression model

Many of response variables from biomedical research studies or from surveys are binary, such as having an HIV positive or negative test result. The logistic regression model is widely used to find the best fitting and yet biologically plausible model to describe the relationship between a binary outcome variable (Y) and a set of explanatory variables (Agresti, 2007; Heeringa et al., 2010; Hosmer & Lemeshow, 2013). A multiple logistic regression model is of the form;

$$\text{g}[\pi(x)] = \text{Logit } [\pi(x)] = log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$\text{and} \qquad \pi(x)] = \left[\frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}}\right]$$

Where π(**x**) denotes the probability of having an HIV positive result (P [Y =1]) with $x_i(i = 1,2,\ldots,p)$ explanatory variables. The regression coefficient, β_i, refers to the effect of x_i on the logit (log odds), adjusted for other x's in the model. Exponentiating β_i gives the adjusted odds ratio which represents the *multiplicative* impact of a one-unit increase in x_i on the odds of the outcome variable being equal to 1, holding all other x's in the model fixed/constant.

When data is from a SRS, the maximum likelihood estimation (MLE) method is used to estimate the unknown parameters β_i. However, when a complex sample design has been used to collect the data, the straightforward application of the MLE method is no longer possible due to violation of the independence of observations assumption due to stratification and clustering as well as the unequal probability of sample selection (Heeringa et al., 2010). Instead, an iterative estimation procedure such as the Newton–Raphson or Fisher Scoring algorithm is used to determine the values of estimated coefficients that maximize the following weighted pseudo-likelihood (PL) function (Binder, 1983; Chambers & Skinner, 2003; Heeringa et al., 2010):

$$\text{PL}(B|x) = \prod_{i=1}^{n}\{\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}\}^{w_i}$$

Besides the pseudo-maximum likelihood estimation (PMLE) technique, Binder et al., 1983, further proposed a sandwich-type variance estimator, a method applying a multivariate version of Taylor's series linearization (TSL) to estimate the sample variance and covariance for the parameter estimates (Binder, 1983; Heeringa et al., 2010). The sandwich-type variance estimator is of the form;

$$var(\hat{B}) = (J^{-1})var[S(\hat{B})](J^{-1})$$

Where J is a matrix of second derivatives with respect to $\hat{B}_j$ of the pseudo log-likelihood for the data [PL($B|x$)] and $S(\hat{B})$ is the variance – covariance matrix for the sample totals of the weighted score function for individual observations used to fit the model. It is the solution of the vector of estimation equation below;

$S(\hat{B}) = \sum_h \sum_\alpha \sum_i w_{h\alpha i}(y_{h\alpha i} - \pi_{h\alpha i}(B))x'_{h\alpha i} = 0$, where h- is a stratum index; α-cluster index within stratum h; i- Individual observation index within cluster α; $w_{h\alpha i}$- sampling weight for observation *i,* $\pi_{h\alpha i}(B)$ is the Pr(Y=1) and $x'_{h\alpha i}$ is a column vector of the p+1 (p predictors plus the intercept) design matrix elements for case i = $[1x_{1,h\alpha i}, \ldots, x_{p,h\alpha i}]'$.

The PMLE approach for parameter estimation together with a linearized estimator of the variance-covariance matrix for the parameter estimates, taking into account the complex sample design features, are the standard method for logistic regression modeling in all of the major software systems that support analysis of complex sample survey data (Heeringa et al., 2010; Skinner & Vallet, 2010). In SAS, the design-based logistic regression procedure is SURVEYLOGISTIC (SAS Institute, 2011) and it takes into account sampling weights, clusters and strata. *Design-based* refers to when statistical analyses of survey data take the survey design (stratification and clustering) and statistical weights into consideration while when such features are ignored and the data is assumed to be collected from a simple random sample, the resulting statistical analyses are termed *model-based* (Hosmer & Lemeshow, 2013).

Model selection

Steps for purposeful selection of variables proposed by Hosmer and Lemeshow were followed when building a design-based logistic regression model (Hosmer & Lemeshow, 2013). Purposeful selection of variables is a more efficient selection method when modelling risk factors rather than fitting a prediction model because it gives the analyst control over the model building process and important confounders can be retained in the model (Bursac, Gauss, Williams, & Hosmer, 2008; Hosmer & Lemeshow, 2013).

Variable Selection Steps

First step: a weighted univariate logistic model was used for examining the association between individual independent variables and HIV infection. Crude or unadjusted odds ratios and modified Wald chi-square p-values were reported for each variable. Any covariate with a bivariate association with HIV infection at significance p-value less than 0.2 was considered important and was included in the multivariable model in step 2.

Second step: an initial weighted multivariable logistic regression model was fit and contribution to the model of each selected independent variable was evaluated. The model with all the important independent variables from step 1 (full model) was compared to a model fitted with one of the independent variables removed at a time (reduced model). The relationship between any of the remaining covariates (or the effect of interest in case of the first model) with HIV infection was confounded by the excluded variable if the percentage change in the estimates for any of the remaining covariates (or the effect of interest) in the full model relative to the reduced model was greater than 10%. The excluded variable was then kept in the model else it was dropped. The process continued until all the independent variables were evaluated. A new multivariable model with selected relevant independent variables was fit. Initially eliminated or not previously added but potential confounding variables at the first step were evaluated by adding each variable to the new multivariable model one at a time and refitting the model while computing the percentage change in estimates. The variable was a confounder and was added back to the model if it resulted in a >10% change in the parameter estimates.

Third step: the assumption of linearity in the logit for any significant continuous independent variables from the second step was checked by plotting the logit against each of the continuous independent variables. If the linearity assumption was violated, the variable was transformed by categorizing it based on different points on the weighted empirical logit plot. The model after this step was the main effects model.

Fourth step and for aim number two: interaction effects were evaluated. In the first model, statistically significant interactions between any two explanatory variables in the main effects model as well as the interaction effect of interest, whether significant or not, were retained in the final model. For the next models, classification trees were used to identify complex interactions among the explanatory variables with a bivariate association with HIV infection (from step 1). The weighted tree was built and pruned to obtain an optimal tree, and the variables leading to the terminal nodes, also referred to as leaves, represented interaction patterns. The detected potential interaction patterns were separately added back to the weighted logistic regression models for further evaluation for statistical significance. Model hierarchy was conserved for the models. The models after this step were the final models.

Interaction (effect-measure modification)

Statistical interaction, defined as deviation from some specified model, occurs when the effect of one factor on the outcome is different within strata of another defined factor. When building statistical models, interaction terms/combination of variables (e.g., circumcision × religion) are considered as potential predictors and if any interaction term is significant, $p < 0.05$, there is interaction on the multiplicative scale hence the interaction term should be retained in the model. In multiplicative models like logistic regression, the coefficient for the interaction term is either negative (submultiplicative) or positive (supramultiplicative) or 0 when the multiplicative assumption is a perfect fit to the data and the interaction term should be dropped (Institute of Medicine (US), 2006; K. J. Rothman, Greenland, & Lash, 2008; J. Rothman K., 2012; Szklo & Nieto, 2000).

One method to detect interactions involves adding all possible n-way interactions using multiplicative interaction terms but for aim 2 of this study, decision trees (classification tree specifically) were used to identify complex linear as well as non-linear variable interactions that were later incorporated into the design-based logistic regression models. Classification trees have an added advantage of uncovering categorical aspects of continuous variables by splitting them at different points at different times during the development of the tree. They are also easy to explain, interpret and can be displayed graphically (James, Witten, Hastie, & Tibshirani, 2013).

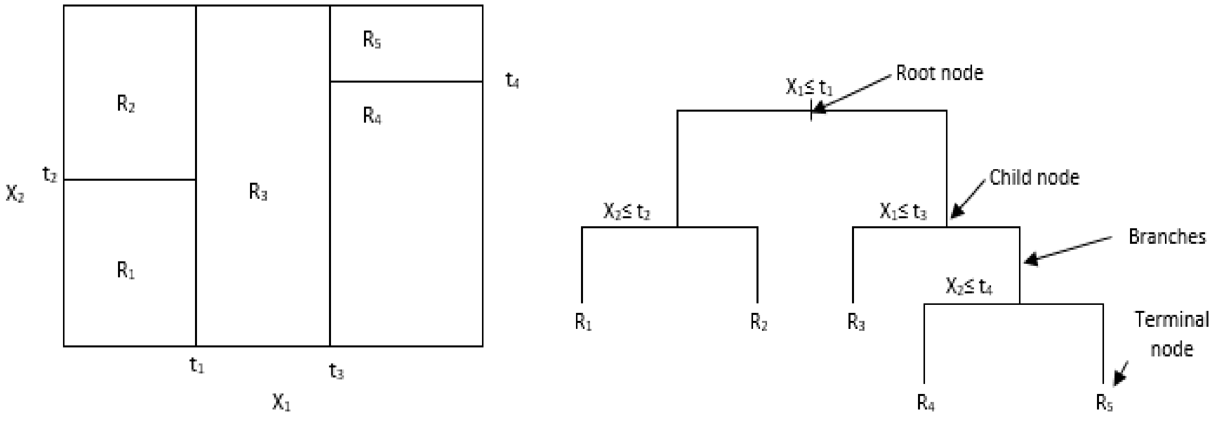
Classification and Regression Trees (CART) analysis

Classification and Regression Trees (CART) are machine-learning algorithms or methods for developing models to predict or explore relationships between independent variables and the response variable (Piper, Loh, Smith, Japuntich, & Baker, 2011; Speiser, Lee, Karvellas, & US Acute Liver Failure Study Group, 2015; Wei-Yin, 2011). CART methods are widely used in medicine, psychology, biology, and computer science. Classification trees are used for categorical outcome variables taking on a finite number of unordered values whereas regression trees are for continuous outcome variables or outcomes with ordered discrete values. CART is a simple yet powerful nonparametric decision tree method. Because of its nonparametric nature, CART has few statistical assumptions and does not require calculations (e.g. transformations) compared to logistic regression. Also, CART allows inclusion of variables with complex interactions, effectively handles independent variables with missing values and outliers, accepts user-specified misclassification costs and class (outcome) probabilities, and offers reliable results that are easy to interpret and simple to use (Speiser et al., 2015; Wei-Yin, 2011).

Tree development

Prior to developing CART models, the data set was split into the training data set (60%) for constructing the models and a testing dataset (40%) for assessing the constructed models with each set having the same proportion of the outcome of interest (HIV infections).

Figure 4: Recursive binary splitting with corresponding decision tree.



Developing a classification tree utilizes a top-down approach, referred to as recursive binary splitting, to grow a large tree on the training data by splitting the predictor space (a set of possible values for the main effects, Xs) into *J* distinct and non-overlapping regions or subgroups R_j (James et al., 2013). Each split (of each X variable at a time) is indicated via two new branches further down on the tree. The split, S, is based on the cut-point (e.g., 1-Male; 2-Female; Missing for gender covariate or Age>=25; Age<25 for age covariate) of the choice of the explanatory variable that results into the tree with the lowest classification error rate when estimating proportions for the class of interest (HIV infections). Classification error rate is defined as the fraction of the training observations in a given region that do not belong to the most common class within the region.

Classification error rate, E = 1-max$_k(\hat{p}_{rk})$

where $\hat{p}_{rk}$ is the proportion of training observations in the rth region that are from the kth class. However, classification error is not sufficiently sensitive for tree-growing hence the Gini index, G = $\sum_{k=1}^{K} \hat{p}_{rk}(1-\hat{p}_{rk})$ or the cross-entropy, D= -$\sum_{k=1}^{K} \hat{p}_{rk}\log(\hat{p}_{rk})$ are preferable when evaluating the quality of a particular split (James et al., 2013). The Gini index, also referred to as a measure of node purity, as well as the cross-entropy will take on a small value (smaller the better) if the *r*th node is pure ($\hat{p}_{rk}$ is close to 0 or 1). Therefore, for each X, a node is split after exhaustively searching over all X and each split that minimizes the total impurity of its two child nodes. Each X variable with *m* distinct values has up to (2^{m-1}-1) splits if unordered and (m-1) for ordered variables (Wei-Yin, 2011).

The repeat process of looking for the best explanatory variable and best cut-point in order to split the data within the regions further continues until a termination criterion is reached. The termination criteria might be reaching a prespecified minimum number of observations in a region (e.g. 10 observations per region or 10% of the training set) or until the relative decrease in impurity is below a prespecified threshold. At this point, terminal nodes or 'leaves' are obtained and the predicted class (e.g. HIV Positive/Negative) along with the class proportions and number of patients are reported. The large tree is then pruned by removing unimportant variable splits in order to obtain a subtree with the least classification error rate using the testing/validation set.

The subset of predictor space (e.g., married × female × age<25 years) corresponding to a given terminal node of the final subtree define a potential interaction term to be assessed in the weighted logistic regression model.

When building the classification tree, sampling weights were incorporated by specifying the weighting variable in the software in order to account for the unequal probability of selection.

Goodness-of-fit and regression diagnostics

After fitting the final models to data, the goodness-of-fit (gof) summary statistics to assess model adequacy in describing the observed outcome as well as logistic regression diagnostics to identify influential observations or patterns of covariates were examined. These methodologies have been developed for SRS assuming independence (zero covariance) between observations and have been included in standard logistic regression programs in most statistical software (Heeringa et al., 2010; Hosmer & Lemeshow, 2013). Archer et al., 2006, extended the standard Hosmer and Lemeshow gof test in order to account for sampling weights, clustering and stratification. They derived a linearized estimator of the covariance matrix to use for the calculation of the modified Wald test for assessing model fit for complex sampling studies. This method, implemented in STATA, has the limitation of not identifying observations where the model does not fit well when the null hypothesis of a good fit is rejected (Archer, Lemeshow, & Hosmer, 2006).

Since the modified gof test is not yet incorporated into SAS, the design-based logistic regression model with the SAS SURVEYLOGISTIC procedure taking into account sampling weights, cluster and strata was used for estimating the coefficients, while for examining model fit and identifying influential observations, the models were refit using only the sampling weights in the standard (model-based) logistic regression program using the SAS LOGISTIC procedure (Heeringa et al., 2010; Hosmer & Lemeshow, 2013). Even though the standard program does not correctly reflect the variances and covariances of the estimates given the complex sample design, the weighted estimates of parameters and predicted probabilities will be identical and serious failures in the model for covariate patterns should be identifiable (Heeringa et al., 2010).

The findings of where the model does not fit well or observations or covariate patterns that exert extreme influence on the overall model fit were implemented in the final design-based analysis in order to devise ways of improving the model fit (Hosmer & Lemeshow, 2013).

The Hosmer–Lemeshow (H-L) gof test statistics and the area under the receiver operating characteristic (ROC) curve, a plot of sensitivity against 1- specificity, were used to assess model gof and discrimination ability. If the model fits the data well enough, the chi-square p-value associated with the H-L gof test statistic should be greater than 0.05 while an area under the ROC curve between 0.7 and 1 signifies great discriminatory power. Influential observations were examined using diagnostic plots and diagnostic statistics. Common influential diagnostics for each covariate patterns include the change in the parameter estimate (DFBETAS or $\Delta\beta_j$), change in Pearson chi-square (DIFCHISQ or ΔX_j^2) and change in deviance (DIFDEV or ΔG_j^2). DFBETAS are used to assess the effect of an individual observation on each estimated parameter of the fitted model. The DFBETAS diagnostic for an observation is the standardized difference in the parameter estimate due to deleting the observation. Any outlying or extreme point may indicate bad data that might require investigating. DIFCHISQ and DIFDEV on the other hand identify observations or covariate patterns that are poorly fit resulting in the disagreement between the data and the predicted values of the fitted model.

The design-based logistic regression model was fit with and without the outlying/influential observations identified in the diagnostic plots and the parameter estimates and model fit statistics, including the Akaike information criterion (AIC) and area under the ROC curve, were compared. The outlying/influential observations were assumed to be erroneous and excluded from the analysis since the estimates changed significantly and the model without the outlying/influential observations was a better fit. The final model after this stage with adjusted odds ratios was reported.

In order to evaluate statistical significance of one or more parameters, the adjusted Wald-type tests were used since complex sample designs invalidate the key assumptions that underlie the *F*-tests or likelihood ratio tests. Furthermore, estimate of the odds ratio (OR) obtained by exponentiating the estimated coefficients and their associated confidence intervals (CIs) were used to quantify the magnitude of the effect of each individual or interaction of explanatory variables (Heeringa et al., 2010).

Missing data

Complex surveys collect large amounts of data and it is not uncommon for one or more explanatory variables to contain missing values. Common reasons for missing data include refusal to respond especially with sensitive questions, the design of the questionnaire where some questions are not applicable to a group of people (e.g. circumcision status only asked of men), respondents failure to recall an event, insufficient knowledge (don't know), and interviewer error. Weighting adjustments for nonresponse only compensate for potential bias due to completely missing data but not for item-missing data for completed cases (Heeringa et al., 2010). Since standard statistical modelling requires a complete data set in which all the response and explanatory variables were recorded (complete case analysis), observations with a missing variable value (incomplete cases) are excluded from the analysis (listwise deletion). Therefore, high rates of missing data can result into biased and misleading conclusions in data analysis.

A number of techniques have been developed to handle data missingness. Among these include; complete case analysis (CC), available case analysis (pairwise deletion), last observation carried forward (LOCF)/last observation carried backward (LOCB), maximum-likelihood using the EM algorithm and simple and multiple imputation methods (Molenberghs & Kenward, 2007; Rubin, 1987). Although these methods are easily implemented, they require assumptions about the data that rarely hold in practice (Pigott, 2001). From the stand point of being able to effectively and practically address item-missing data in a survey dataset, a missing at random (MAR) mechanism is the more reasonable assumption (Heeringa et al, 2010). MAR means that missingness depends only on the observed variables. Multiple imputation (MI), a technique that replaces each missing value with two or more acceptable values representing a distribution of possibilities, was used (Heeringa et al., 2010; Rubin, 1987). This is due to the ability of the multiple imputation process to incorporate statistically sophisticated techniques and draw from distributions of "plausible" values while accounting for the variability introduced by the process of selecting a value for the missing data point (Rubin, 1987).

Prior to imputing missing values, item-missing data percentages and patterns for the variables included in the first main effects model (model 1) were examined and reported. Missing data patterns provide information about the amount and structure of missing data thereby categorizing missingness as arbitrary or a more specialized pattern such as monotone missing data. A data set is said to have an arbitrary missing pattern when missingness is interspersed among full data values while monotone missing pattern is when if a variable X_j is missing for individual i, all subsequent variables X_k, k>j, are all missing for the individual i.

Besides the missing data pattern, the imputation method to use greatly depends on the type of variables with missing data as well as variables to use while developing the imputation models. The default method of Markov Chain Monte Carlo (MCMC) is most appropriate for continuous arbitrary missing data while for categorical variables, a monotone missing data pattern offers the widest range of imputation method options, including logistic regression (for binary or ordinal response variable) and discriminant function (for binary or nominal response variable) methods. For continuous variables with monotone missing patterns, a parametric method that assumes multivariate normality (such as regression) or a nonparametric method that uses propensity scores should be used (Rubin, 1987; SAS Institute, 2011; Schafer, 1997).

Multiple Imputation steps

In order to assess the effects of missing data on the estimates, main effects in model 1 with item-missing data were multiply imputed, interaction effects re-created as combination variables and a new model fit on the full dataset. Estimates and their 95% CIs as well as the difference between the upper and lower 95% CIs (also referred to as precision) in the new model and in the model excluding item-missing data (complete case analysis/listwise deletion approach) were compared. The item-missing data was multiply imputed according to the steps below;

First step: missing values in these variables were imputed *5* times (in order to achieve a relative efficiency of at least 0.9 (Rubin, 1987) to create one dataset with *5* complete concatenated datasets along with an automatic variable (_imputation_), which served as an identifier for analysis of each imputation. During imputation, multivariate normal (MVN) distribution was assumed. Since the MVN is a purely continuous distribution, any non-continuous variables were imputed as continuous and at the end of the process, their values were rounded off and bounded so that they match the format and range of the observed discrete values (Schafer, 1997). If both categorical and continuous variables contained missing values, a two-step process to first impute data by MCMC method in order to produce a monotone missing pattern and then employ imputation by logistic regression for imputation of the subsequent missing data (with categorical variables) would be employed. Interaction terms in the first final model were recreated at this step.

