2011
2012

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Statistical Methodology for HIV Serodiscordance among Couples: The case of Mozambique

Promotor :
Prof. dr. Marc AERTS
Prof. dr. Niel HENS

Adelino Jose C Juga

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
KNOWLEDGE IN ACTION

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt

2011
2012

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Statistical Methodology for HIV Serodiscordance among Couples: The case of Mozambique

Promotor :
Prof. dr. Marc AERTS
Prof. dr. Niel HENS

Adelino Jose C Juga

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
KNOWLEDGE IN ACTION

Certification

This is to certify that this report was written by **Adelino José Chingore Juga** under our supervision.

Signiture: _____ Date: _____

Prof. Dr. Marc Aerts **Internal Supervisor**

Signiture: _____ Date: _____

Prof. Dr. Niel Hens **Internal Supervisor**

Thesis submitted in partial fulfilment of the requirements for the degree of Master of Statistics: Biostatistics.

Signiture: _____ Date: _____

Adelino José Chingore Juga **Student**

Abstract

Even though several promising global efforts have been made in increasing effective treatment and prevention programs, the number of people living with human immuno deficiency virus is still high. Across countries including Mozambique, a substantial proportion of couples with human immuno deficiency virus infection is discordant. Hence the human immuno deficiency virus prevalence of serodiscordance among heterosexual couples, who are often in stable partnerships but unaware of both partner's serostatus is high. In this study risk factors associated with serodiscordance among couples in Mozambique was investigated. Cross-sectional data based on a national-representative sample in Mozambique was used. Several statistical models such as Alternating Logistic Regression and Generalized Linear Mixed Model (univariate binary outcome), Baseline Category Logit and Generalized Linear Mixed Model (multicategory outcome), Alternating Logistic Regression, Bivariate Dale Model, Generalized Linear Mixed Model (bivariate binary outcomes) were applied motived by the nature of outcomes and by the design of the study. Statistical findings revealed that prevalence, sexual transmission infectious disease, union number and wealth index were risk factor associated human immuno deficiency virus serodiscordance in Mozambique. Moreover, men were at higher risk of human immuno deficiency virus infection than the women, hence couples were more likely to be male than female discordant. Heterogeneity or variation between couples, Enumeration Areas as well as between provinces in probability of couples being Human Immuno Deficiency Virus positive was observed. Furthermore, there is positive and strong association between the two serostatus (woman and man) in sense that when one member within a couple is human immuno deficiency virus positive another member was at high risk acquiring the human immuno deficiency virus infection.

Keywords: Alternating Logistic Regression, Bivariate Dale Model, Baseline Category Logit, Generalized Linear Mixed Model, Human Immuno Deficiency Virus, Serodiscordance.

Acknowledgements

First of all, I dedicate this work to LORD JESUS CHRIST and I thank Him for giving me life, health, wisdom and strength to complete successfully this Master program. I also thank all my brothers in Christ that days and nights remember me in their prayers.

Successful completion of this thesis has been through the support of a number of individuals. First of all, I am very grateful to my supervisors; Prof. Dr. Marc Aerts and Prof. Dr. Niel Hens for their valuable and continuous the guidance, by supplying thoughtful suggestions for improving this report. In particular, I thank the Health Ministry of Mozambique (MISAU) and the DHS project for providing me the data for this Master thesis.

A word of gratitude to all those involved in the Desafio Program (Mozambique and Belgium) in special the project leaders of Biostatistics and Modelling (P7), Rafica Abdulrazac, MSc and Prof. Dr. Marc Aerts. In particular, I thank the Flemish Interuniversity Council (VLIR) for the scholarship which has enabled me to pursue this valuable Master program. Not forgetting Annick Verheylezoon, Martine Michaels and Katrien Sterken.

Special thanks to all my lecturers for valuable teaching of Statistics, also to my colleagues in the Master of Statistics 2010-2012 for providing an intellectually environment during the program. Osvaldo Loquiha, Hélio Langa and António Rungo, thanks for your friendship and encouragement.