Second step: The design-based logistic regression model was fitted on each of the concatenated imputed datasets by specifying the imputation identifier variable as a domain variable (usually as a by statement in standard statistical methods). Parameter estimates and covariance matrices (for computing standard errors) for each of the models were saved to an external dataset for use in the next step.

Third step: parameter estimates were combined into a single set of statistics by taking their arithmetic mean over the *m* imputations, $\bar{\beta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\beta}_i$. Averaging the estimates dampens the variation thus increasing efficiency and decreases sampling variation. Combining results from the analyses accounts for the variance adjustments due to complex sampling design and multiple imputation variability in order to fully correct the variance estimates thereby giving valid statistical inferences. The adjusted variances are as below;

1) The within imputation variance (V_W) is the arithmetic mean of the sampling variances ($\hat{V}_i$) from each of the *m* imputed datasets, $V_W = \frac{1}{m}\sum_{i=1}^{m}\hat{V}_i$.
2) The between imputation variance (V_B) is the measure of variability in the parameter estimates across the *m* imputed datasets, $V_B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\beta}_i - \bar{\beta})^2$.
3) The total variance (V_T) is computed as; $V_T = V_W + V_B + \frac{V_B}{m}$ and the standard error as; $\sqrt{V_T}$ (Heeringa et al., 2010; Rubin, 1987; von Hippel, 2009).

The SAS MI procedure was used to examine the missing data patterns as well as to create multiply imputed datasets on which design-based logistic regression models were fit for each imputation set. The SAS MIANALYZE procedure was used to combine results of the analyses of the imputations in order to generate valid statistical inferences.

Statistical Power

At the design stage of the 2011 UAIS, it was estimated that 26,870 adults would be required for the survey based on the 2004-05 Uganda adults' HIV prevalence rate of 6.4%, a 10% relative error, a design effect of 1.69 and a response rate of 92%. However, the survey covered 12,153 women and 9,588 men age 15-59 (MOH et al., 2012).

During model fitting, at least 10 outcome events (HIV positive cases) per covariate were required in order to detect an actual effect. Therefore, considering our sample size of at least 10,000 observations and an HIV prevalence of 6.4%, we anticipated that we would be able to incorporate utmost 64 variables into the logistic regression model and thus we would have enough power to test the 3-way interaction terms, adjusted for other sexual behavior, socio-demographic, and biological variables.

Software

Data management and statistical analysis was done using Statistical Analysis System (SAS) software version 9.3 (SAS Institute, 2011). Salford Predictive Modeler® (SPM) with CART® 7.0, Salford Systems, San Diego, CA was used to generate a weighted classification tree for the detection of association patterns. A Type-I error level of 0.05 was used.

Human Subjects, Animal Subjects, or Safety Considerations

The student completed an Ethics in Public Health & Healthcare course and the CITI Protection of Research Subjects/Human Subjects training course. This research only involved secondary de-identified data from the 2011 Uganda AIDs and Indicator survey which was freely accessible and available online. An Internal Review Board (IRB) with exemption status was approved by the UTHSC Committee for Protection of Human Subjects (protocol # HSC-SPH-15-0571, see Appendix A).

RESULTS

Exploratory Data analysis

The original survey data set included information on approximately 22,000 participants. After restricting our analysis to respondents age 15-59 years who: (1) reported ever having sexual intercourse; (2) volunteered to provide a blood specimen for HIV and syphilis testing; and (3) have either an HIV sero-positive or sero-negative result in the HIV dataset, the analytic sample included 18,395 respondents (not adjusting for sampling weights): 7857 men and 10,538 women.

Table 2 presents weighted counts and percentages of selected characteristics of respondents included in our study. It was observed that the majority (56.7%) of respondents were females and the overall mean (SE) age was 32.8 (0.11). Majority of the respondents were aged 22-39 years (55.02%), had a primary education (58.07%), married or cohabiting (71.37%) and were catholic (41.45%).

Thirty-one percent of all respondents reported having had sex with a non-marital/cohabiting partner or with a commercial sex worker within the last 12 months. Consistent condom use (always used condoms) among these respondents was at 17.86% while 44% reported to never having used a condom. Among the males, 27.96% were circumcised and the median (Interquartile range (IQR)) age at circumcision was 11.6 (0.25-55). Of all the respondents, 31.63% had a high self-perceived risk of getting HIV. Females (37.23%) had the highest percentage of respondents with a high self-perceived risk of getting HIV followed by circumcised males (26.67%) and uncircumcised males (23.91%) had the least number of respondents with a high self-perceived risk of getting HIV.

Overall, 21.07% (3879/18410) reported alcohol use during sex in the last 12 months and 80.4% reported living in the rural area, 29.1% (5357/18410) reported having had an STI, 3.04% (559/18410) reported engaging in commercial or exchange sex and 23.6% (4351/18410) reported having more than one sex partners during the same time frame. Furthermore, the mean (SE) and median (IQR) total number of lifetime sex partners was 4.63 (0.096) and 2.03 (1.00-3.78) respectively while 44.12% (8088/18334) reported having had their first sexual intercourse when aged less than 17 years.

Table 2: Background characteristics of the study population

Variable	Categories	Total (%)
SEX OF RESPONDENT		
	Male	7968 (43.28%)
	Female	10441 (56.72%)
AGE (YEARS)	Mean (SE)	32.8 (0.11)
AGE GROUP (YEARS)		
	15-21	3151 (17.12%)
	22–39	10129 (55.02%)
	40–59	5129 (27.86%)
RELIGION		
	Catholic	7618 (41.45%)
	Anglican/Protestant	6253 (34.03%)
	Muslim	2360 (12.84%)
	Other Christians	718 (3.91%)
	Other religions	93 (0.51%)
	Pentecostal	1334 (7.26%)
MARITAL STATUS		
	Never married	2853 (15.50%)
	Married/cohabiting	13139 (71.37%)
	Divorced/separated	1650 (8.96%)
	Widowed	766 (4.17%)
EDUCATION		
	Higher	1213 (6.59%)
	Secondary	4041 (21.95%)
	Primary	10691 (58.07%)
	No education	2463 (13.38%)
TYPE OF PLACE OF RESIDENCE		
	Rural	14805 (80.42%)
	Urban	3604 (19.58%)
GEOGRAPHICAL LOCATION		
	Western	4593 (24.95%)
	Central	4030 (21.89%)
	Eastern	3853 (20.93%)
	Kampala	1314 (7.14%)
	Northern	4619 (25.09%)
WEALTH INDEX		
	Poorest	3301 (17.93%)
	Poorer	3506 (19.04%)
	Middle	3457 (18.78%)
	Richer	3643 (19.79%)
	Richest	4500 (24.45%)
CONDOM USE IN PAST 12 MONTHS		
	Always	1019 (17.86%)
	Not Always	2174 (38.10%)
	Never	2513 (44.04%)
SELF-PERCEIVED RISK OF GETTING HIV		
	Low	10571 (68.37%)
	High	4890 (31.63%)
MALE RESPONDENT CIRCUMCISED?		
	No	5740 (72.04%)
	Yes	2227 (27.96%)

HIV prevalence

From table 3, it is observed that the overall prevalence of HIV in this study population was 8.15% (1500/18410) with females (9.08%) having a higher HIV prevalence than males (6.93%). Similarly, the prevalence of HIV was highest among respondents who were widowed (23.84%) or divorced/separated (15.94%), lived in an urban area (10.17%), consumed alcohol during sex in last 12 months (10.13%), and engaged in commercial or exchange sex in the same time frame (16.66%). HIV prevalence among respondents who were previously HIV negative or those that did not know their HIV sero-status by the time of the survey was 4.9% (869/17779). The prevalence of HIV increased with increasing total lifetime number of sex partners however the prevalence was almost the same for respondents who reported having multiple sex partners (8.78%) and those who did not report having multiple sex partners (7.96%) in the past 12 months. Furthermore, high HIV rates were observed among respondents who reported not consistently using (11.42%) or never (8.8%) used condoms when compared to those who always (6.53%) used condoms when having sex with non-spousal/cohabiting partners in the last 12 months. In addition, HIV rates were high among respondents who had first sexual intercourse before their 17th birthday (8.94%), had a high self-perceived risk of getting HIV (6.37%), had an STI in the past 12 months (12.73%) and among non-circumcised males when compared to circumcised males (7.75% vs 4.84%, *p*<0.0001).

From table 4, it is observed that more respondents with a low compared to those with a high self-perceived risk of getting HIV were less likely to consistently use condoms when having sex with a non-spousal/cohabiting partner in the last 12 months. Uncircumcised males with a high self-perceived risk of getting HIV when compared to those with a low self-perceived risk of getting HIV were more likely to engage in commercial or exchange sex. In contrast, engaging in commercial or exchange sex for money or gifts in the past 12 months was approximately the same across the two risk groups for females and circumcised males. It is further observed that more males, both circumcised and uncircumcised, with a low self-perceived risk of getting HIV than males with a high self-perceived risk of getting HIV were more likely to use alcohol during sex, have an STI, have more total lifetime number of sex partners and have multiple sex partners in the past 12 months. On the contrast, there was no profound difference across the two risk groups for females hence evidence of risk compensation mostly on the part of males. However, a big difference in HIV infections between the risk groups was observed among the uncircumcised males compared with circumcised males.

Table 3: Subject characteristics by HIV results

	HIV Negative (N=16909)	HIV Positive (N=1500)	Overall (N=18410)	Crude OR [95% CI]	P-value
SEX OF RESPONDENT					<.0001
Male	7416 (93.07%)	552 (6.93%)	7968	1	
Female	9493 (90.92%)	948 (9.08%)	10441	1.34 (1.196 - 1.502)	
AGE (YEARS)					
Mean (SE)	32.61 (0.113)	34.91 (0.308)	32.80 (0.107)		
AGE GROUP (YEARS)					<.0001
15-21	3023 (95.94%)	127 (4.06%)	3151	1	
22–39	9231 (91.14%)	897 (8.86%)	10129	2.30 (1.843 - 2.859)	
40–59	4654 (90.73%)	475 (9.27%)	5129	2.41 (1.916 - 3.038)	
RELIGION					0.058
Total	n = 16879	n = 1498	n = 18378		
Catholic	6958 (91.34%)	659 (8.66%)	7618	1	
Anglican/Protestant	5729 (91.62%)	524 (8.38%)	6253	0.97 (0.826 - 1.127)	
Muslim	2213 (93.79%)	146 (6.21%)	2360	0.70 (0.557 - 0.876)	
Other Christians	661 (92.02%)	57 (7.98%)	718	0.91 (0.661 - 1.266)	
Other religions	83 (89.62%)	9 (10.38%)	93	1.22 (0.508 - 2.939)	
Pentecostal	1232 (92.41%)	101 (7.59%)	1334	0.87 (0.654 - 1.148)	
MARITAL STATUS					<.0001
Never married	2741 (96.06%)	112 (3.94%)	2853	1	
Married/cohabiting	12197 (92.83%)	942 (7.17%)	13139	1.88 (1.517 - 2.342)	
Divorced/separated	1387 (84.06%)	262 (15.94%)	1650	4.62 (3.612 - 5.920)	
Widowed	584 (76.16%)	182 (23.84%)	766	7.63 (5.846 - 9.969)	
MARITAL STATUS (CART dichotomization)					<.0001
Never/Married/Cohabiting	14938 (93.40%)	1054 (6.60%)	15993	1	
Divorced/Separated/Widow	1971 (81.56%)	445 (18.44%)	2416	3.20 (2.760 – 3.716)	
EDUCATION					<.0001
Higher	1155 (95.24%)	57 (4.76%)	1213	1	
Secondary	3754 (92.88%)	287 (7.12%)	4041	1.53 (1.095 - 2.149)	
Primary	9747 (91.17%)	943 (8.83%)	10691	1.94 (1.429 - 2.624)	
No education	2252 (91.42%)	211 (8.58%)	2463	1.88 (1.338 - 2.634)	
TYPE OF PLACE OF RESIDENCE					<.0001
Rural	13671 (92.34%)	1133 (7.66%)	14805	1	
Urban	3237 (89.83%)	366 (10.17%)	3604	1.37 (1.150 - 1.621)	

GEOGRAPHICAL LOCATION					<.0001
Western	4178 (90.97%)	414 (9.03%)	4593	1	
Central	3604 (89.44%)	425 (10.56%)	4030	1.19 (0.973 - 1.453)	
Eastern	3641 (94.50%)	211 (5.50%)	3853	0.59 (0.453 - 0.759)	
Kampala	1198 (91.19%)	115 (8.81%)	1314	0.97 (0.724 - 1.307)	
Northern	4286 (92.79%)	332 (7.21%)	4619	0.78 (0.626 - 0.978)	
WEALTH INDEX					0.003
Poorest	3076 (93.16%)	225 (6.84%)	3301	1	
Poorer	3254 (92.83%)	251 (7.17%)	3506	1.05 (0.828 - 1.339)	
Middle	3200 (92.56%)	257 (7.44%)	3457	1.09 (0.855 - 1.401)	
Richer	3300 (90.59%)	343 (9.41%)	3643	1.42 (1.121 - 1.788)	
Richest	4077 (90.60%)	423 (9.40%)	4500	1.41 (1.127 - 1.772)	
CONDOM USE IN PAST 12 MONTHS					<.0001
Total	n = 5170	n = 536	n = 5706		
Always	952 (93.47%)	66 (6.53%)	1019	1	
Not Always	1925 (88.58%)	248 (11.42%)	2174	1.84 (1.370 - 2.482)	
Never	2292 (91.20%)	221 (8.80%)	2513	1.38 (0.999 - 1.908)	
AGE AT FIRST SEXUAL INTERCOURSE (YEARS)					0.0034
Total	n = 16843	n = 1490	n = 18334		
Mean (SE)	17.21 (0.035)	17.0 (0.089)	17.19 (0.034)		
<17 Years	7366 (91.06%)	722 (8.94%)	8088	1.21 (1.067 - 1.374)	
>= 17 Years	9477 (92.50%)	768 (7.50%)	10245	1	
TOTAL LIFETIME NUMBER OF SEX PARTNERS					<.0001
Total	n = 16417	n = 1428	n = 17846		
1	4486 (95.43%)	215 (4.57%)	4701		
2-3	6714 (92.13%)	573 (7.87%)	7287	1.78 (1.462 - 2.172)	
4+	5217 (89.07%)	640 (10.93%)	5857	2.56 (2.109 - 3.109)	
TOTAL LIFETIME NUMBER OF SEX PARTNERS (CART dichotomization)					<.0001
< 3	8328 (94.32%)	501 (5.68%)	8830	1	
3+	8088 (89.72%)	927 (10.28%)	9016	1.90 (1.673 – 2.168)	
MULTIPLE SEX PARTNERS IN PAST 12 MONTHS?					0.096
Total	n = 16909	n = 1500	n = 18410		
No	12940 (92.04%)	1118 (7.96%)	14059	1	
Yes	3969 (91.22%)	382 (8.78%)	4351	1.11 (0.981 - 1.264)	
CONCURRENT SEXUAL RELATIONSHIP					
Point					0.087
Total	n = 1925	n = 188	n = 2113		
No	1430 (90.43%)	151 (9.57%)	1581	1	
Yes	494 (93.05%)	36 (6.95%)	531	0.71 (0.475 - 1.051)	
Cumulative					0.004
Total	n = 1925	n = 188	n = 2113		
No	395 (87.12%)	58 (12.88%)	453	1	
Yes	1530 (92.18%)	129 (7.82%)	1659	0.57 (0.394 - 0.836)	

ALCOHOL USE AT LAST SEX IN 12 MONTHS?						<.0001
	No	13423 (92.38%)	1107 (7.62%)	14530	1	
	Yes	3486 (89.87%)	392 (10.13%)	3879	1.37 (1.170 - 1.594)	
HAD AN STI IN PAST 12 MONTHS?						<.0001
	No	12234 (93.73%)	818 (6.27%)	13052	1	
	Yes	4675 (87.27%)	682 (12.73%)	5357	2.18 (1.923 - 2.474)	
HAD COMMERCIAL SEX IN PAST 12 MONTHS?						<.0001
	No	16443 (92.12%)	1407 (7.88%)	17850	1	
	Yes	466 (83.34%)	93 (16.66%)	559	2.34 (1.785 - 3.056)	
SELF-PERCEIVED RISK OF GETTING HIV						<.0001
	Total	n = 14728	n = 733	n = 15462		
	Low	10149 (96.01%)	422 (3.99%)	10571	1	
	High	4578 (93.63%)	311 (6.37%)	4890	1.63 (1.351 - 1.976)	
MALE RESPONDENT CIRCUMCISED?						<.0001
	Total	n = 7416	n = 552	n = 7968		
	No	5295 (92.25%)	444 (7.75%)	5740	1	
	Yes	2120 (95.16%)	107 (4.84%)	2227	0.61 (0.457 - 0.802)	

Table 4: Selected factors against gender and self-perceived risk of HIV infection

Risk factor	**Female (N= 8554)**		**Male**			
			Circumcised (N = 1950)		**Uncircumcised (N= 4956)**	
	Perceived Risk		**Perceived Risk**		**Perceived Risk**	
	High, n (%)	**Low, n (%)**	**High, n (%)**	**Low, n (%)**	**High, n (%)**	**Low, n (%)**
Total, *N = 15462*	3185 (37.23%)	5370 (62.77%)	520 (26.67%)	1431 (73.33%)	1185 (23.91%)	3772 (76.09)
HIV INFECTION						
Positive	186 (44.40%)	233 (55.60%)	28 (46.15%)	33 (53.85%)	96 (38.24%)	155 (61.76%)
Negative	2998 (36.86%)	5136 (63.14%)	491 (26.01%)	1397 (73.99%)	1088 (23.14%)	3615 (76.86%)
AGE (YEARS)						
15-21	618 (36.81%)	1060 (63.19%)	80 (22.93%)	269 (77.07%)	169 (23.78%)	544 (76.22%)
22–39	1937 (40.87%)	2803 (59.13%)	322 (30.01%)	751 (69.99%)	676 (25.01%)	2029 (74.99%)
40–59	629 (29.47%)	1505 (70.53%)	117 (22.30%)	410 (77.70%)	338 (22.03%)	1197 (77.97%)
EDUCATION						
No education	492 (33.17%)	991 (66.83%)	30 (25.76%)	87 (74.24%)	80 (23.41%)	262 (76.59%)
Primary	1982 (39.53%)	3033 (60.47%)	279 (27.90%)	722 (72.10%)	708 (24.74%)	2156 (75.26%)
Secondary	573 (35.70%)	1031 (64.30%)	161 (25.82%)	462 (74.18%)	297 (22.88%)	1001 (77.12%)
Higher	137 (30.47%)	312 (69.53%)	48 (23.61%)	158 (76.39%)	98 (22.00%)	350 (78.00%)
MARITAL STATUS						
Never married	358 (34.35%)	685 (65.65%)	108 (22.55%)	373 (77.45%)	243 (24.26%)	761 (75.74%)
Married/cohabiting	2437 (39.76%)	3692 (60.24%)	369 (27.74%)	962 (72.26%)	842 (23.36%)	2765 (76.64%)
Widowed	103 (20.65%)	396 (79.35%)	2 (14.34%)	12 (85.66%)	12 (29.80%)	30 (70.20%)
Divorced/separated	286 (32.45%)	595 (67.55%)	39 (32.50%)	82 (67.50%)	85 (28.55%)	214 (71.45%)

CONDOM USE IN PAST 12 MONTHS						
n =	n = 747	n = 1179	n = 304	n = 643	n = 562	n = 1367
Always	89 (31.53%)	194 (68.47%)	49 (25.32%)	146 (74.68%)	103 (24.40%)	319 (75.60%)
Not Always	292 (38.31%)	471 (61.69%)	112 (35.75%)	202 (64.25%)	245 (31.03%)	544 (68.97%)
Never	364 (41.56%)	513 (58.44%)	142 (32.57%)	294 (67.43%)	214 (29.84%)	503 (70.16%)
ALCOHOL USE AT LAST SEX IN PAST 12 MONTHS?						
No	2479 (36.08%)	4392 (63.92%)	407 (24.87%)	1229 (75.13%)	835 (22.08%)	2949 (77.92%)
Yes	705 (41.91%)	977 (58.09%)	112 (35.94%)	201 (64.06%)	349 (29.83%)	821 (70.17%)
HAD AN STI IN PAST 12 MONTHS?						
No	1846 (33.53%)	3661 (66.47%)	384 (24.20%)	1205 (75.80%)	861 (21.66%)	3115 (78.34%)
Yes	1338 (43.92%)	1708 (56.08%)	134 (37.51%)	224 (62.49%)	323 (33.06%)	655 (66.94%)
HAD COMMERCIAL SEX IN PAST 12 MONTHS?						
No	3050 (36.82%)	5234 (63.18%)	486 (25.82%)	1397 (74.18%)	1112 (22.97%)	3729 (77.03%)
Yes	134 (49.77%)	135 (50.23%)	33 (50.48%)	32 (49.52%)	73 (63.45%)	42 (36.55%)
TOTAL LIFETIME NUMBER OF SEX PARTNERS						
Total	n = 3167	n = 5343	n = 483	n = 1342	n = 1118	n = 3595
1	1060 (33.60%)	2095 (66.40%)	22 (11.99%)	166 (88.01%)	99 (16.14%)	517 (83.86%)
2-3	1543 (37.81%)	2538 (62.19%)	116 (23.35%)	380 (76.65%)	328 (20.55%)	1268 (79.45%)
4+	563 (44.28%)	709 (55.72%)	344 (30.23%)	795 (69.77%)	690 (27.62%)	1810 (72.38%)
MULTIPLE SEX PARTNERS IN PAST 12 MONTHS?						
No	2693 (36.55%)	4675 (63.45%)	239 (21.51%)	875 (78.49%)	685 (20.80%)	2610 (79.20%)
Yes	491 (41.46%)	694 (58.54%)	280 (33.52%)	555 (66.48%)	499 (30.09%)	1160 (69.91%)

Univariate analysis

Table 3 also shows unadjusted (crude) odds ratios and associated 95% confidence intervals as well as p-values depicting the magnitude and significance of associations between single independent variables with HIV infection. Significant independent variables are pertinent for development of multivariate models and for building classification trees.

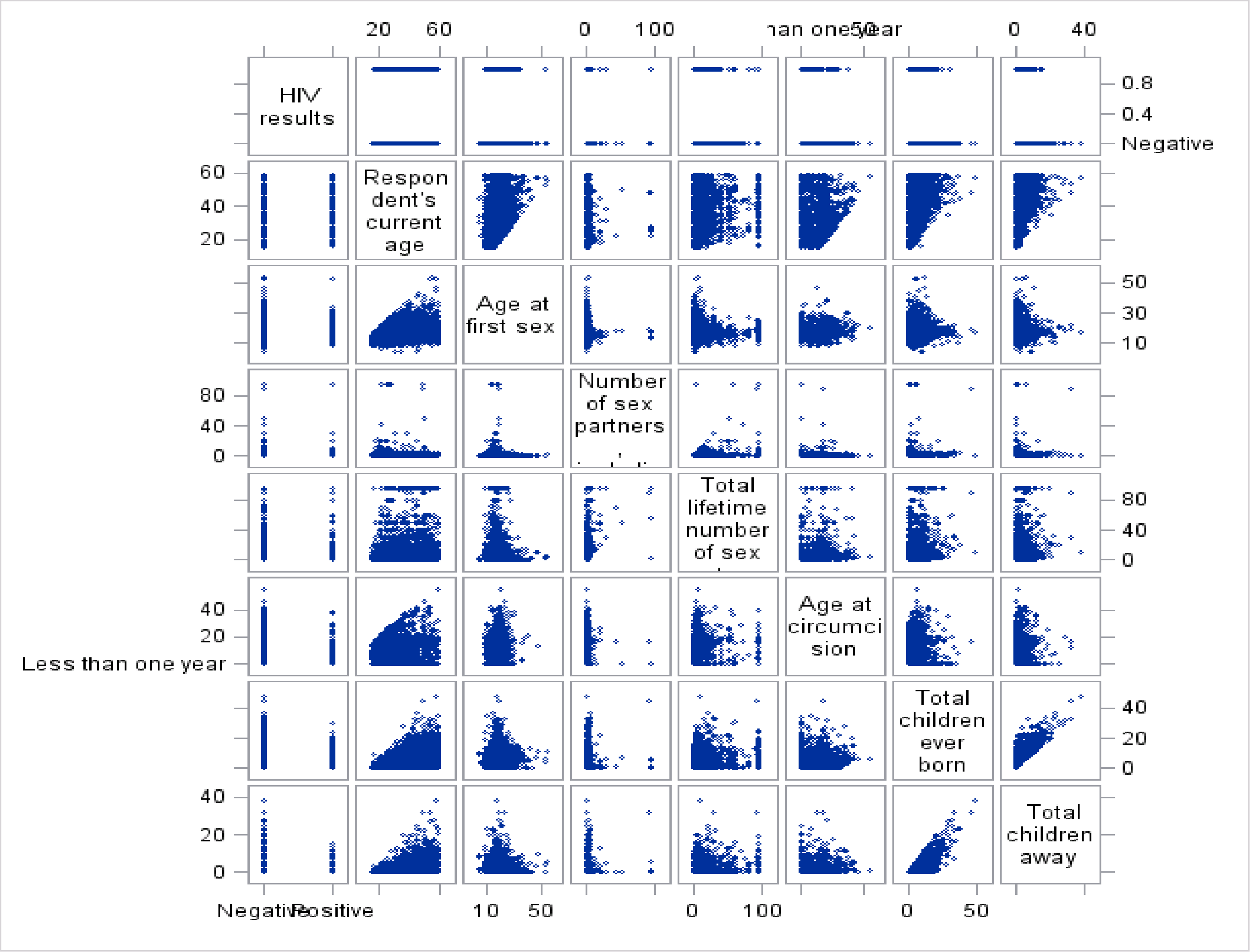
From table 3, HIV infection was significantly associated with sex of respondent, self-perceived risk of getting HIV and male circumcision status. The odds of HIV infection were 1.34 times higher among females compared with males (OR =1.34, 95% CI: 1.196 - 1.502). Compared to uncircumcised males, circumcised males had 39% lower odds of HIV infection (OR = 0.61, 95% CI: 0.457 – 0.802). Among respondents with a high self-perceived risk of getting HIV, the odds of HIV infection were 1.63 times higher when compared with those with a low self-perceived risk of getting HIV (OR = 1.63, 95% CI: 1.351 - 1.976).

Other factors that were significantly associated with high (OR>1.7) HIV infections included: being in age groups 22-39 (OR = 2.30, 95% CI: 1.84 - 2.86) or 40-59 (OR = 2.41, 95% CI: 1.92 – 3.04) years when compared with 15 – 21 years; widowed (OR = 7.63, 95% CI: 5.85 – 9.97) or divorced/separated (OR = 4.62, 95% CI: 3.61 – 5.92) or married/cohabiting (OR = 1.88, 95% CI: 1.52 – 2.34) compared with never married; having a primary (OR = 1.94, 95% CI: 1.43 – 2.62) or no education (OR = 1.88, 95% CI: 1.34 – 2.63) compared with a higher education; not always using condoms (OR = 1.84, 95% CI: 1.37 – 2.48) compared with respondents who always used condoms; having 2-3 (OR = 1.78, 95% CI: 1.46 – 2.17) or 4 or more (OR = 2.56, 95% CI: 2.11 –3.11) total lifetime number of sex partners compared to having only one lifetime sex partner at the time of the survey; having an STI (OR = 2.18, 95% CI: 1.92 – 2.47); and engaging in commercial or exchange sex for money or gifts (OR = 2.34, 95% CI: 1.79 –3.06). In contrast, factors (besides being circumcised) that were significantly associated with low HIV infection included: being Muslim compared to being catholic (OR = 0.70, 95% CI: 0.56 – 0.88); living in the eastern (OR = 0.59, 95% CI: 0.45 – 0.76) or northern (OR = 0.78, 95% CI: 0.63 – 0.98) parts of the country compared with the western.

Having multiple partners in the past 12 months was not found to be associated with HIV infection (OR = 1.11, 95% CI: 0.98 – 1.26).

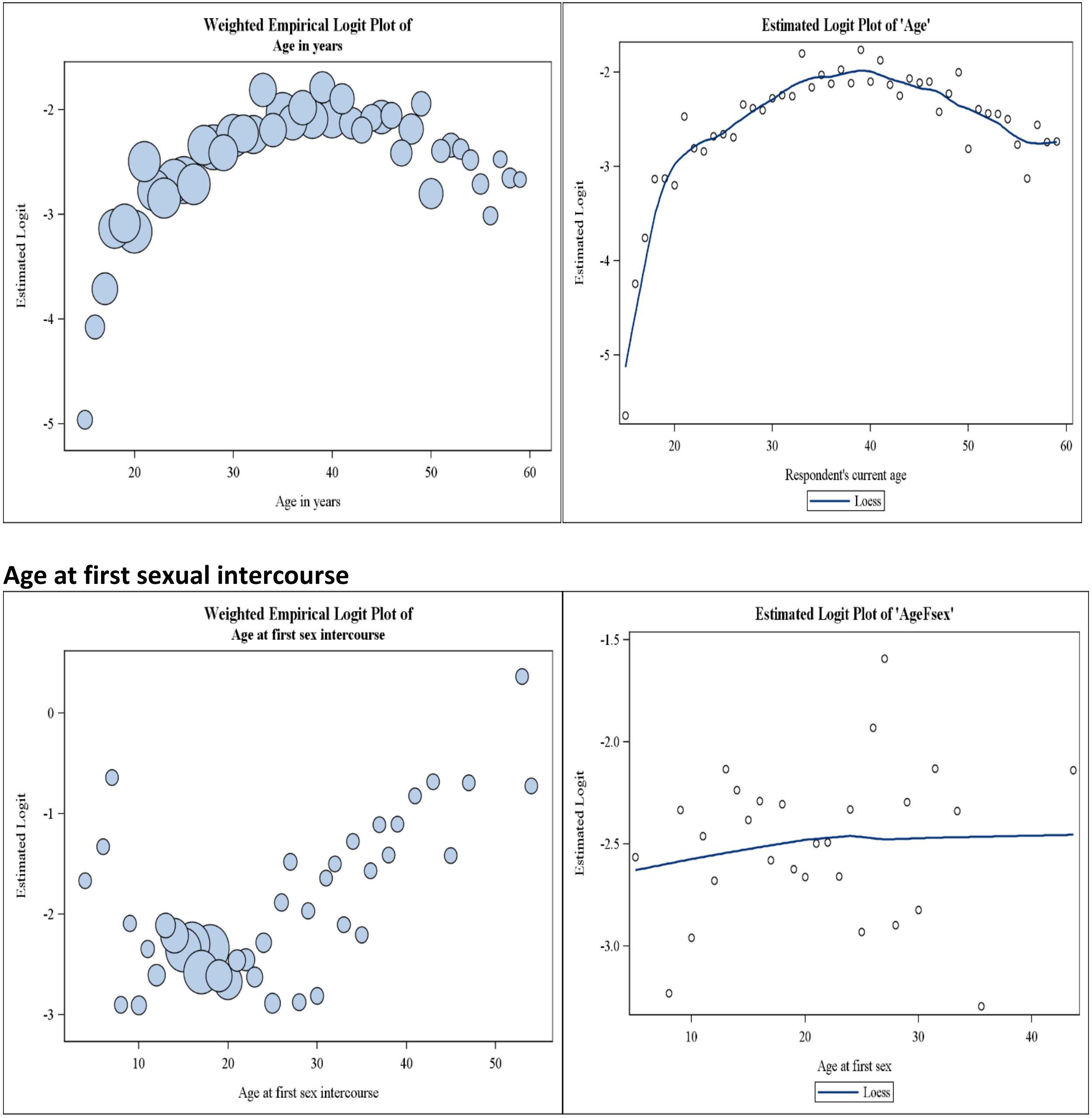
Model Assumptions

Figure 5: Scatter plot matrix to assess correlation coefficients among variables



A Pearson correlation analysis revealed that total number of children ever born was significantly positively correlated with age (Rho = 0.73, *p* <0.0001) and with the number of biological children born away from home (Rho = 0.78, *p* <0.0001). Total number of children was therefore excluded from the analysis in order to avoid likely collinearity and convergence problems during model building. Furthermore, continuous variables, age and age at first sex, in the main effects model were evaluated for linearity in the logit using weighted empirical logit plots.

Figure 6: Weighted empirical logit plots for assessing linearity for age and age at first sexual intercourse in years



Age of respondent

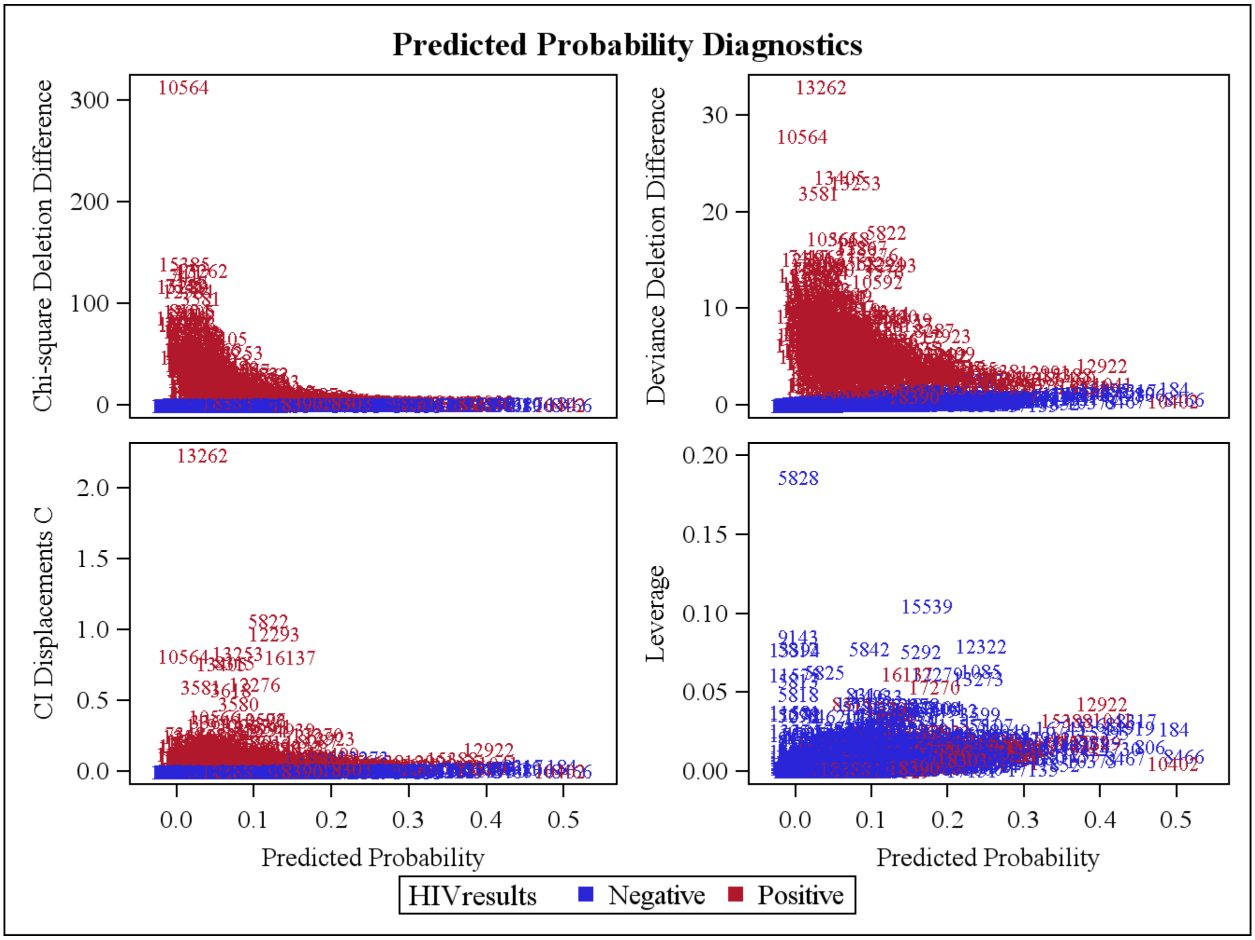
Age at first sexual intercourse

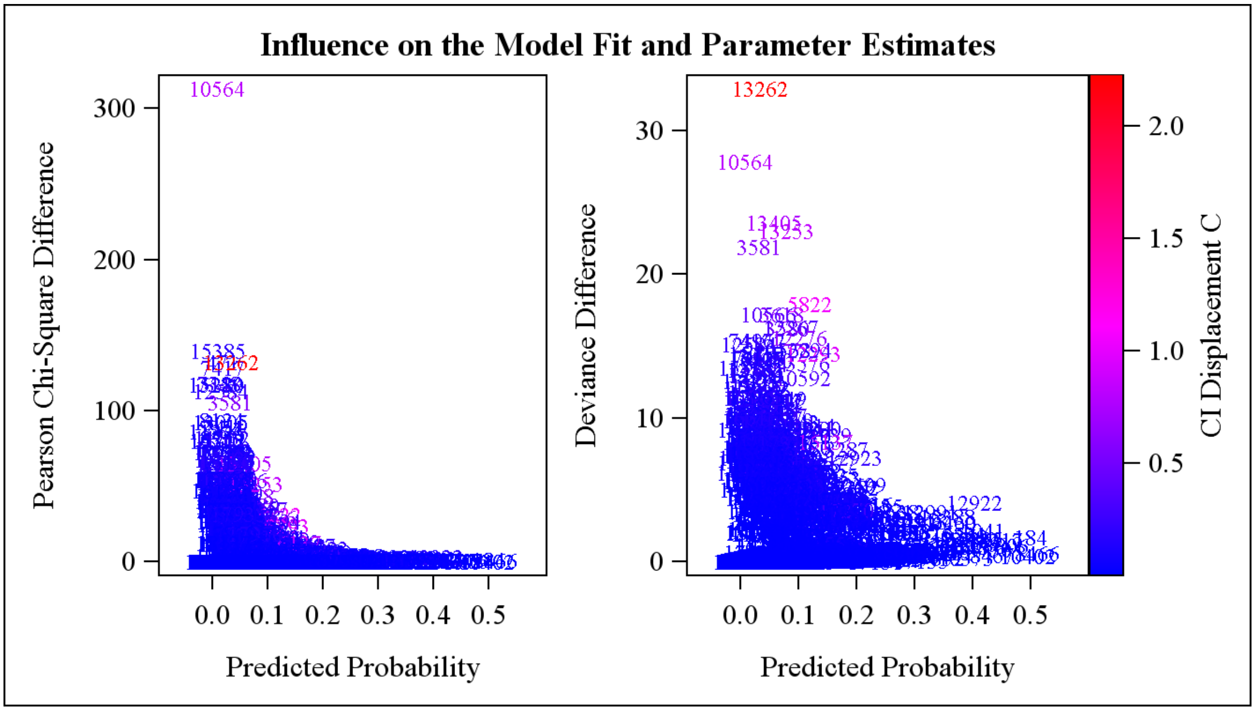
The weighted empirical logit plots with a loess curve revealed a fairly linear relationship for age at first sexual intercourse and a curvilinear relationship for age with HIV infection. Therefore, age was categorized into 15-21, 22-39 and 40-59 years while age at first sexual intercourse treated as a linear independent variable. The bubbles added to the scatter plots represent the population units at a specific age and age at first sexual intercourse, respectively.