Finally thanks to all my family, in special to Graça (wife), Raquel (daughter), Isabel (mother), Jacinta (sister), Sérgio (brother), Gracinda (sister) and António (brother) for their support and patience during this 2 years of the Master program.

Adelino Juga
University of Hasselt
Belgium, September, 2012

List of Abbreviations

AIDS - Acquired Immuno Deficiency Syndrome

ALR - Alternating Logistic Regression

BCL - Baseline Category Logit

BDM - Bivariate Dale Model

BPM - Bivariate Probit Model

EA - Enumeration Area

GEE - Generalized Estimating Equations

GLM - Generalized Linear Models

GLMM - Generalized Linear Mixed Model

HIV- Human Immuno Deficiency Virus

HCV - Hepatitis C virus

INSIDA - Inquérito Nacional de Prevalência, Riscos Comportamentais e Informação sobre o HIV e SIDA

PLHIV - People living with HIV

PSU - Primary Sampling Unit

LR - Likelihood ratio

OR - Odds Ratios

STID - Sexually Transmitted Infection Disease

SSU - Secondary Sampling Unit

SAA - Sub-Saharan Africa

SE - Standard Error

TSU - Tertiary Sampling Units

WHO - World Health Organization

GLOSSARY

Binomial: Having two possible values, e.g., a variable with two categories

Cohabitation: Living together as if a married couple

Concordant: Both members of a couple having the same HIV status

Concordant negative: Both members of a couple being HIV-negative

Concordant positive: Both members of a couple being HIV-positive

Discordant: Two members of a couple having different HIV-status; one is HIV-positive while the other is HIV-negative

Female discordant: A couple in which the woman is HIV-positive and the man is HIV-negative

Male discordant: A couple in which the man is HIV-positive and the woman is HIV-negative

Multinomial: Having several possible values, e.g., a variable with three categories

Polygamy: A type of marital union in which one man has two or more wives

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

Since the Human Immuno Deficiency Virus (HIV) was first diagnosed, it has infected around 65 million individuals worldwide. Even though several promising global efforts have been made in increasing effective treatment and prevention programs, the number of people living with HIV (PLHIV) is still high and more than 25 million of people have already died due to the Acquired Immuno Deficiency Syndrome (AIDS). In Sub-Saharan Africa (SAA) region, parts of Asia, Central America and the Caribbean, HIV has become a 'generalized' epidemic, since more than 1% of the population are HIV-positive [22].

Out of approximately 40 million PLHIV, a third (12.35 million) is living in the ten countries of SAA region, thus is the region most affected by the epidemic, consequently the epicentre of the HIV and AIDS pandemic worldwide. HIV/AIDS has had a profound impact on economic growth, income and poverty in this region. In some countries, the economy was projected to shrink by 30% in 2010 due to HIV/AIDS. At household level, HIV/AIDS increases economic vulnerability due to additional costs of care, treatment and support as well as funerals. In addition, the life expectancy at birth has dropped to below 40 years in many SAA countries, along with a negative socio-economic impact due to a decline in the productive work force, and increase in dependency ratios [6].

HIV/AIDS is increasingly feminized, since there are more than twice as many young adult women infected than men. It has been reported that the major determinant of HIV spread is gender inequality linked to risky livelihoods, forced mobility for economic reasons in a context of food insecurity and increasing impoverishment. High levels of sexual and gender-based violence and exploitation including intergenerational sex have been observed. Weakened education, health, administration and other public services imply that poor, vulnerable people have reduced access to essential support [6].

Estimates of HIV transmission rates in African discordant couples who do not know their HIV statuses range from 20% to 25% per year. HIV counselling and testing specially adapted to couples including education about HIV discordance rather than an individual focused HIV risk assessment and counselling could help to reduce HIV transmission. This underscores that stable HIV discordant African couples are a critical target for counselling and testing and for the evaluation of new prevention interventions [6]. HIV negative individuals in discordant relationships remain an especially vulnerable population for acquiring the virus.

The risk of transmission per coital act differs from couple to couple depending on their characteristic and within a couple the risk of transmission can fluctuate over time. Evidence from

literature suggests that there are several factors that influence the HIV risk infection within couples [19]. In the SAA region, heterosexual exposure is the main mode of HIV transmission from infected woman to unaffected man or vice versa [13], and it involves a complex interaction between biologic and behavioural factors. A HIV negative partner in a discordant couple can become HIV positive through sexual acquisition from outside the marriage. Non-spouse partner was observed as primary risk factor associated with this type of acquisition, and this risk could be mediated through condom use [12]. Another way of infection from an HIV negative partner in a discordant couple to HIV positive was through nonsexual transmission, such as unsafe medical injections, surgery, blood transfusions, or tattoo or scarification procedures.

Despite the empirical evidence pointing to their programmatic importance, serodiscordant couples are often overlooked or, at best, only vaguely addressed in many national prevention plans. This omission may stem not only from the sensitivity surrounding HIV within couples but also from misperceptions about the extent of serodiscordance and failure to understand that it is possible to prevent transmission within a stable union once one partner has become infected [9].

In Mozambique, the proportion of couples where the man is HIV positive and the woman is HIV negative (male discordant couples) is similar to the proportion of couples in which the woman is HIV positive and the man is HIV negative (female discordant couples). Estimates from INSIDA pointed that 85% of all couples were concordant negative, i.e., both HIV negative. 15% of cohabiting couples were affected by HIV, either one or both members were HIV positive. In 5% of couples, both members were HIV positive (concordant positive), while in 10% of couples, one member was HIV positive whilst the other were HIV negative (discordant) [12]. Specifically there were approximately 433,000 discordant couples in 2009.

Furthermore, these estimates showed that both members of couples have been tested for HIV in 11% of all couples and 15% were discordant couples. This means that there were at least 368,000 couples in Mozambique who were discordant but do not know it. According to the estimated rates of HIV transmission within discordant couples published in the scientific literature, this population was at high risk for new HIV infections [8]. Risk factors associated with HIV serodiscordance in Mozambique are not well-studied and documented, it is against this background that this study aims to:

- Investigate a relationship between the HIV status of woman and man within a couple;
- Investigate risk factors associated with the HIV status of woman and man;
- Investigate whether the association of the HIV status of woman and man depends on certain factors, i.e., does the relationship change in certain subgroups;
- Investigate risk factors associated with HIV serodiscordance among couples.

The remainder of this report is organized as follow; The study design and data used are introduced in Section 2. In Section 3, the statistical methodology is explained while the application to INSIDA data is presented in Section 4. The discussions and conclusions are given in Section 5 Sample weight calculations are presented in the appendix.

1.2 Literature Review

As the HIV/AIDS epidemic has matured in many countries including Mozambique, it is believed that the proportion of new infections occurring within couples has risen. In recent years, there has been increasing interest in how HIV is spread within stable sexual partnerships or couples. Across countries, a substantial proportion of couples with any HIV infection is discordant [10]. In Africa, several studies pointed at a high prevalence of HIV serodiscordance among heterosexual couples, who are often in stable partnerships but unaware of both partner's HIV status [13]. In most of the studies that have been conducted on HIV serodiscordance, bivariate analysis, multivariate logistic regression for binary and multinomial response has been applied [11], [19]. However in Zambia and Rwanda one mathematical model for adults was developed and predicted that 55.1% to 92.7% of new heterosexually acquired HIV infections could occur within cohabiting discordant couples [7].