Multivariable analysis

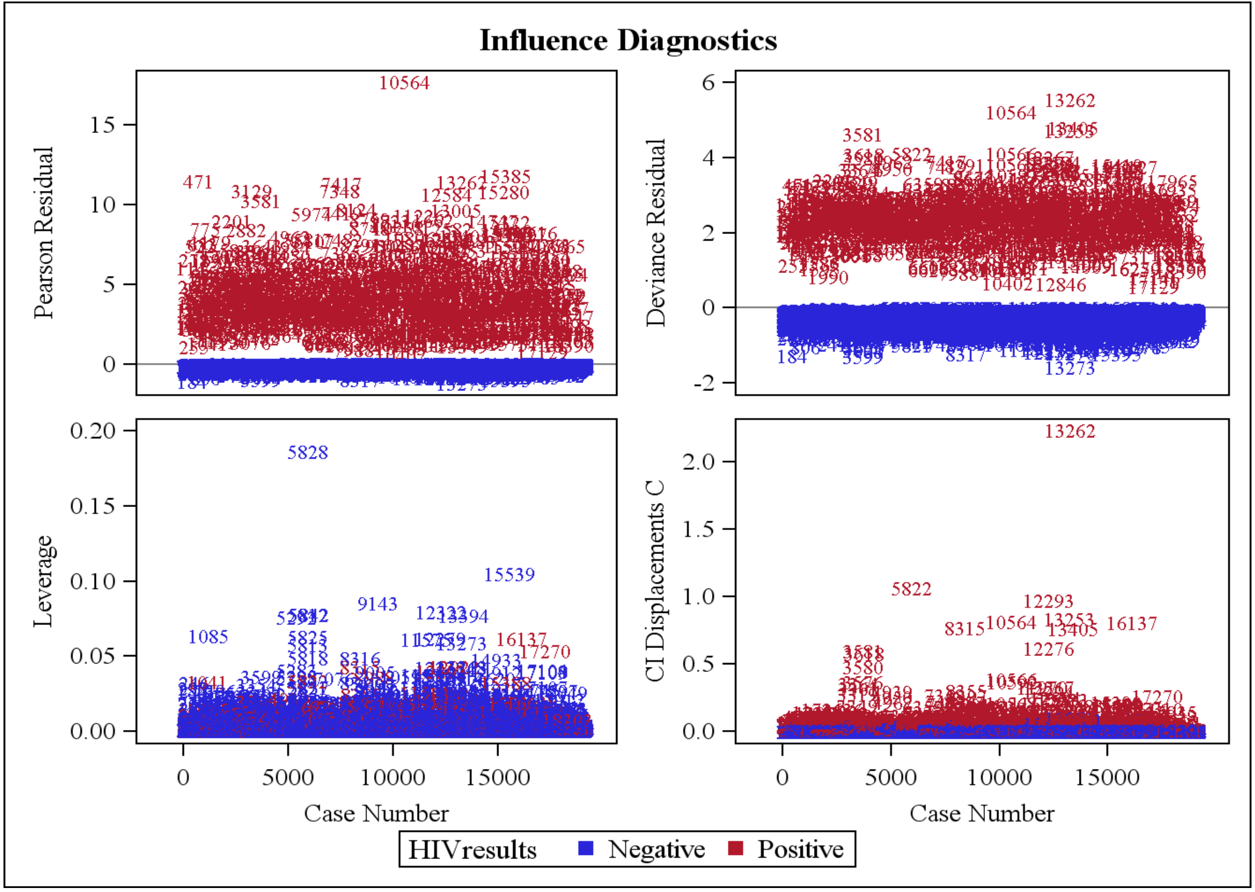
Additional analysis was carried to fit a final best fitting model and assess its goodness of fit.

Figure 7: Diagnostic plots for the weighted multivariable logistic regression model

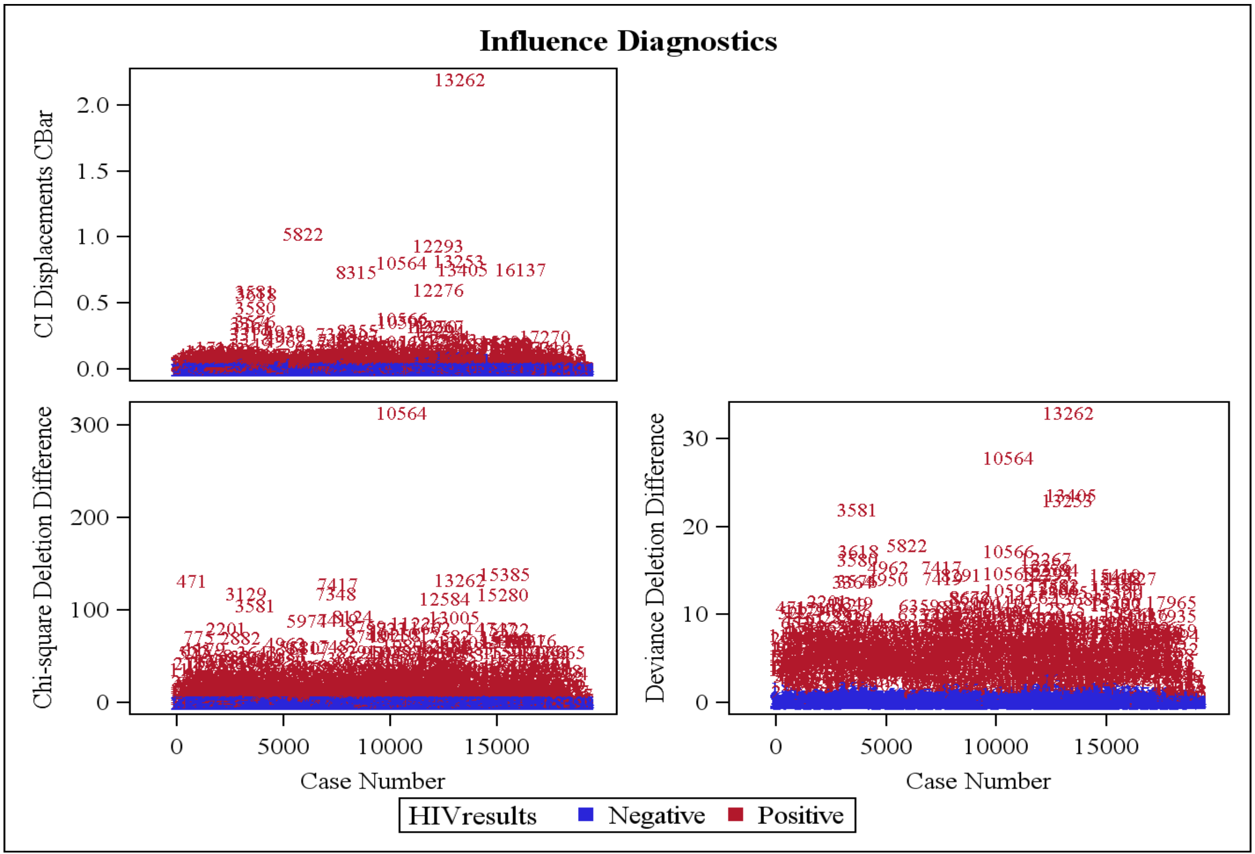




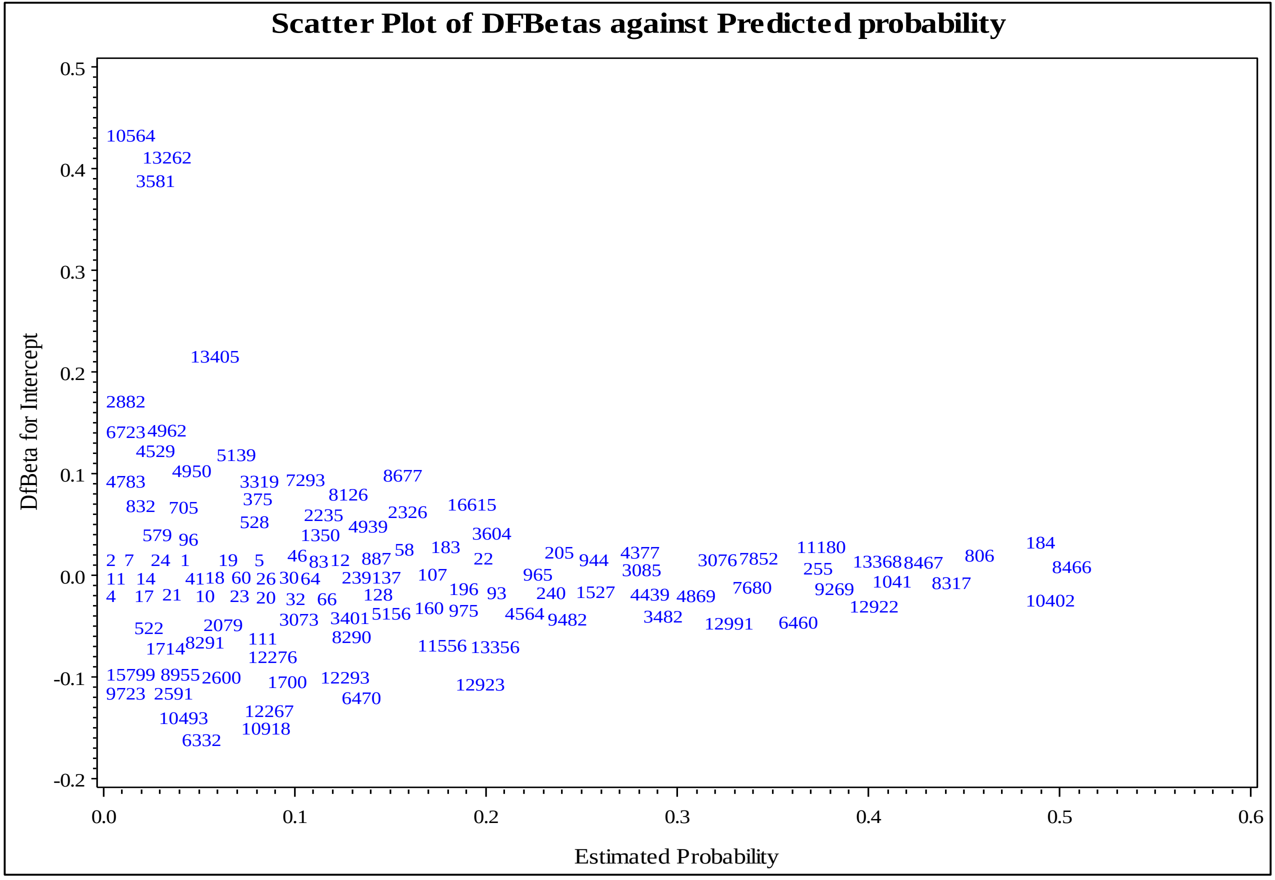
Influence on the Model Fit and Parameter Estimates



Influence Diagnostics



Influence Diagnostics



Scatter Plot of DFBetas against Predicted probability

Diagnostic plots revealed covariate patterns (observations: 3581, 5828, 10564, 13262, 13253, 13405 and 15539) that contributed heavily to the disagreement between the data and the predicted values of the fitted model (poorly fit) and caused instability in the selected parameter estimates. These observations were excluded from the final model reported in table 5 below:

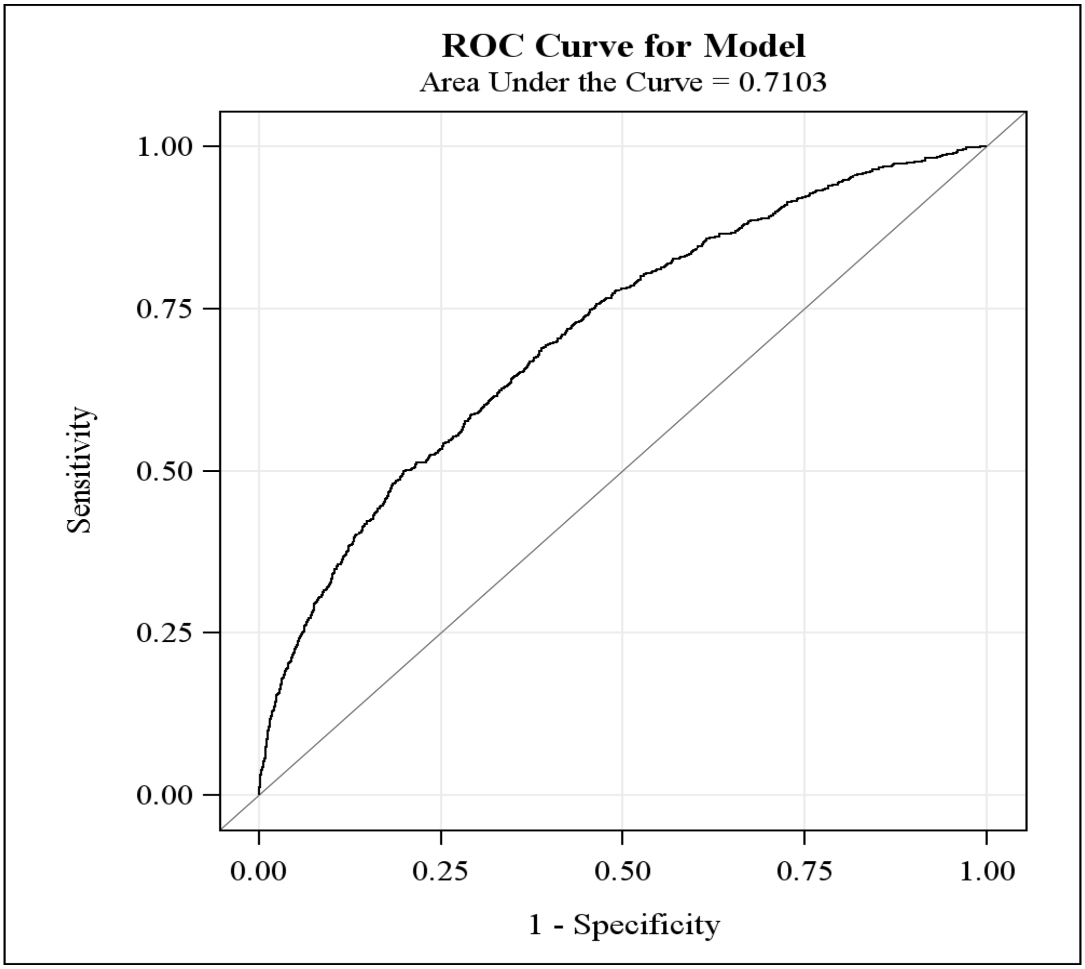
Table 5: Factors associated with HIV infection based on Multivariable weighted logistic regression model (Model 1)

Effect	Adj. OR	95% CI		P-value
		Lower	Upper	
Marital Status				<.0001
Never married (ref)	1			
Married/cohabiting	1.554	1.081	2.236	
Divorced/separated	2.097	1.348	3.261	
Widowed	3.737	2.24	6.233	
COERCED TO HAVE SEX IN THE LAST 12 MONTHS				0.0576
No (ref)	1			
Yes	1.524	0.987	2.356	
ALCOHOL USE AT LAST SEX IN 12 MONTHS?				0.0268
No (ref)	1			
Yes	1.275	1.028	1.581	
HAD COMMERCIAL SEX IN PAST 12 MONTHS?				0.0008
No (ref)	1			
Yes	2.072	1.356	3.166	
HAD AN STI IN PAST 12 MONTHS?				<.0001
No (ref)	1			
Yes	1.601	1.328	1.93	
TOTAL LIFETIME NUMBER OF SEX PARTNERS				<.0001
01 (ref)				
2-3	1.563	1.174	2.08	
4+	1.981	1.466	2.677	
NUMBER OF BIOLOGICAL CHILDREN AWAY FROM HOME				0.0035
None (ref)	1			
1 - 2	1.423	1.146	1.767	
3 - 4	1.318	1.006	1.725	
5+	1.006	0.654	1.547	
Gender × self-perceived risk of getting HIV				0.0018
High vs Low self-perceived risk for Female	1.2293	0.9689	1.5598	
High vs Low self-perceived risk for Circumcised	1.5444	0.8248	2.8921	
High vs Low self-perceived risk for Uncircumcised	1.6525	1.2057	2.2649	
Age group × Ever tested for HIV				0.0275
Age 22-39 vs 15-21 for Never tested	1.8853	1.1033	3.2216	
Age 22-39 vs 15-21 for Ever tested	0.7986	0.5544	1.1503	
Age 40-59 vs 15-21 for Never tested	1.1123	0.577	2.1444	
Age 40-59 vs 15-21 for Ever tested	0.4712	0.2944	0.754	

Ever tested for HIV × Age at first sexual intercourse				0.0159
Gender of household head × Education				0.2497
Female vs Male headed HHD for No Education	1.763	0.9893	3.1418	
Female vs Male headed HHD for Primary Education	1.6418	1.248	2.1598	
Female vs Male headed HHD for Secondary	1.1152	0.7558	1.6457	
Female vs Male headed HHD for Higher Education	0.9333	0.343	2.5396	
Type of place of residence × Wealth Index				<.0001
Urban vs Rural for Poorest	1.9192	0.5431	6.7812	
Urban vs Rural for Poorer	8.54E-06	3.92E-06	0.000019	
Urban vs Rural for Middle	1.4507	0.5309	3.9644	
Urban vs Rural for Richer	2.3123	1.3611	3.9282	
Urban vs Rural for Richest	0.8916	0.6162	1.2903	

ref: reference group; HHD: head of household

Figure 8: ROC curve from the weighted multivariable logistic regression model



From the ROC analysis, the area under the ROC curve of 0.7103 (71.03%) depicted that the model had good discrimination capability to distinguish respondents who were HIV positive from those who were HIV negative. Furthermore, the Hosmer and Lemeshow goodness-of-fit test was not significant (P-value>0.05) hence we failed to reject the null hypothesis of good fit and concluded that the model fit was adequate (H-L χ^2 = 8.692; p-Value = 0.369).

After adjusting for other factors, the results from design-based logistic regression analysis (table 5) are as follows: The odds of HIV infection among respondents with a high self-perceived risk of getting HIV compared with respondents with a low self-perceived risk of contracting HIV varied by gender. High odds of HIV infection were observed among the uncircumcised males (adj. OR = 1.65, 95% CI: 1.21 – 2.26) and the association was significant. Non-significant association between self-perceived risk of getting HIV with HIV infection was observed for females (adj. OR = 1.23, 95% CI: 0.97 – 1.56) and circumcised males (adj. OR = 1.54, 95% CI: 0.82 – 2.89).

The primary research hypothesis of interest was whether circumcised males with a high self-perceived risk of contracting HIV were more likely to be HIV positive than uncircumcised males with a low self-perceived risk of contracting HIV. In the final model, a *contrast* statement in SAS SURVEYLOGISTIC procedure was used in order to compare the groups of interest. It was observed that the effect of interest, gender and self-perceived risk of getting HIV interaction effect was statistically significant (χ^2=25.530; p-value = 0.0018). Circumcised males with a high self-perceived risk of contracting HIV had 14.5% lower odds of HIV infection when compared to uncircumcised males with a low self-perceived risk of contracting HIV (adj. OR = 0.855, 95% CI: 0.48 – 1.53). However, the association was not significant. On the contrary, circumcised males with a high self-perceived risk of contracting HIV had 49% lower odds of HIV infection when compared with uncircumcised males with a high self-perceived risk of contracting HIV (adj. OR = 0.51, 95% CI: 0.28 –0.92) and the association was significant.

Other factors that remained statistically significantly associated with HIV infection after adjustment were: marital status (married/cohabiting, divorced/separated or widowed), alcohol consumption during sex in past 12 months, engaging in commercial or exchange sex for money or gifts, having an STI, having more than 2 total lifetime number of sex partners, and having 1-2 (adj. OR =1.42, 95% CI: 1.15 – 1.77) or 3-4 (adj. OR =1.32, 95% CI: 1.01 – 1.73) biological children outside the home (table 5). From the analysis, we also observed several significant interaction effects as described below.

The association between age in years and HIV infection varied depending on whether the respondent has ever been tested for HIV or not. For respondents that have never tested for HIV, the odds of HIV infection were higher among respondents aged 22-39 years (adj. OR =1.89, 95% CI: 1.10 – 3.22) compared to their counterparts aged 15 – 21 years. For respondents that have ever tested for HIV, relative to those aged 15-21 years, the adjusted odds of HIV infection were 53 times lower among respondents aged 40-59 years (adj. OR =0.47, 95% CI: 0.29 – 0.75). Furthermore, the association between age at first sexual intercourse in years with HIV infection varied depending on whether the respondent has ever been tested for HIV or not. As age at first sexual intercourse increased, the odds of HIV infection increased among respondents who have never tested for HIV and decreased for those that have ever tested for HIV (Appendix B).

The association between gender of household head and HIV infection varied depending on the level of education. The adjusted odds of HIV infection were 64% higher among female-headed households (adj. OR =1.64, 95% CI: 1.25 – 2.16) compared to male headed households for respondents with a primary education as the highest level of education and not significant for other education categories.

The association between type of place of residence and HIV infection varied depending on the respondent's wealth index. For richer respondents, the odds of HIV infection were higher among urban dwellers (adj. OR =2.31, 95% CI: 1.36 – 3.93) compared to rural dwellers.

Missing data analysis

Due to missing values for one or more explanatory variables, data for 14,962 out of 18,395 (81%) respondents were used in the analysis. During model building, variables with a missing data rate of more than 20% where excluded from the multivariable model (step 2). For the variables in the final model, proportions of missing data are summarized below;

Table 6: Proportion of missing data (weighted)

Variable label	Total N = 18,410	Missing values Count (%)
HIV results	18410	0
Age group	18410	0
Marital status	18410	0
Age at first sexual intercourse (C)	18330	80 (0.43%)
Type of place of residence	18410	0
Forced/coerced to have sex against your will in last 12 months?	18410	0
Alcohol consumption during sex in last 12 months?	18410	0
Highest educational level	18410	0
Ever been tested for HIV?	18410	0
Gender of household head	18410	0
Had Commercial sex in last 12 months?	18410	0
Had an STI in last 12 months?	18410	0
Total lifetime number of sex partners(C)	17894	516 (2.80%)
Total children away(C)	18410	0
Wealth index	18410	0
Gender-Circumcision status	18410	0
Are the chances that you can get HIV high or low?	15428	2982 (16.20%)

From the above table, few variables from the final model had missing values.

Table 7 displays the comparison of parameter estimates from the first final model 1 (with missing data excluded) and the model with missing data imputes. It is observed that the estimates did not vary much between the two models. A paired t-test statistical method was further used to compare the estimates from the 2 models. Based on the t-test results (t = -1.06, *p* = 0.297), assuming estimates were normally distributed, we failed to reject the null hypothesis of a difference between estimates and we therefore conclude that there was no statistically significant difference between the estimates from the final model with missing data excluded and the model with missing data imputed. Any slight difference in the SE and p-values might be due to the difference in the sample sizes used in the two models.

Table 7: Comparison of the final model with missing data imputed (model 3) and the model with missing data excluded (model 1)

Effect	**Model 1 - Missing data excluded (complete case analysis)**					**Model 3- Missing data imputed (Multiple imputation analysis)**				
	β	**SE (β)**	**OR[95% CI]**	**Precision**	**P-value**	**β**	**SE (β)**	**OR[95% CI]**	**Precision**	**P-value**
Intercept	-5.52	0.518		2.0286	<.0001	-5.21	0.3438		1.3528	<.0001
Gender × self-perceived risk of getting HIV status										
Female-Low - Risk(ref)										
Female - High-Risk	0.21	0.122	1.23(0.972-1.565)	0.4765	0.085	0.1	0.093	1.10(0.918-1.324)	0.3715	0.2993
Circumcised Male - High-Risk	-0.09	0.287	0.91(0.520-1.603)	1.1257	0.7514	-0.19	0.223	0.83(0.536-1.283)	0.8855	0.4029
Circumcised Male - Low-Risk	-0.53	0.21	0.59(0.391-0.890)	0.8216	0.0119	-0.61	0.186	0.54(0.377-0.781)	0.7366	0.0013
Uncircumcised Male - High-Risk	0.58	0.184	1.78(1.238-2.551)	0.7227	0.0018	0.38	0.135	1.47(1.127-1.909)	0.5288	0.0046
Uncircumcised Male-Low-Risk	0.09	0.129	1.09(0.849-1.409)	0.506	0.4868	0.02	0.097	1.02(0.844-1.235)	0.3812	0.833
Marital Status										
Never married (ref)										
Divorced/separated	0.77	0.224	2.17(1.400-3.364)	0.8768	0.0005	0.92	0.161	2.51(1.833-3.450)	0.6322	<0.0001
Married/cohabiting	0.46	0.185	1.58(1.100-2.274)	0.7264	0.0134	0.24	0.146	1.27(0.954-1.690)	0.5717	0.1013
Widowed	1.33	0.264	3.78(2.251-6.346)	1.0363	<.0001	1.64	0.177	5.17(3.654-7.326)	0.6955	<.0001
COERCED TO HAVE SEX IN THE LAST 12 MONTHS										
Yes vs No	0.42	0.22	1.53(0.993-2.354)	0.863	0.0537	0.21	0.148	1.23(0.924-1.650)	0.5799	0.1538
ALCOHOL USE AT LAST SEX IN 12 MONTHS?										
Yes vs No	0.25	0.109	1.28(1.036-1.589)	0.4279	0.0226	0.29	0.083	1.34(1.139-1.577)	0.3252	0.0004
HAD COMMERCIAL SEX IN PAST 12 MONTHS?										
Yes vs No	0.72	0.217	2.05(1.342-3.138)	0.8493	0.0009	0.48	0.162	1.62(1.181-2.230)	0.6357	0.0028
HAD AN STI IN PAST 12 MONTHS?										
Yes vs No	0.47	0.095	1.59(1.323-1.920)	0.3722	<.0001	0.61	0.068	1.84(1.607-2.097)	0.2663	<.0001
TOTAL LIFETIME NUMBER OF SEX PARTNERS										
1 (ref)										
2 to 3	0.43	0.144	1.53(1.155-2.031)	0.5646	0.0031	0.43	0.105	1.54(1.253-1.893)	0.4125	<.0001
4+	0.64	0.151	1.89(1.406-2.543)	0.5925	<.0001	0.84	0.113	2.32(1.856-2.896)	0.4454	<.0001

NUMBER OF BIOLOGICAL CHILDREN AWAY FROM HOME										
0 (ref)										
1 - 2	0.34	0.11	1.41(1.133-1.745)	0.4321	0.002	0.27	0.075	1.31(1.134-1.519)	0.2929	0.0003
3 - 4	0.25	0.138	1.29(0.983-1.689)	0.5412	0.0664	0.12	0.102	1.12(0.921-1.371)	0.3979	0.2513
5+	-0.03	0.22	0.97(0.632-1.496)	0.8626	0.8978	-0.14	0.139	0.87(0.661-1.142)	0.5466	0.3144
Type of place of residence × Wealth Index										
Rural-Poorest (ref)										
Rural-Middle	0.19	0.175	1.20(0.855-1.695)	0.6846	0.2891	0.14	0.125	1.15(0.900-1.468)	0.4891	0.2636
Rural-Poorer	0.23	0.164	1.26(0.912-1.733)	0.6419	0.1617	0.15	0.12	1.16(0.915-1.465)	0.471	0.2237
Rural-Richer	0.34	0.17	1.41(1.007-1.963)	0.6676	0.0454	0.29	0.122	1.33(1.050-1.696)	0.4793	0.0182
Rural-Richest	0.58	0.198	1.78(1.206-2.623)	0.7773	0.0037	0.56	0.154	1.74(1.287-2.358)	0.6054	0.0003
Urban-Middle	0.58	0.498	1.79(0.676-4.754)	1.9511	0.2411	0.28	0.42	1.32(0.581-3.011)	1.6459	0.5062
Urban-Poorer	-11.46	0.403	0.00(0.000-0.000)	1.5794	<.0001	-1.85	1.146	0.16(0.017-1.490)	4.4922	0.107
Urban-Poorest	0.69	0.641	1.99(0.566-6.978)	2.5124	0.2841	0.88	0.369	2.40(1.163-4.951)	1.4482	0.0178
Urban-Richer	1.2	0.279	3.32(1.926-5.739)	1.0919	<.0001	1.08	0.241	2.95(1.840-4.733)	0.945	<.0001
Urban-Richest	0.45	0.197	1.58(1.072-2.316)	0.7707	0.0208	0.51	0.14	1.66(1.259-2.184)	0.5507	0.0003
Ever tested for HIV × Age at first sexual intercourse										
	0.01	0.102	1.01 (0.828-1.234)	0.398	0.9147	0.04	0.069	1.04(0.906-1.185)	0.2688	0.6047
Age group × Ever tested for HIV										
Ever Tested - Age15-21 (ref)										
Ever Tested - Age22–39	-0.23	0.184	0.79(0.553-1.138)	0.7208	0.208	0.33	0.139	1.39(1.056-1.821)	0.5446	0.0186
Ever Tested - Age40–59	-0.75	0.234	0.47(0.300-0.750)	0.9172	0.0014	0.34	0.162	1.40(1.021-1.928)	0.635	0.0365
Never Tested - Age15-21	-0.46	0.363	0.63(0.310-1.285)	1.4213	0.205	-0.61	0.286	0.54(0.309-0.948)	1.122	0.0319
Never Tested - Age22–39	0.29	0.301	1.34(0.741-2.412)	1.1799	0.3346	0.14	0.226	1.15(0.737-1.788)	0.8859	0.5405
Never Tested - Age40–59	-0.14	0.324	0.87(0.461-1.646)	1.2718	0.671	-0.31	0.246	0.74(0.454-1.191)	0.9653	0.2113
Gender of household head × Education										
Higher Education–Male HHD (ref)										
Higher Education – Female HHD	-0.06	0.51	0.95(0.348-2.571)	1.9985	0.9141	0.39	0.322	1.48(0.786-2.778)	1.2628	0.2255
Secondary– Female HHD	0.93	0.327	2.53(1.330-4.797)	1.2826	0.0046	0.82	0.233	2.27(1.440-3.588)	0.9131	0.0004

Secondary – Male HHD	0.81	0.299	2.25(1.254-4.053)	1.1733	0.0066	0.73	0.221	2.08(1.352-3.211)	0.8647	0.0009
Primary– Female HHD	1.43	0.306	4.16(2.283-7.578)	1.1996	<.0001	1.27	0.232	3.57(2.264-5.630)	0.9109	<.0001
Primary – Male HHD	0.93	0.302	2.53(1.401-4.579)	1.1845	0.0021	0.94	0.222	2.56(1.659-3.963)	0.871	<.0001
No education – Female HHD	1.16	0.347	3.18(1.607-6.273)	1.3617	0.0009	1.33	0.27	3.79(2.229-6.430)	1.0596	<.0001
No education – Male HHD	0.6	0.336	1.81(0.938-3.508)	1.3185	0.0766	0.85	0.239	2.34(1.464-3.735)	0.9367	0.0004

*Precision = Upper – Lower (95% Confidence bounds for estimates).

ref: reference group; HHD: head of household

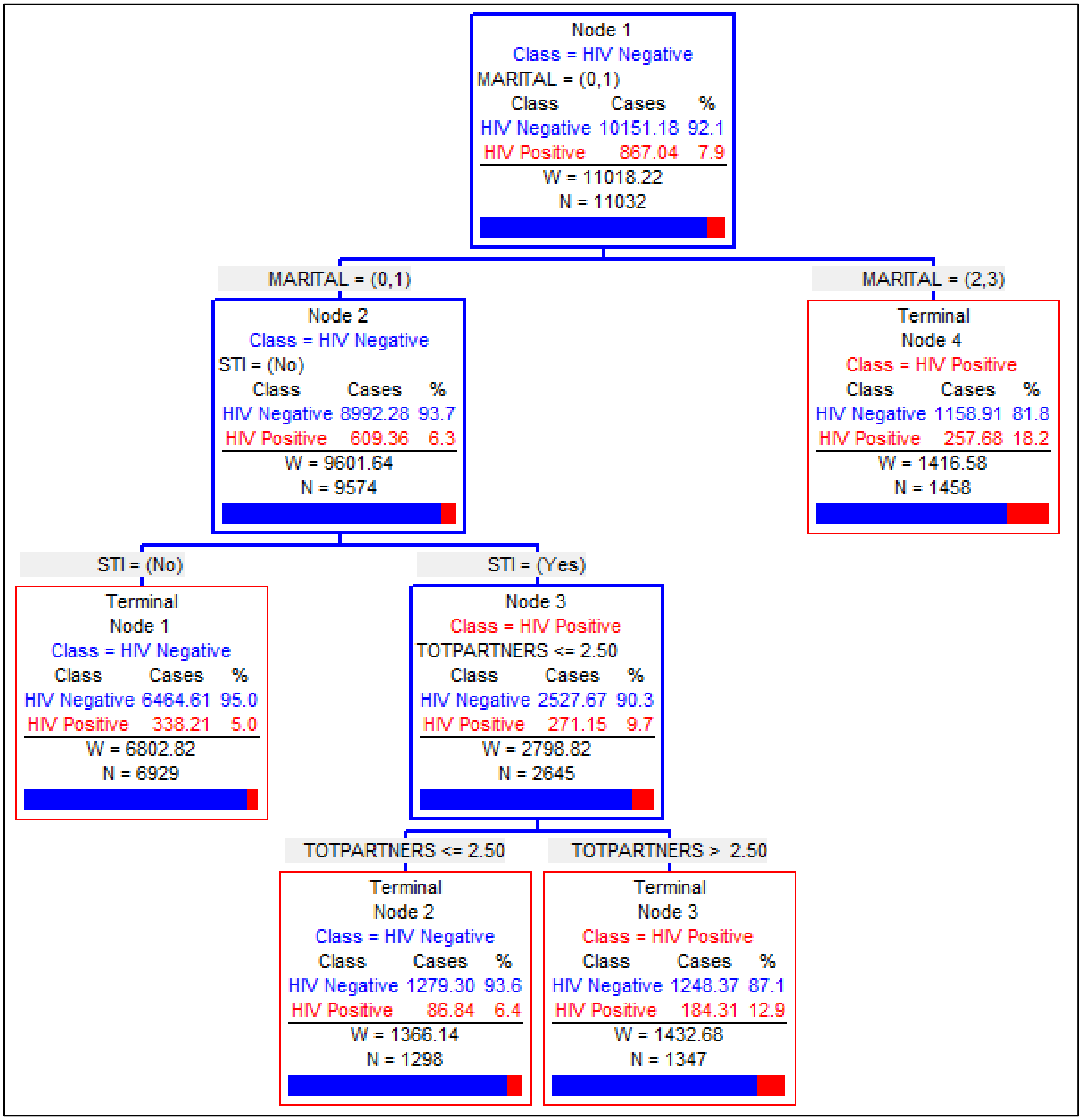
Note: Any interaction effect in final model one was recreated as a new variable (a combination of two variables) for convenience especially with fitting the multiple imputation model. Reference categories changed and the ORs in this table are exponentials for the obtained estimates hence they should not be used for interpretation purposes.

CART analysis

Classification and regression tree (CART) analysis was performed to identify complex higher level interaction effects that were associated with HIV infection. The final result from CART analysis is a decision tree, a classification tree in our case, and the optimal tree is one that minimizes the relative error rate. Figure 9 shows the optimal tree built from the training/learn data set and evaluated using the validation/ test data set. An initial split for the tree was on marital status, and 4 terminal nodes/sub-groups were formed. Variables identified that were associated with HIV infections included marital status, having an STI in past 12 months, and total lifetime number of sex partners.

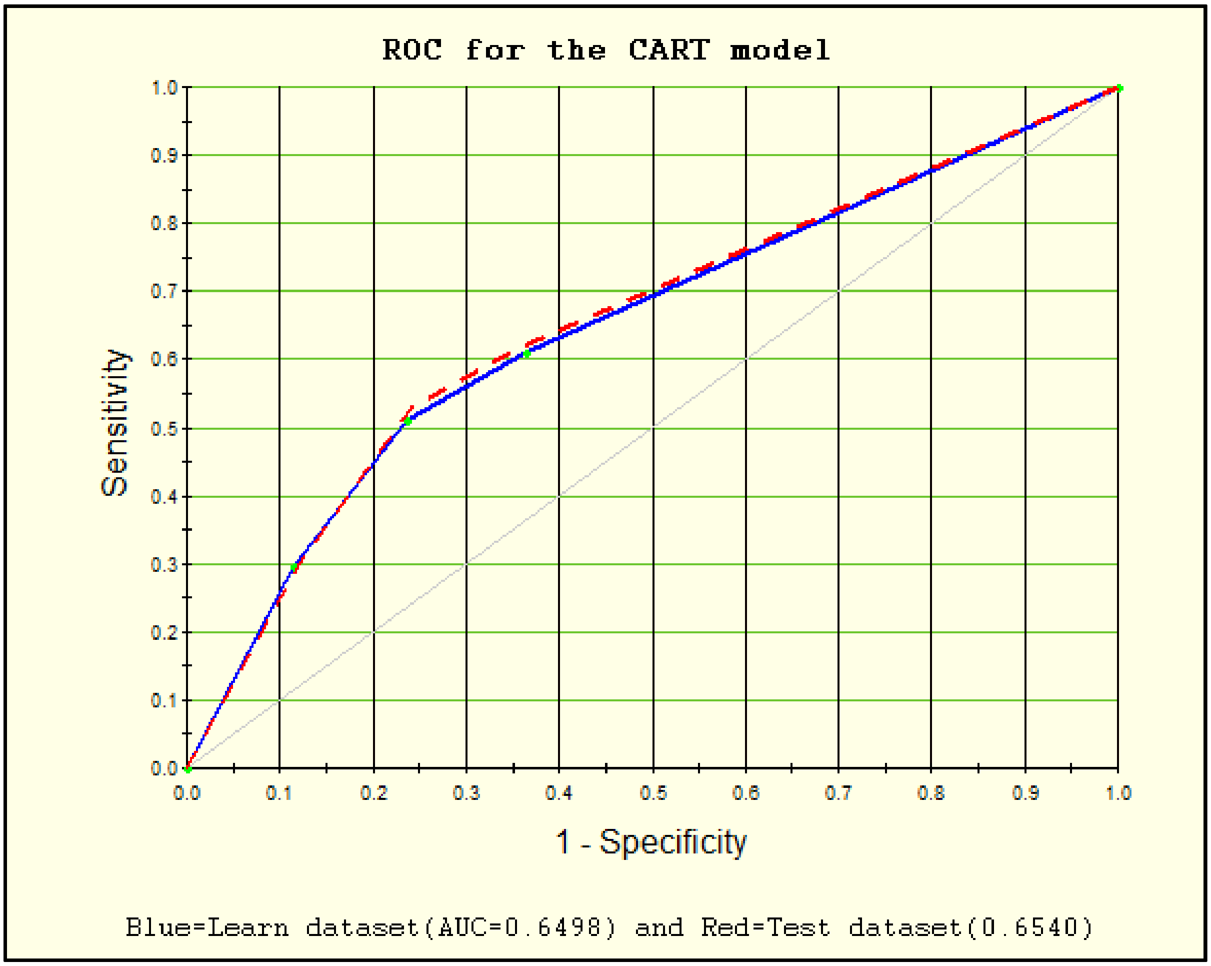
CART analysis revealed four possible/potential interaction patterns (marital status × having an STI, marital status × Total lifetime number of sex partners, having an STI × Total lifetime number of sex partners and Marital status × having an STI × Total lifetime number of sex partners) that were further evaluated for statistical significance using design-based logistic regression model. The variables marital status and total lifetime number of sex partners were dichotomized based on the cut-points from the classification tree (see description under table 3).