In Mozambique a few studies have been conducted on HIV serodiscordance among couples. Among these include national demograph surveys in wich some information about HIV is collected and one national survey so called INSIDA, which carried out in 2009. The INSIDA 2009 collected information on HIV serostatus, risk behaviors, and other background characteristics, allowing cohabitating couples to be matched and analyzed together. This survey had two main objectives: (1) to estimate the number of discordant couples and to provide useful information about these couples, and (2) to identify risk factors that could help to protect HIV negative partner from becoming infected within a marital relationship. An HIV test was performed in both couples'members, thus two outcomes from the same couple were observed and this couple could be classified as concordant negative, female discordant, male discordant or concordant positive. In their analysis they (INSIDA 2009) applied bivariate associations and two (for binary and multicategory responses) multivariate logistic regression models. In a logistic regression model with binary responses they compare concordant positive couples with all discordant couples, regardless of whether the discordant couple is male or female, while in a multiple-category model comparison between concordant positive couples and male as well as female discordant couples was performed separately [8].

In literature there are many approaches to model two outcomes from the same subject with respect to certain covariates, while accounting for dependence between both outcomes, such as the INSIDA 2009 data used in this report. These include the bivariate dale model (BDM) and

bivariate probit model (BPM) which are the two most often used [11], [17]. BDM was applied to joint modeling the HVC and HIV co-infection among injecting drug users in Italy and Spain using individual cross-sectional data [5] as well as in estimation of new joint and conditional epidemiological parameters for multisera data [11]. For the INSIDA data, the BDM can model the marginal joint distribution of the HIV status of women and men. Parameter estimates are expressed in log-odds ratios and are interpreted in exactly the same manner as ordinary logistic regression. Futhermore, the BDM allows one to infer the association between HIV status of woman and man, and to model variation in this association as a function of covariates [15].

For multicategory outcomes, several models have also been developed. The most often used are the baseline category logit (BCL) model (for nominal response), the cumulative logit models, the adjacent-categories logits, the continuation-ratio logits (for ordinal response) [1], the Generalized Estimating Equations(GEE) for clustered data. In case of the INSIDA 2009 data, the GEE is most suitable, since the data are clustered which allows to model the marginal probability of couple being female discordant compared to male discordant. In addition to that, other models allows to take into account the heterogeneity present in data including random effects since couples are nested within household, households within Enumeration Area (EA) and final EAs within provinces.

2 Study Design and Data

2.1 Study Design

The data used in this report results from a cross-sectional study, based on a national-representative sample of individuals of the INSIDA survey in Mozambique carried out in 2009. One of the outcomes of interest was the HIV status of cohabiting couples. Cohabiting couples refers to those couples for which both partners were present at the time of the survey. Each identified the other as husband/wife or living together as husband and wife, and consented to an HIV test for which results were obtained [8].

Stratification and clustering were applied to ensure that for each province inference was possible with nearly the same precision. Moreover, a two-stage sampling method was also applied to access individuals within households. EAs, households and individuals were Primary Sampling Units (PSU), Secondary Sampling Units (SSU) and Tertiary Sampling Units (TSU) respectively. The 11 provinces were considered as stratums. In stage one, 270 EAs were selected in all provinces from a total of 45000 EAs defined according to the cartography of general census, 2007 (See figure 1). Out of the selected 270 EAs, 122 were urban and 148 rural. Then a fixed number of households were systematically selected within each EA in the second stage. A total of 22 household from urban EAs and 24 from rural EAs were selected [8].

A Household questionnaire was administered to every selected household, which included a complete list of household individuals and their respective age. Each individual listed in the household questionnaire was assigned a unique household line number within a household. Men and women aged between 15-64 years were eligible to participate in an individual interview and to provide a blood sample for the HIV test. During the individual interviews, respondents were asked whether they were married or not. If so and if the husband/wife was named in the household questionnaire, the household line number of the husband/wife was recorded in the individual's questionnaire. To identify couples, each woman that was married or who lived with her husband, an attempt was made to match her with her husband/partner using his household line number. A confirmation was then made by checking the man's interview information that was named by the woman. If a man was polygamy (if has more than one wife) he may appear in the database multiple times, one for each wife [8].