Figure 9: Weighted CART model for HIV data



Key: Marital: 0 = "Never married"; 1 = "Married/cohabiting"; 2 = "Widowed"; 3 = "Divorced/separated"
W = Weighted count and N = Unweighted count

Figure 10: ROC for the CART model



The area under the ROC curve was used to evaluate the extent of misclassification CART commits in allocating the dataset cases to terminal nodes. The CART model on the learn/training data set had an area under the ROC curve of 65%, approximately equal to that of the test data set (figure 10). Furthermore, the CART model had a classification accuracy of 74.3% and 73.9% on the learn/training and test/validation data sets respectively.

Table 8: Multivariable weighted logistic regression models for CART analysis

Effect		Category		Model 4*			Model 5**			Model 6***		
				β	SE (β)	P-value	β	SE (β)	P-value	β	SE (β)	P-value
Intercept				-4.2665	0.2158	<.0001	-4.4954	0.2672	<.0001	-4.4000	0.2631	<.0001
Marital Status (CART)	Divorced/Separated/Widowed			0.7304	0.1960	0.0002	0.5806	0.1467	<.0001	0.6164	0.1463	<.0001
Had an STI in past 12 months?	Yes			0.1967	0.1514	0.1938	0.4478	0.1029	<.0001	0.3853	0.1070	0.0003
Total lifetime number of sex partners												
	2-3						0.4711	0.1410	0.0008			
	4+						0.6981	0.1524	<.0001			
Total lifetime sex partners (CART)	3+			0.3435	0.1222	0.0049				0.4205	0.1174	0.0003
Number of biological children away from home												
	1-2			0.3821	0.0927	<.0001	0.3649	0.1034	0.0004	0.2703	0.1271	0.0335
	3-4			0.2592	0.1220	0.0336	0.2437	0.1198	0.0420	0.0339	0.1746	0.8463
	5+			-0.0708	0.1526	0.6425	-0.0933	0.1829	0.6100	-0.8936	0.3092	0.0039
Coerced to have Sex in past 12 months (Yes)				0.4249	0.1766	0.0161	0.4114	0.2153	0.0560	0.3976	0.2184	0.0687
Consumed alcohol during Sex in past 12 months (Yes)				0.2466	0.0915	0.0070	0.2352	0.1064	0.0270	0.0550	0.1527	0.7189
Education												
	No education			0.0837	0.2350	0.7219	0.1033	0.2637	0.6951	0.3057	0.2623	0.2439
	Primary			0.4730	0.2034	0.0201	0.4667	0.2359	0.0479	0.6289	0.2329	0.0069
	Secondary			0.3967	0.2132	0.0628	0.3838	0.2727	0.1593	0.4654	0.2677	0.0822
Ever Tested for HIV?	No			0.2964	0.0838	0.0004	0.3008	0.0953	0.0016	0.3287	0.0971	0.0007
Gender-Circumcision status												
	Male-Circumcised			-0.4783	0.1508	0.0015	-0.5177	0.1702	0.0024	-0.8083	0.2771	0.0035
	Male-Uncircumcised			-0.0225	0.0971	0.8166	-0.0535	0.1224	0.6623	-0.2662	0.1657	0.1081
Self-perceived risk of getting HIV	High			0.3553	0.0815	<.0001	0.3400	0.1055	0.0013	0.3049	0.1217	0.0122
Engaged in commercial Sex	Yes			0.6844	0.1606	<.0001	0.6826	0.2113	0.0012	0.7651	0.3443	0.0263
Type of place of residence	Urban									0.3601	0.1279	0.0049
Marital Status(CART) × Had an STI	Divorced/Separated/Widowed	Yes		0.2089	0.3360	0.5340	0.0915	0.2196	0.6770	0.1118	0.2220	0.6144
Marital Status(CART) × Total lifetime sex partners(CART)												

Effect		Category		Model 4*			Model 5**			Model 6***		
				β	SE (β)	P-value	β	SE (β)	P-value	β	SE (β)	P-value
	Divorced/Separated/Widowed	3+		-0.2189	0.2553	0.3912						
Had an STI in past 12 months × Total lifetime sex partners (CART)		Yes	3+	0.3952	0.1875	0.0351						
Marital Status(CART) × Had an STI × Total lifetime sex partners (CART)												
	Divorced/Separated/Widowed	Yes	3+	-0.1607	0.4167	0.6998						
Gender × # Biological children away from home												
	Male-Circumcised	1-2								0.3237	0.4509	0.4728
	Male-Uncircumcised	1-2								0.2867	0.2118	0.1758
	Male-Circumcised	3-4								0.0859	0.5062	0.8653
	Male-Uncircumcised	3-4								0.6639	0.2842	0.0195
	Male-Circumcised	5+								1.8391	0.4976	0.0002
	Male-Uncircumcised	5+								1.3174	0.3942	0.0008
Self-perceived risk of getting HIV × Consumed alcohol during Sex												
	High	Yes								0.3885	0.2066	0.0601
Self-perceived risk of getting HIV × Engaged in commercial Sex												
	High	Yes								-0.9698	0.3552	0.0063
Had an STI in past 12 months × Engaged in commercial Sex												
	Yes	Yes								0.7653	0.4039	0.0581
AIC				5476.473			5472.585			5441.687		
H-L GOF fit statistic (P-value)				5.402 (0.7139)			5.5998 (0.6920)			11.8886 (0.1562)		
Area under the ROC curve (Overall)				0.682			0.683			0.692		

*Model 4: Includes variables from the optimal tree, interaction effects amongst them plus any potential confounders

** Model 5: Includes variables from the optimal tree further pruned to exclude the total number of lifetime sex partners plus any potential confounders

*** Model 6: Includes variables from the optimal tree, the rest of the potential confounders and all potential interactions

From table 8, it is observed that all interaction patterns from CART analysis, except having an STI × total lifetime number of sex partners, were not statistically significant (Table 9, Model 4). When other interaction terms among the confounders were allowed into the model (model 6), significant interaction terms identified included: 1) self-perceived risk of getting HIV × engaging in commercial sex (p-value = 0.0063), 2) self-perceived risk of getting HIV × alcohol consumption during sex (p-value = 0.0601), and 3) gender × number of biological children away from home (p-value = 0.0018).

DISCUSSION

This thesis was aimed at exploring sexual behavior, socio-demographic, biological and any other factors associated with HIV infection among a Ugandan population age 15 – 59 years with a history of sexual activity. Particular emphasis was on gender (with categories; female, circumcised male, and uncircumcised male) and self-perceived risk of getting HIV (with categories; low and high) combination as the interaction effect of interest. The data used was from the 2011 Uganda AIDs Indicator Survey (UAIS) and the outcome variable was HIV status of the respondent.

The analysis involved fitting design-based logistic regression models, evaluation of bias due to data missingness and application of a data mining technique, specifically CART models, to identify complex higher level interaction effects that later were evaluated using the design-based logistic regression model. To my knowledge, this is the first study to use CART models to analyze data from the 2011 UAIS as well as to explore the association between HIV infection and an interaction effect consisting of a combination of gender and self-perceived risk of getting HIV.

Findings from this thesis agreed with results from many other studies that HIV infections are fueled by high-risk sexual behaviors (Bolton, 1992; Chimoyi & Musenge, 2014; Kibira, Nansubuga, Tumwesigye, Atuyambe, & Makumbi, 2014; Konde-Lule et al., 1997; Lema, Katapa, & Musa, 2008; Serwadda et al., 1992; Tumwesigye et al., 2012). In this study, we found that the prevalence of HIV among respondents who have ever had sex was 8.15%, much higher than in the general Uganda adult population in 2011 (6.7%). This is due to the fact that the analysis excluded respondents who were not at risk of getting HIV (reported never had sex). In both unadjusted and adjusted analyses, HIV risk was highest among respondents who were currently or previously (divorced/separated or widowed) married, used alcohol during sex, engaged in commercial or exchange sex for money or gifts, had an STI, had 2 or more total lifetime number of sex partners, and had 1-4 biological children outside the home. To my knowledge, no study has extensively explored interaction effects nor explored the association between the number of biological children begotten outside the home with HIV infections in Uganda.

Our main interaction effect of interest was significant. Self-perceived risk of getting HIV was significantly associated with HIV infection for uncircumcised male respondents only. The odds of HIV infection among uncircumcised males with a high self-perceived risk of getting HIV were higher when compared to uncircumcised male respondents with a low self-perceived risk of getting HIV. A likely explanation for this could be based on findings from table 4. According to results in table 4, there was evidence for risk compensation mostly among male respondents. More respondents with a low compared to those with a high self-perceived risk of getting HIV were less likely to consistently use condoms yet more males, both circumcised and uncircumcised, with a low self-perceived risk of getting HIV than males with a high self-perceived risk of getting HIV were more likely to use alcohol during sex, have an STI, have more total lifetime number of sex partners and have multiple sex partners in the past 12 months. Furthermore, unlike the other gender categories, uncircumcised males with a high self-perceived risk of getting HIV were more likely to engage in commercial or exchange sex. HIV infections were also highest among uncircumcised males with a low self-perceived risk of getting HIV.

In addition, our findings show that circumcised males with a high self-perceived risk of contracting HIV had significantly lower odds of HIV infection when compared with uncircumcised males with a high self-perceived risk of contracting HIV. Our findings concur with previous research that has shown a protective effect of male circumcision (R. H. Gray et al., 2007; Kibira et al., 2014).

Using classification trees, our study identified four complex higher level potential interaction effects/patterns in relationship with HIV infection. These included: 1) marital status × STI, 2) marital status × Total lifetime number of sex partners, 3) having an STI × Total lifetime number of sex partners and 4) Marital status × having an STI × Total lifetime number of sex partners. In design-based logistic regression models adjusting for potential confounders, only one interaction effect (having an STI × total lifetime number of sex partners) was found to be statistically significant. Furthermore, main effects and interactions selected in the final CART model 6 were different compared to those in the first model (non-CART model 1). The possible explanation could be due to residual confounding introduced into the logistic regression model after dichotomization of continuous and categorical variables based on CART cut-points.

Limitations

The study faced a number of limitations. First, due to the design of the study, it was hard to address temporality or compare results across various studies. The study would have gained more validity if we were able to determine trends in high-risk sexual behaviors over time or exactly when the respondent tested HIV positive. Second, self-reporting of sensitive information like sexual behaviors and perceived risk of getting HIV is likely to bias the results. This is due to social pressures that discourage accurate reporting of information. Third, household surveys are prone to recall bias as well as difficulties in interpreting questions. Fourth, because of a high rate of missingness, important factors including concurrent sexual partnerships were excluded from multivariable analysis. Lastly, not being able to extensively explore the joint effect of circumcision with alcohol use on HIV infection was a limitation. The strengths of the study included large sample size, use of CART methodology and fitting significant models with good ROC.

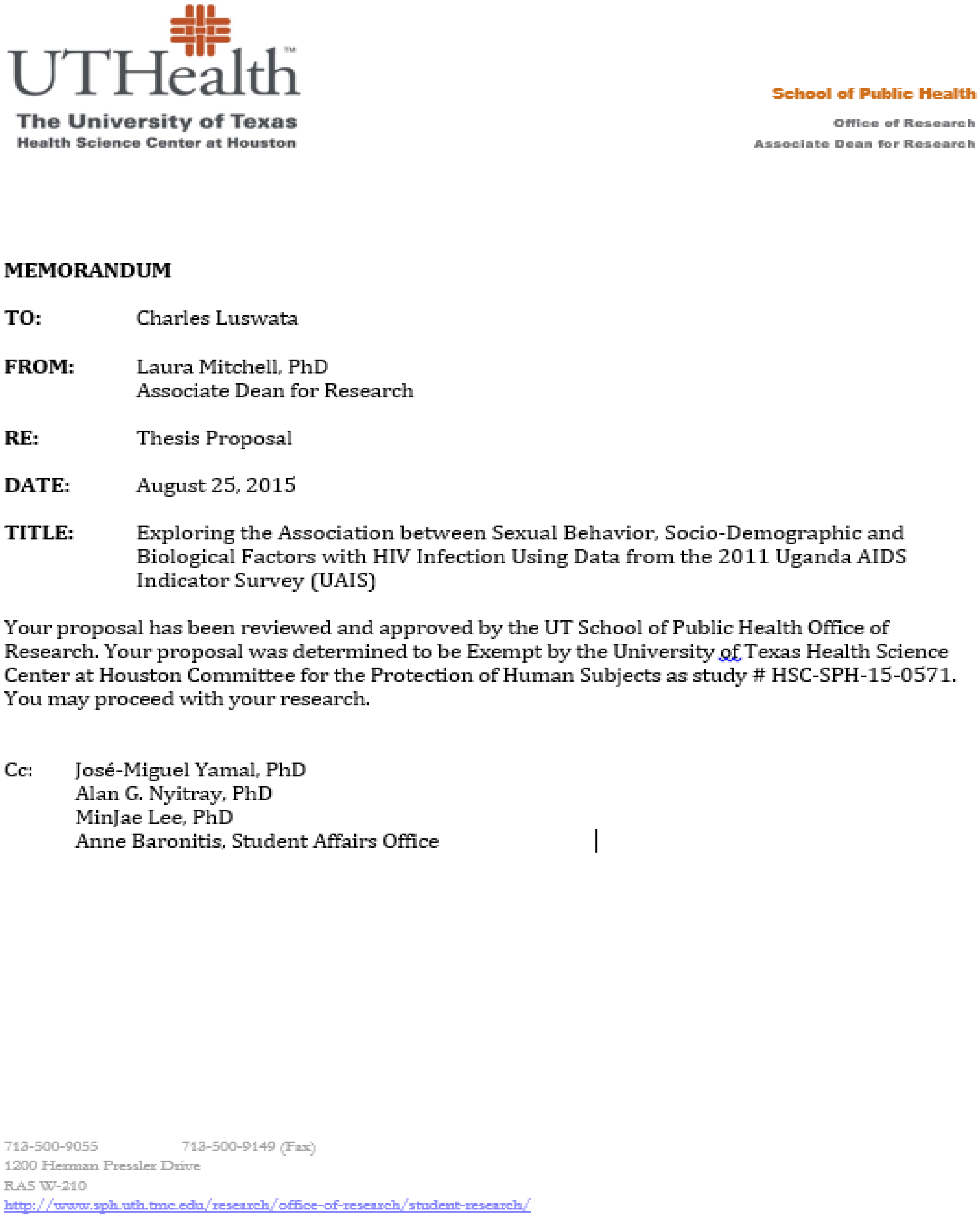
CONCLUSION

In summary, we observed that self-perceived risk of getting HIV was associated with HIV infections but for uncircumcised males only. The study was also able to identify sexual behavior, socio-demographic, biological and other factors as well as interaction effects among the factors that were associated with HIV infection among a sexually active Ugandan population aged 15 – 59 years. Furthermore, the prevalence of HIV was higher than that was seen in the general Uganda adult population in 2011. Programs for the control and treatment of HIV/AIDS should therefore assess and raise awareness of the risk as well as encourage prevention strategies like PrEP, condom use or abstinence among non-married individuals, being faithful among the married or those in committed relationships, routine testing of HIV, treatment of sexually transmitted infections and circumcision among males.

Since all variables in the classification tree were selected in the (first) final design-based logistic regression model as main effects, we can use CART methodology and other data mining techniques to explore associations or predict the outcome of interest as well as categorize continuous variables in complex sample surveys.

APPENDICES

Appendix A: The Institutional Review Board (IRB) Approval Letter



UTHealth™
The University of Texas
Health Science Center at Houston

School of Public Health
Office of Research
Associate Dean for Research

MEMORANDUM

TO: Charles Luswata

FROM: Laura Mitchell, PhD
Associate Dean for Research

RE: Thesis Proposal

DATE: August 25, 2015

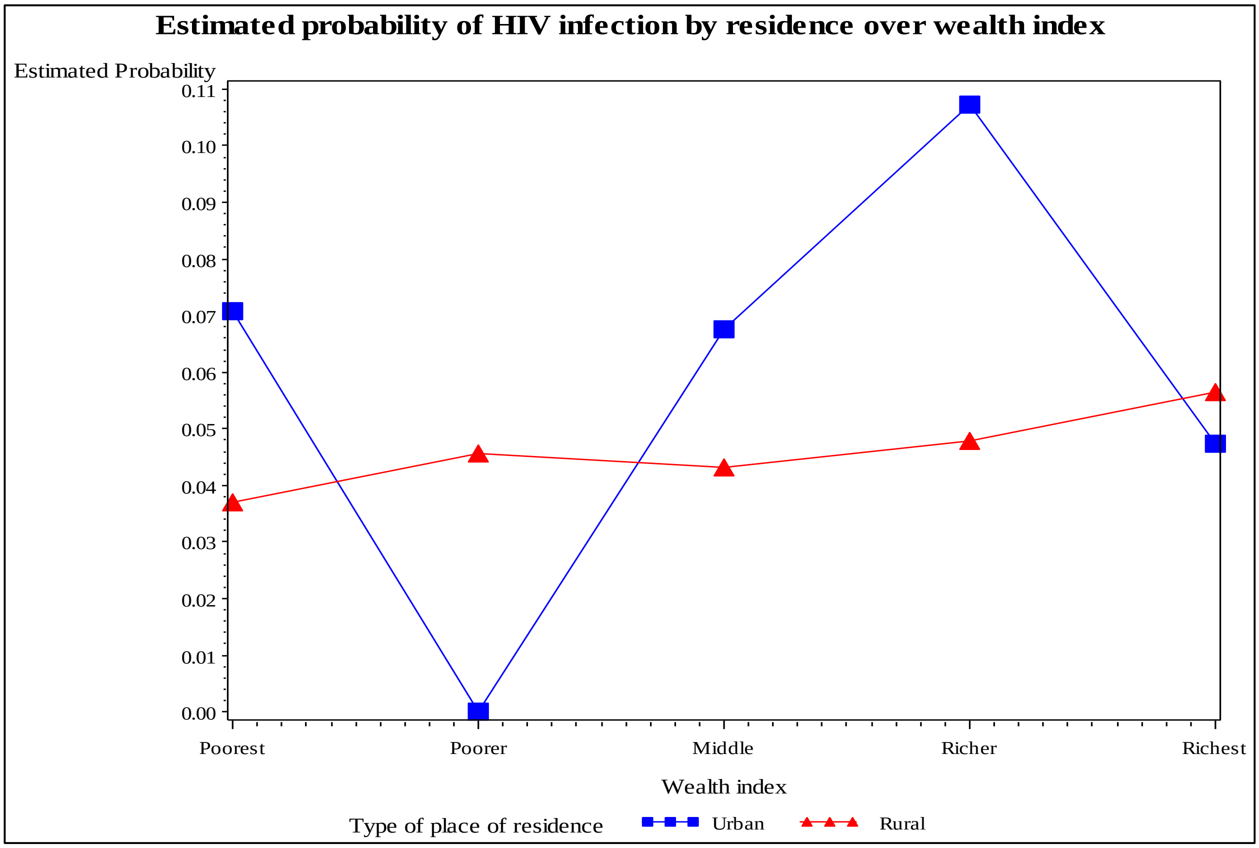
TITLE: Exploring the Association between Sexual Behavior, Socio-Demographic and Biological Factors with HIV Infection Using Data from the 2011 Uganda AIDS Indicator Survey (UAIS)

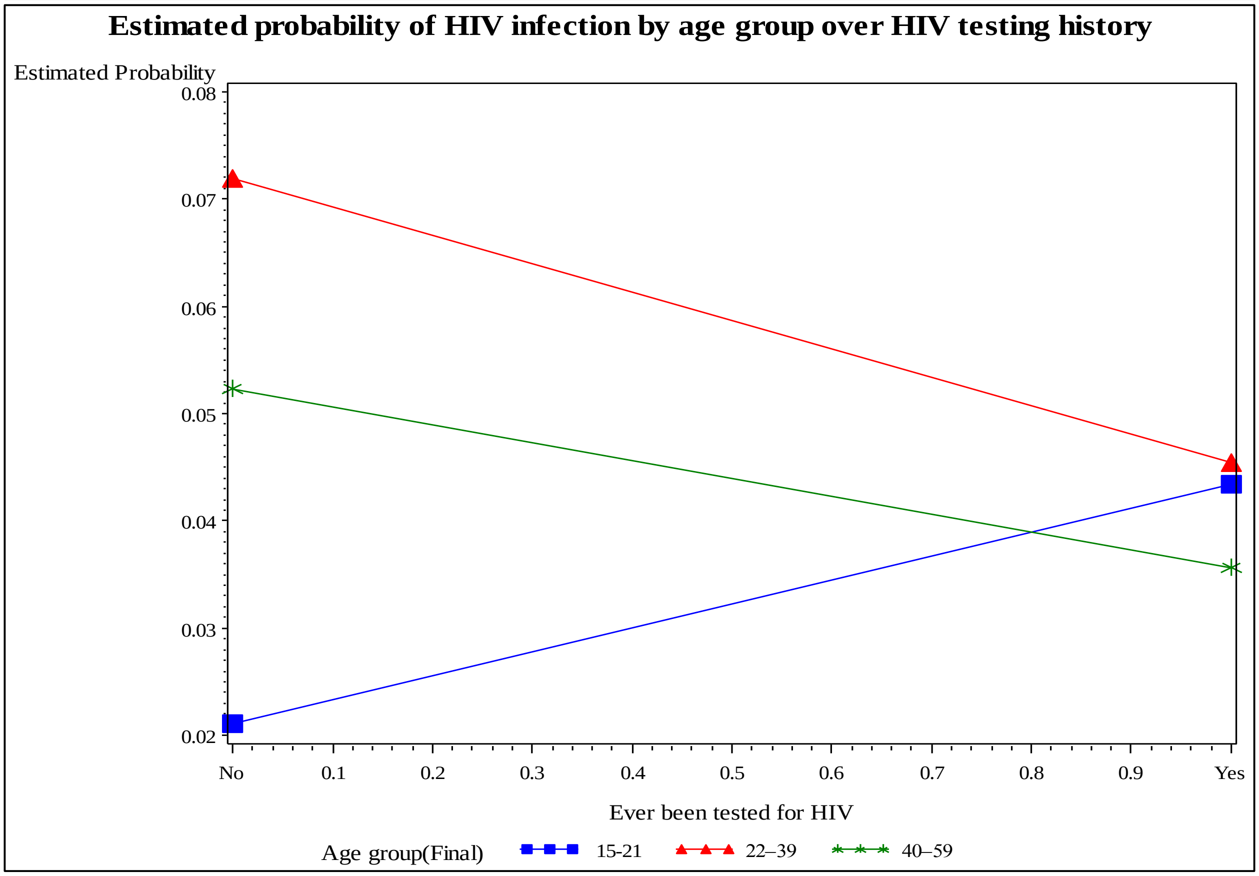
Your proposal has been reviewed and approved by the UT School of Public Health Office of Research. Your proposal was determined to be Exempt by the University of Texas Health Science Center at Houston Committee for the Protection of Human Subjects as study # HSC-SPH-15-0571. You may proceed with your research.

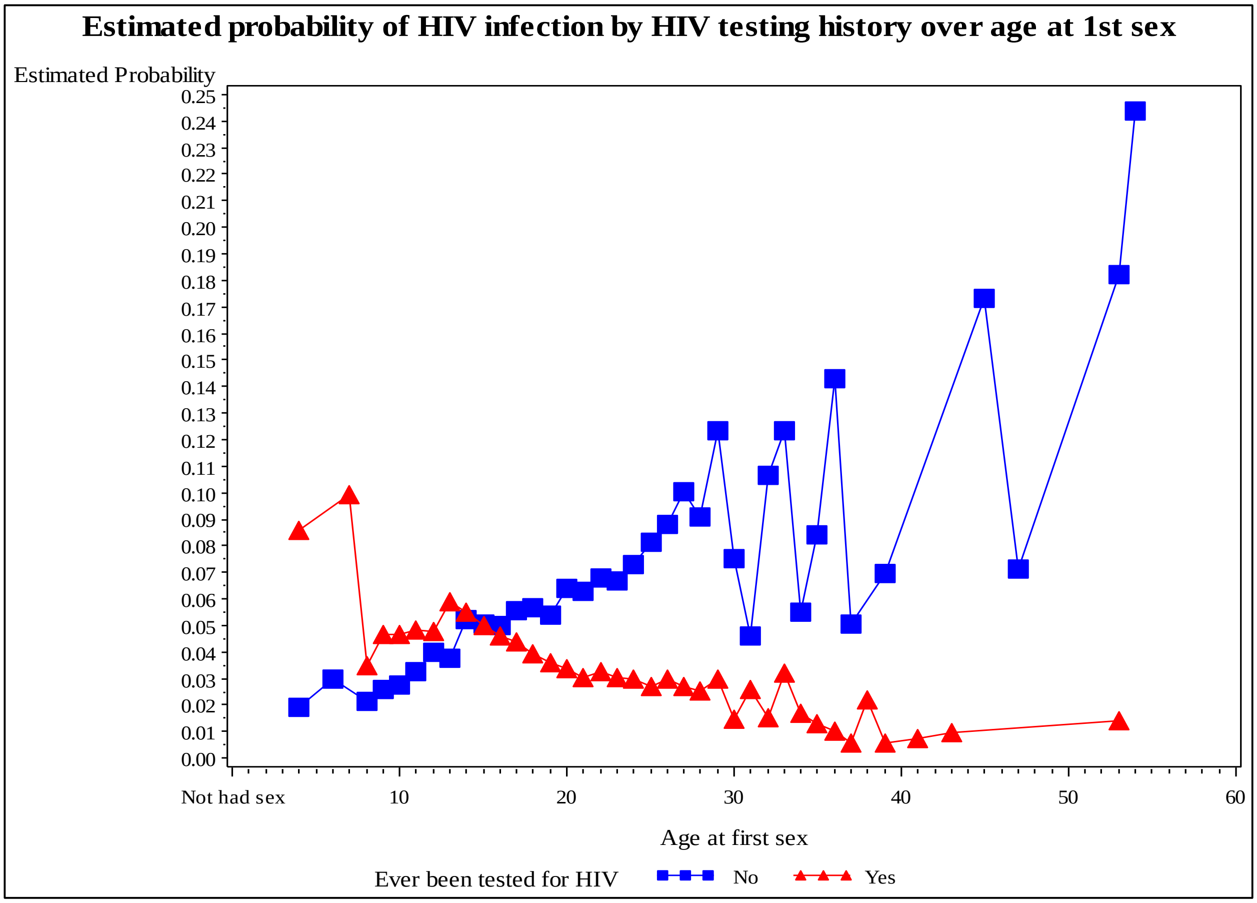
Cc: José-Miguel Yamal, PhD
Alan G. Nyitray, PhD
MinJae Lee, PhD
Anne Baronitis, Student Affairs Office

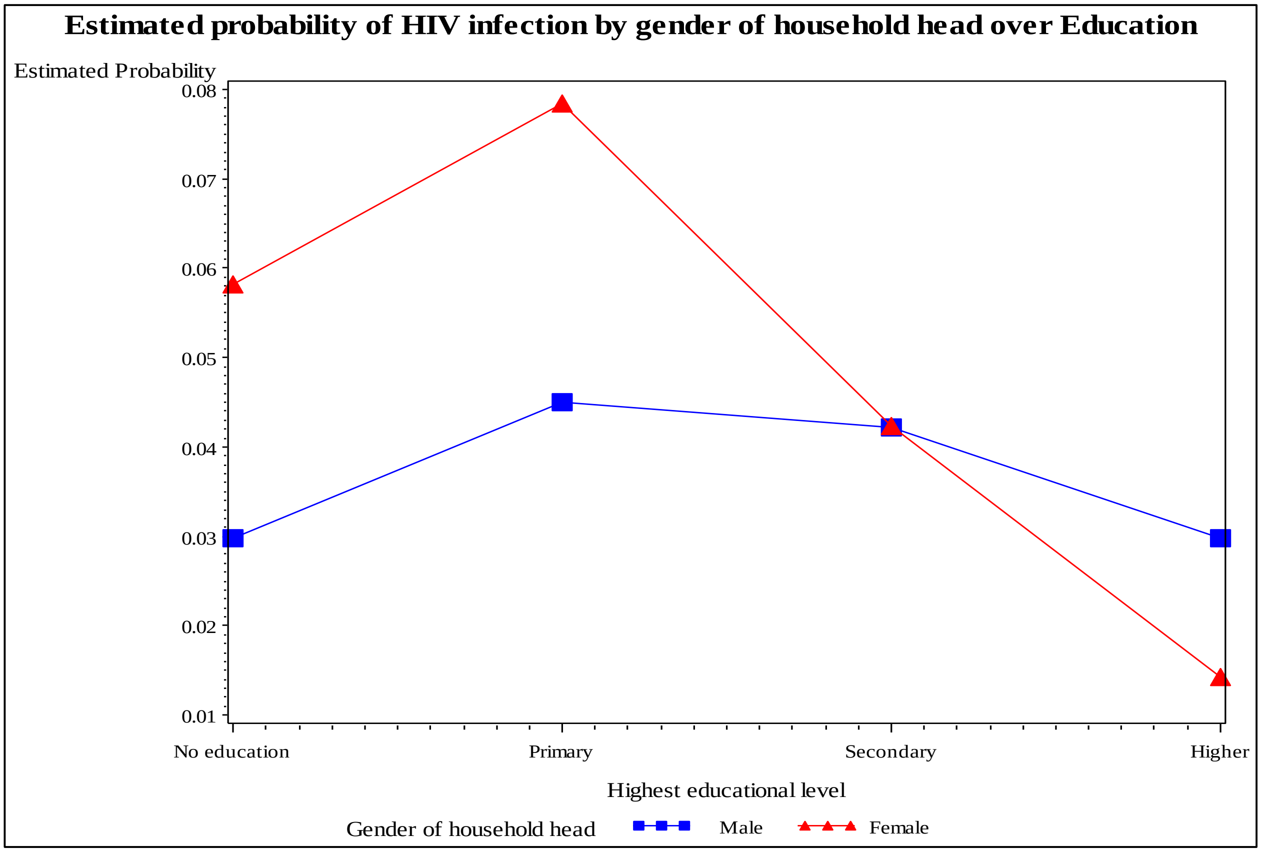
713-500-9055 713-500-9149 (Fax)
1200 Herman Pressler Drive
RAS W-210
http://www.sph.uth.tmc.edu/research/office-of-research/student-research/

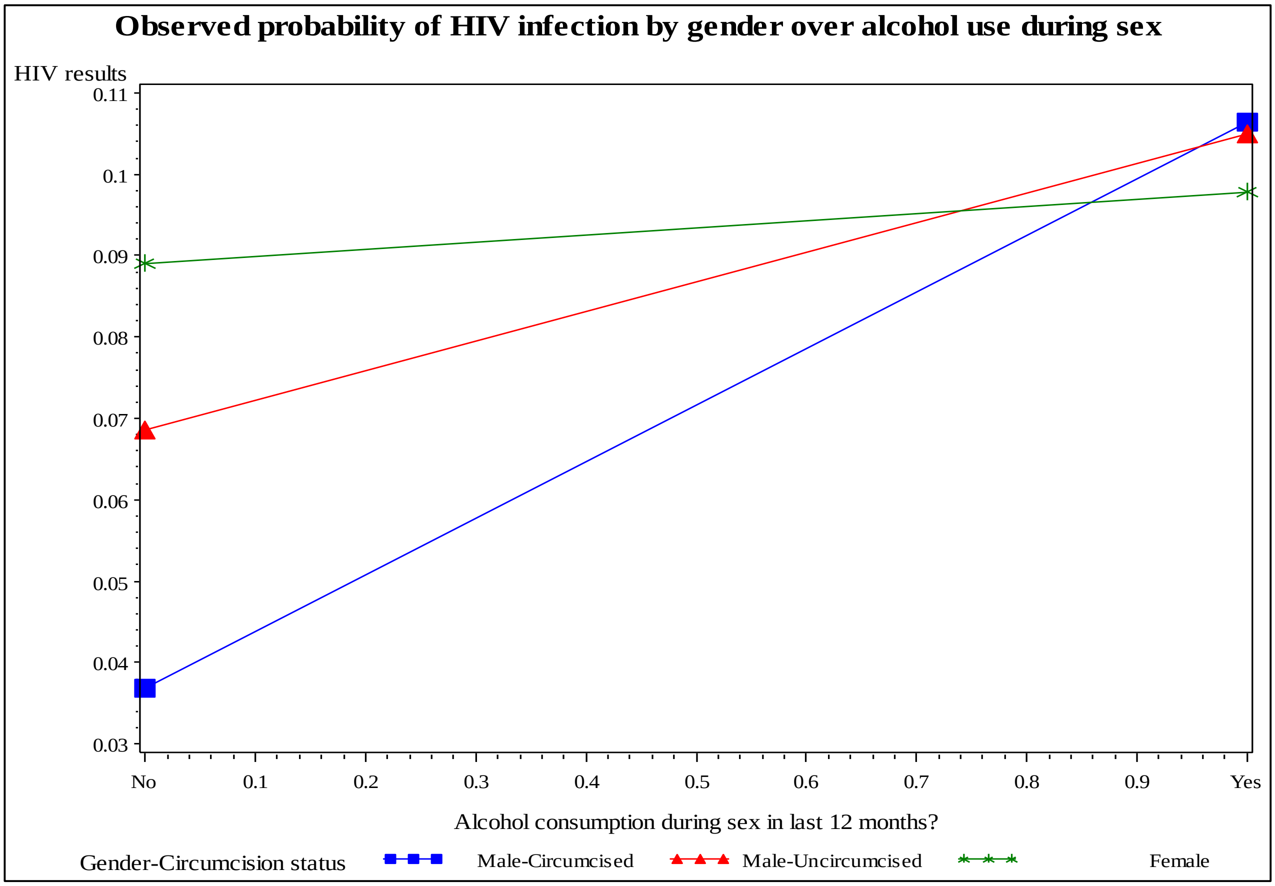
Appendix B: Plots of estimated probability of HIV infections



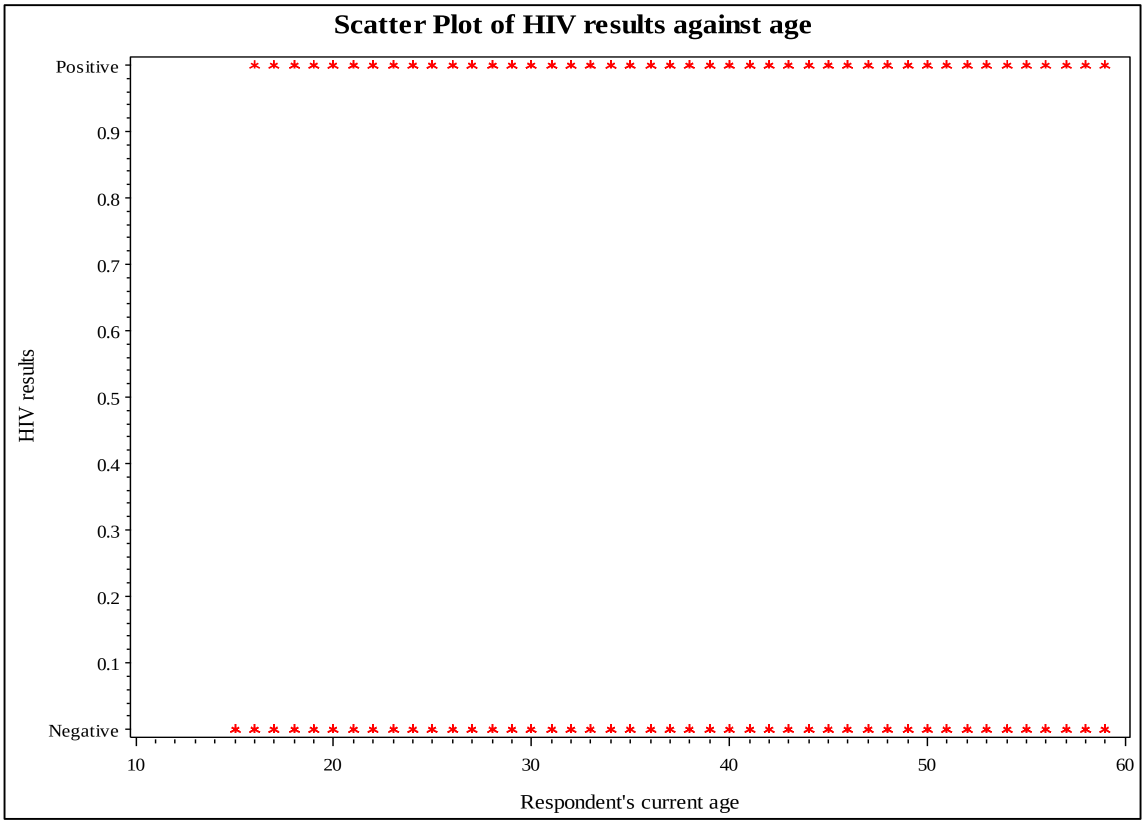


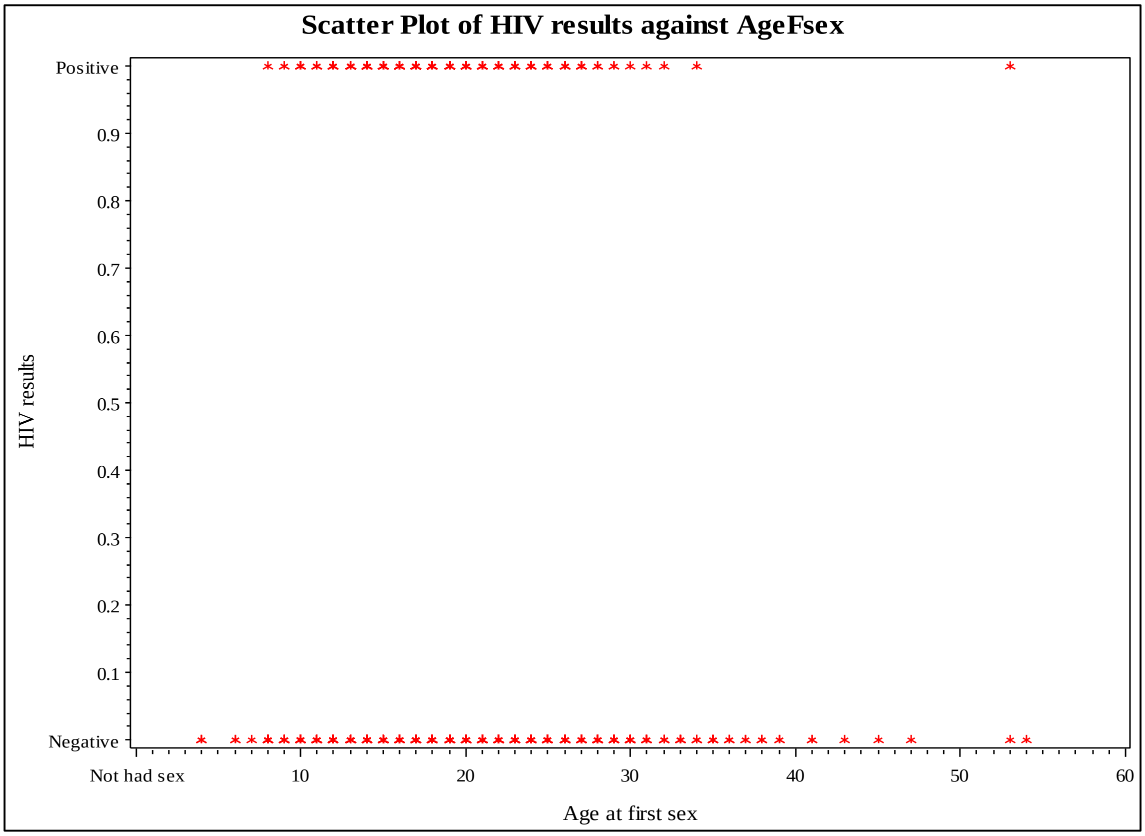

Estimated probability of HIV infection by HIV testing history over age at 1st sex
Estimated Probability
0.25
0.24
0.23
0.22
0.21
0.20
0.19
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Not had sex
10
20
30
40
50
60
Age at first sex
Ever been tested for HIV
No
Yes