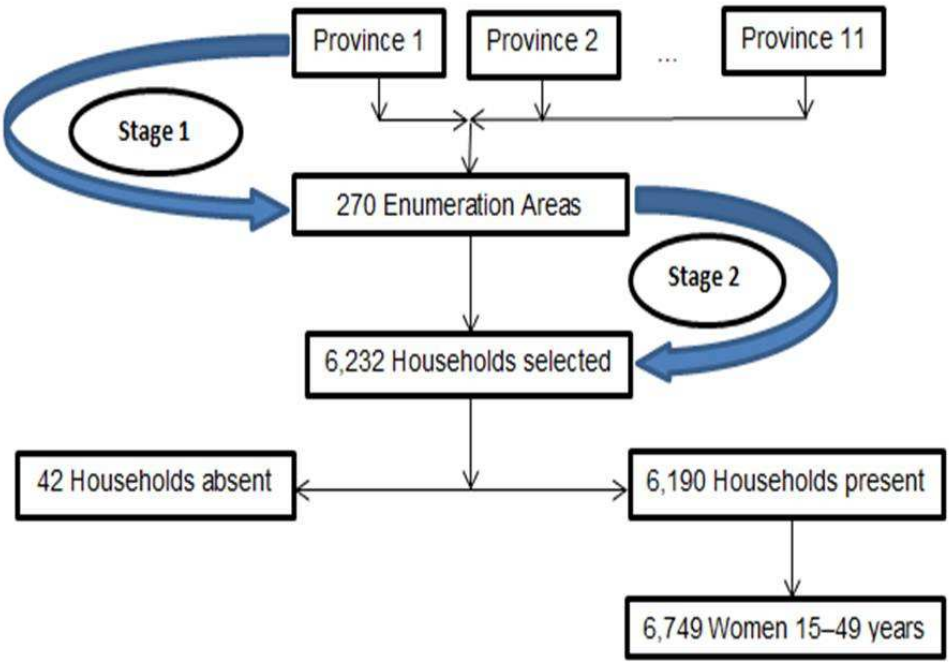

Figure 1: Survey Design

2.2 Data

In this report, a couple used as the unit of analysis; however it was not the unit of selection in the sample. Out of 2639 wife and 2490 husband tested for HIV (Figure 2), 2466 couples were successfully matched; hence their information was used in this report. With regards to member of wives, 2731(86.73%) of men declared that they had only one wife, 358(11.37%) had one extra wife, 23(0.73%) had two extra wives and only 9(0.29%) had three wives. On other hand, 10(0.16%) out of selected 6190 households had two couples and only one household had three couples.

Samples weights were obtained and used to ensure the actual representativeness of the sample at the national level, since the allocation of the sample to the different provinces and to their urban-rural areas was not proportional. More details about sample weights calculation can be found in the appendix.