Estimated probability of HIV infection by gender of household head over Education
Estimated Probability
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
No education
Primary
Secondary
Higher
Highest educational level
Gender of household head
Male
Female



Appendix C: Checking for complete separation




Scatter Plot of HIV results against AgeFsex
HIV results
Positive
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Negative
Not had sex
10
20
30
40
50
60
Age at first sex

Appendix D: Patterns of missing data

Missing Data Patterns													
Group	HIVRisk	AgeGrp	Marital	AgeFsex_C	Residence	CoercedSex	DrunkSex	Educ	EverTested	HHDGender	PaidSex	STI	TotPartners_C
1	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X	.
3	X	X	X	.	X	X	X	X	X	X	X	X	X
4	.	X	X	X	X	X	X	X	X	X	X	X	X
5	.	X	X	X	X	X	X	X	X	X	X	X	.
6	.	X	X	.	X	X	X	X	X	X	X	X	X

Missing Data Patterns												
					Group Means							
Group	ChildAway_C	WealthIndex	Freq	Percent	HIVRisk	AgeGrp	Marital	AgeFsex_C	Residence	CoercedSex	DrunkSex	Educ	EverTested
1	X	X	14982	81.47	1.689094	3.054265	1.045588	0.443532	1.805033	0.029636	0.203978	1.234081	0.649913
2	X	X	367	2.00	1.716621	3.784741	1.100817	0.427793	1.719346	0.024523	0.264305	1.337875	0.523161
3	X	X	58	0.32	1.724138	2.603448	1.362069	.	1.758621	0.034483	0.155172	1.120690	0.620690
4	X	X	2811	15.29	.	3.205621	1.195304	0.464959	1.825685	0.028815	0.241551	1.070082	0.658840
5	X	X	149	0.81	.	3.691275	1.302013	0.436242	1.738255	0.026846	0.355705	1.234899	0.604027
6	X	X	22	0.12	.	2.727273	1.272727	.	1.727273	0	0.181818	1.090909	0.681818

Missing Data Patterns						
	Group Means					
Group	HHDGender	PaidSex	STI	TotPartners_C	ChildAway_C	WealthIndex
1	1.266920	0.026565	0.268656	1.063810	0.765986	3.087105
2	1.158038	0.054496	0.277929	.	1.411444	3.561308
3	1.517241	0.034483	0.327586	0.517241	0.396552	2.948276
4	1.315902	0.029171	0.317325	1.074351	0.803629	2.889363
5	1.201342	0.174497	0.308725	.	1.087248	3.120805
6	1.409091	0	0.227273	0.909091	0.454545	3.090909

REFERENCES

Agresti, A. (2007). *An introduction to categorical data analysis* (Second Edition ed.). New Jersey: John Wiley & Sons, Inc.

Archer, K. J., Lemeshow, S., & Hosmer, D. W. (2006). Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design . *Computational Statistics & Data Analysis,* , 4450-4464. doi: 10.1016

Auvert, B., Taljaard, D., Lagarde, E., Sobngwi-Tambekou, J., Sitta, R., & Puren, A. (2005). Randomized, Controlled Intervention Trial Of Male Circumcision For Reduction Of Hiv Infection Risk: The Anrs 1265 Trial. *Plos Medicine, 2*(11), E298.

Bailey, R. C., Moses, S., Parker, C. B., Agot, K., Maclean, I., Krieger, J. N., Et Al. (2007). Male Circumcision For Hiv Prevention In Young Men In Kisumu, Kenya: A Randomised Controlled Trial. *Lancet (London, England), 369*(9562), 643-656.

Bankole, A., Ahmed, F. H., Neema, S., Ouedraogo, C., & Konyani, S. (2007). Knowledge of correct condom use and consistency of use among adolescents in four countries in sub-saharan africa. *African Journal of Reproductive Health, 11*(3), 197-220.

Baral, S., Trapence, G., Motimedi, F., Umar, E., Iipinge, S., Dausab, F., & Beyrer, C. (2009). HIV prevalence, risks for HIV infection, and human rights among men who have sex with men (MSM) in malawi, namibia, and botswana. *PloS One, 4*(3), e4997. doi:10.1371/journal.pone.0004997 [doi]

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review, 50*, 279-292.

Bolton, R. (1992). AIDS and promiscuity: Muddles in the models of HIV prevention. *Medical Anthropology, 14*(2-4), 145-223. doi:10.1080/01459740.1992.9966072 [doi]

Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine, 3*, 17-0473-3-17. doi:10.1186/1751-0473-3-17 [doi]

Cassell, M. M., Halperin, D. T., Shelton, J. D., & Stanton, D. (2006). Risk compensation: The achilles' heel of innovations in HIV prevention? *BMJ (Clinical Research Ed.), 332*(7541), 605-607. doi:332/7541/605 [pii]

CDC. (1981). *Pneumocystis pneumonia.* (No. 30). Los Angeles: MMWR.

CDC[Internet]. (2014). About HIV/AIDS. Retrieved from http://www.cdc.gov/hiv/basics/whatishiv.html

Cederbaum, J. A., Gilreath, T. D., & Barman-Adhikari, A. (2014). Perceived risk and condom use among adolescents in sub-saharan africa: A latent class analysis. *African Journal of Reproductive Health, 18*(4), 26-33.

Chambers, R. L., & Skinner, C. J. (2003). *Analysis of survey data*. Sussex, England: John Wiley & Sons Ltd.

Chimoyi, L. A., & Musenge, E. (2014). Spatial analysis of factors associated with HIV infection among young people in uganda, 2011. *BMC Public Health, 14*, 555-2458-14-555. doi:10.1186/1471-2458-14-555 [doi]

Cohen, J., & Tate, T. (2006). The less they know, the better: Abstinence-only HIV/AIDS programs in uganda. *Reproductive Health Matters, 14*(28), 174-178.

Cohen, M. S., Hellmann, N., Levy, J. A., DeCock, K., & Lange, J. (2008). The spread, treatment, and prevention of HIV-1: Evolution of a global pandemic. *The Journal of Clinical Investigation, 118*(4), 1244-1254. doi:10.1172/JCI34706 [doi]

Eaton, L. A., & Kalichman, S. (2007). Risk compensation in HIV prevention: Implications for vaccines, microbicides, and other biomedical HIV prevention technologies. *Current HIV/AIDS Reports, 4*(4), 165-172.

Epstein, H. (2008). *The invisible cure* (First ed.). New York: Farrar, Straus and Giroux.

Gray, P. B. (2004). HIV and islam: Is HIV prevalence lower among muslims? *Social Science & Medicine (1982), 58*(9), 1751-1756. doi:10.1016/S0277-9536(03)00367-8 [doi]

Gray, R. H., Kigozi, G., Serwadda, D., Makumbi, F., Watya, S., Nalugoda, F., . . . Wawer, M. J. (2007). Male circumcision for HIV prevention in men in rakai, uganda: A randomised trial. *Lancet, 369*(9562), 657-666. doi:S0140-6736(07)60313-4 [pii]

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis*. Florida: Chapman & Hall/CRC.

Hladik, W., Barker, J., Ssenkusu, J. M., Opio, A., Tappero, J. W., Hakim, A., . . . Crane Survey Group. (2012). HIV infection among men who have sex with men in kampala, uganda--a respondent driven sampling survey. *PloS One, 7*(5), e38143. doi:10.1371/journal.pone.0038143 [doi]

Hosmer, D. W., & Lemeshow, S. (2013). *Applied logistic regression* (Third Edition ed.). New Jersey: John Wiley & Sons, Inc.

Hrdy, D. B. (1987). Cultural practices contributing to the transmission of human immunodeficiency virus in africa. *Reviews of Infectious Diseases, 9*(6), 1109-1119.

Institute of Medicine (US). (2006). Committee on Assessing interactions among social, behavioral, and genetic factors in health; doi:NBK19929 [bookaccession]

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York: Springer.

Janier, M., Perroud, A. M., Revuz, J., Wechsler, J., Feuilhade, M., Poirier, J., . . . Touraine, R. (1984). Acquired immunodeficiency syndrome, kaposi's disease and cerebral toxoplasmosis in a young man. review of the literature apropos of a case. [Syndrome d'immunodeficience acquise, maladie de Kaposi et toxoplasmose cerebrale chez un homme jeune. Revue de la litterature a propos d'un cas] *Annales De Dermatologie Et De Venereologie, 111*(1), 11-23.

Kamarulzaman, A. (2013). Fighting the HIV epidemic in the islamic world. *Lancet, 381*(9883), 2058-2060. doi:10.1016/S0140-6736(13)61033-8 [doi]

Kibira, S. P., Nansubuga, E., Tumwesigye, N. M., Atuyambe, L. M., & Makumbi, F. (2014). Differences in risky sexual behaviors and HIV prevalence of circumcised and uncircumcised men in uganda: Evidence from a 2011 cross-sectional national survey. *Reproductive Health, 11*(1), 25-4755-11-25. doi:10.1186/1742-4755-11-25 [doi]

Kiwanuka, N., Ssetaala, A., Nalutaaya, A., Mpendo, J., Wambuzi, M., Nanvubya, A., . . . Sewankambo, N. K. (2014). High incidence of HIV-1 infection in a general population of fishing communities around lake victoria, uganda. *PloS One, 9*(5), e94932. doi:10.1371/journal.pone.0094932 [doi]

Koh, K. C., & Yong, L. S. (2014). HIV risk perception, sexual behavior, and HIV prevalence among men-who-have-sex-with-men at a community-based voluntary counseling and testing center in kuala lumpur, malaysia. *Interdisciplinary Perspectives on Infectious Diseases, 2014*, 236240. doi:10.1155/2014/236240 [doi]

Konde-Lule, J. K., Wawer, M. J., Sewankambo, N. K., Serwadda, D., Kelly, R., Li, C., . . . Kigongo, D. (1997). Adolescents, sexual behaviour and HIV-1 in rural rakai district, uganda. *AIDS (London, England), 11*(6), 791-799.

Lema, L. A., Katapa, R. S., & Musa, A. S. (2008). Knowledge on HIV/AIDS and sexual behaviour among youths in kibaha district, tanzania. *Tanzania Journal of Health Research, 10*(2), 79-83.

Manda, S. O., Lombard, C. J., & Mosala, T. (2012). Divergent spatial patterns in the prevalence of the human immunodeficiency virus (HIV) and syphilis in south african pregnant women. *Geospatial Health, 6*(2), 221-231.

Matovu, J., & Ssebadduka, N. (2013). Knowledge, attitudes & barriers to condom use among female sex workers and truck drivers in uganda: A mixed-methods study. *African Health Sciences, 13*(4), 1027-1033. doi:10.4314/ahs.v13i4.24 [doi]

Maughan-Brown, B., & Venkataramani, A. S. (2012). Learning that circumcision is protective against HIV: Risk compensation among men and women in cape town, south africa. *PloS One, 7*(7), e40753. doi:10.1371/journal.pone.0040753 [doi]

MOH, ICF International, CDC, USAID, & WHO. (2012). *Uganda AIDS indicator survey 2011. calverton, maryland, USA: MOH and macro international inc.* ().

Molenberghs, G., & Kenward, M. (2007). *Missing data in clinical studies*. West Sussex, England: John Wiley & Sons, Ltd.

Murphy, E. M., Greene, M. E., Mihailovic, A., & Olupot-Olupot, P. (2006). Was the "ABC" approach (abstinence, being faithful, using condoms) responsible for uganda's decline in HIV? *PLoS Medicine, 3*(9), e379. doi:05-PLME-PMD-0444R1-APPEAL [pii]

Ntozi, J. P., Najjumba, I. M., Ahimbisibwe, F., Ayiga, N., & Odwee, J. (2003). Has the HIV/AIDS epidemic changed sexual behaviour of high risk groups in uganda? *African Health Sciences, 3*(3), 107-116.

Pigott, T., D. (2001). A review of methods for missing data. *Educational Research and Evaluation, 7*, 353-383.

Piper, M. E., Loh, W. Y., Smith, S. S., Japuntich, S. J., & Baker, T. B. (2011). Using decision tree analysis to identify risk factors for relapse to smoking. *Substance use & Misuse, 46*(4), 492-510. doi:10.3109/10826081003682222 [doi]

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology* (Third Edition ed.). Philadelphia: Lippincott Williams & Wilkins.

Rothman, J., K. (2012). *Epidemiology: An Introduction* (2 edition ed.) Oxford University Press.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc.

SAS Institute. (2011). *SAS/STAT ® 9.3 user's guide.* (). Cary, NC: SAS Publishing.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Serwadda, D., Mugerwa, R. D., Sewankambo, N. K., Lwegaba, A., Carswell, J. W., Kirya, G. B., . . . Clayden, S. A. (1985). Slim disease: A new disease in uganda and its association with HTLV-III infection. *Lancet, 2*(8460), 849-852. doi:S0140-6736(85)90122-9 [pii]

Serwadda, D., Wawer, M. J., Musgrave, S. D., Sewankambo, N. K., Kaplan, J. E., & Gray, R. H. (1992). HIV risk factors in three geographic strata of rural rakai district, uganda. *AIDS (London, England), 6*(9), 983-989.

Sharp, P. M., Bailes, E., Chaudhuri, R. R., Rodenburg, C. M., Santiago, M. O., & Hahn, B. H. (2001). The origins of acquired immune deficiency syndrome viruses: Where and when? *Philosophical Transactions of the Royal Society of London.Series B, Biological Sciences, 356*(1410), 867-876. doi:10.1098/rstb.2001.0863 [doi]

Sharp, P. M., & Hahn, B. H. (2011). Origins of HIV and the AIDS pandemic. *Cold Spring Harbor Perspectives in Medicine, 1*(1), a006841. doi:10.1101/cshperspect.a006841 [doi]

Skinner, C., & Vallet, L., A. (2010). Fitting log-linear models to contingency tables from surveys with complex sampling designs: An investigation of the clogg-eliason approach. *Sociological Methods Research, 39:1 83-108*

Smolak, A. (2010). Contextual factors influencing HIV risk behaviour in central asia. *Culture, Health & Sexuality, 12*(5), 515-527. doi:10.1080/13691051003658135 [doi]

Speiser, J. L., Lee, W. M., Karvellas, C. J., & US Acute Liver Failure Study Group. (2015). Predicting outcome on admission and post-admission for acetaminophen-induced acute liver failure using classification and regression tree models. *PloS One, 10*(4), e0122929. doi:10.1371/journal.pone.0122929 [doi]

Suresh, K., & Chandrashekara, S. (2012). Sample size estimation and power analysis for clinical research studies. *Journal of Human Reproductive Sciences, 5*(1), 7-13. doi:10.4103/0974-1208.97779 [doi]

Szklo, M., & Nieto, F. J. (2000). *Epidemiology beyond the basic*. Maryland: Aspen.

The DHS Program. (2012). The uganda AIDS indicator survey. Retrieved from http://dhsprogram.com/publications/publication-AIS10-AIS-Final-Reports.cfm

Tumwesigye, N. M., Atuyambe, L., Wanyenze, R. K., Kibira, S. P., Li, Q., Wabwire-Mangen, F., & Wagner, G. (2012). Alcohol consumption and risky sexual behaviour in the fishing communities: Evidence from two fish landing sites on lake victoria in uganda. *BMC Public Health, 12*, 1069-2458-12-1069. doi:10.1186/1471-2458-12-1069 [doi]

Uganda AIDS commission (UAC). *Global AIDS response progress report: Uganda. 2012.* Kampala

Uganda Bureau of Statistics (UBOS) [Internet]. (2013). Uganda in figures. Retrieved from http://www.ubos.org/onlinefiles/uploads/ubos/pdf%20documents/Facts%20and%20Figures%202013.pdf

Uganda Bureau of Statistics (UBOS) and Macro International Inc. (2012). *Uganda demographic and health survey 2011. calverton, maryland, USA: UBOS and macro international inc.* ().

UNAIDS. (2012). *UNAIDS report on the global AIDS epidemIc.* ().

von Hippel, P., T. (2009). How to impute interactions, squares and other transformed variables. *Sociological Methodological, 39*, 265-291. doi:10.1111

Walusaga, H. A., Kyohangirwe, R., & Wagner, G. J. (2012). Gender differences in determinants of condom use among HIV clients in uganda. *AIDS Patient Care and STDs, 26*(11), 694-699. doi:10.1089/apc.2012.0208 [doi]

Wawer, M. J., Serwadda, D., Musgrave, S. D., Konde-Lule, J. K., Musagara, M., & Sewankambo, N. K. (1991). Dynamics of spread of HIV-I infection in a rural district of uganda. *BMJ (Clinical Research Ed.), 303*(6813), 1303-1306.

Weiss, H. A., Hankins, C. A., & Dickson, K. (2009). Male circumcision and risk of HIV infection in women: A systematic review and meta-analysis. *The Lancet Infectious Diseases, 9*(11), 669-677. doi:10.1016/S1473-3099(09)70235-X [doi]

Wei-Yin, L. (2011). Classification and regression trees. *2011 John Wiley & Sons, Inc., 1*, 14-23.

Wertheim, J. O., & Worobey, M. (2009). Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. *PLoS Computational Biology, 5*(5), e1000377. doi:10.1371/journal.pcbi.1000377 [doi]

Westercamp, N., Agot, K., Jaoko, W., & Bailey, R. C. (2014). Risk compensation following male circumcision: Results from a two-year prospective cohort study of recently circumcised and uncircumcised men in nyanza province, kenya. *AIDS and Behavior, 18*(9), 1764-1775. doi:10.1007/s10461-014-0846-4 [doi]

WHO, UNICEF, & UNAIDS. (2011). *Global HIV/AIDS response, progress report. 2011*. Geneva: World Health Organization.