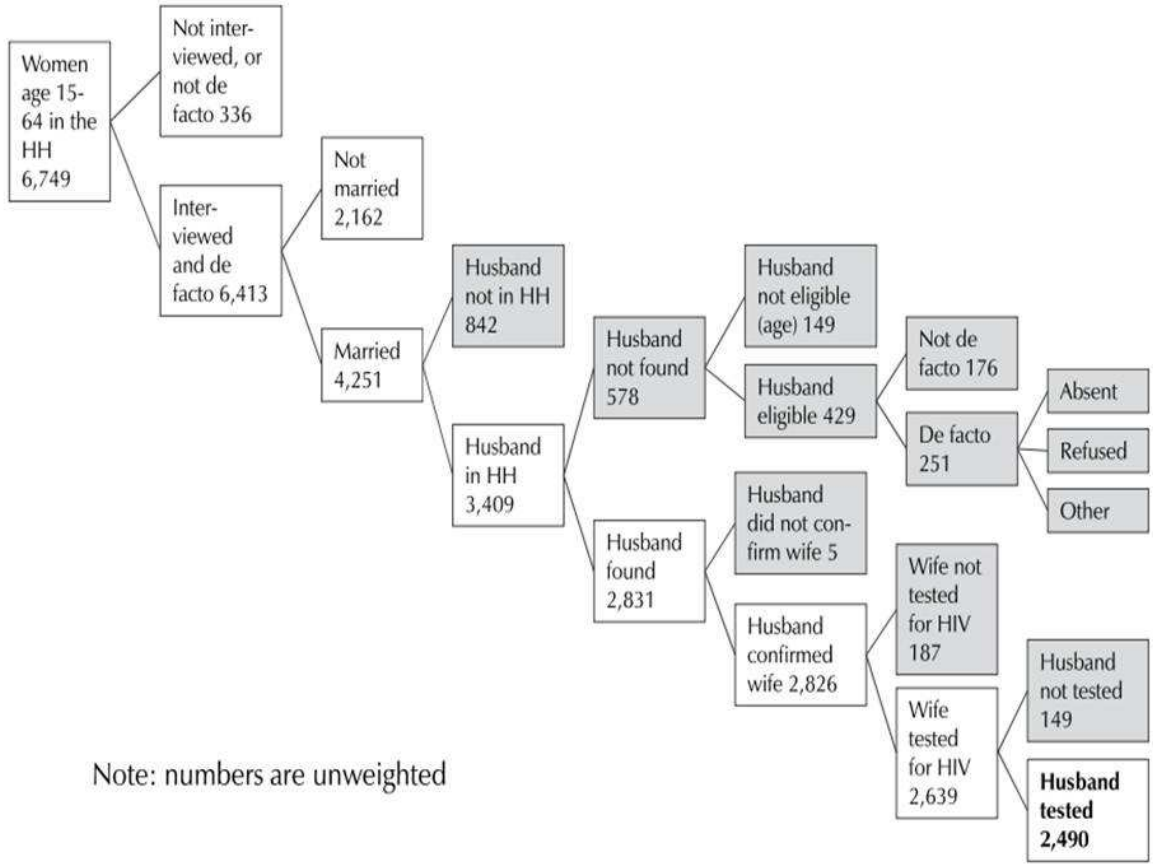

Figure 2: Composition of Couple File [8]

2.3 Descriptions of Variables

The HIV statuses of each member within a couple are the two first outcomes of interest. The HIV status of each member was dichotomized in binary response (Y_1, Y_2), since the HIV results had 4 categories (HIV negative, infected by HIV1, infected by HIV2, co-infection of the HIV1 and HIV2). From the two individual HIV statuses (Y_1, Y_2), three additional responses were derived. Y_3 was restricted to only positive couples, which is binary was created, where zero (0) refers to a concordant positive couple and one (1) to a discordant couple (female or male discordant). Furthermore, a multicategorical variable Y_4 was created; zero (0) refers to a concordant negative couple, (1) to a female discordant couple, (2) to a male discordant and (3) to a concordant positive couple. Finally a multicategorical variable Y_5 was as well obtained restricted only to positive couples where zero (0) refers to a female discordant couple, (1) to a male discordant and (2) to a concordant positive (See Table 1).

Table 1: Descriptions of Response Variables

Variable	Description	Categories	Measured level
Y_1	HIV status of Woman	0: HIV Negative	Individual
		1: HIV Positive	
Y_2	HIV status of Man	0: HIV Negative	Individual
		1: HIV Positive	
Y_3	Serodiscordance of couple	0:$(Y_1 = 1; Y_2 = 1)$-Concordant positive	
		1:$(Y_1 = 1; Y_2 = 0)$-Discordant	Couple
		1:$(Y_1 = 0; Y_2 = 1)$-Discordant	
Y_4	Serodiscordance of couple	0:$(Y_1 = 0; Y_2 = 0)$-Concordant Negative	
		1:$(Y_1 = 1; Y_2 = 0)$-Female Discordant	Couple
		2:$(Y_1 = 0; Y_2 = 1)$-Male Discordant	
		3:$(Y_1 = 1; Y_2 = 1)$-Concordant Positive	
Y_5	Serodiscordance of couple	0:$(Y_1 = 1; Y_2 = 0)$-Female Discordant	
		1:$(Y_1 = 0; Y_2 = 1)$-Male Discordant	Couple
		2:$(Y_1 = 1; Y_2 = 1)$-Concordant Positive	

Relevant covariates used in all analyses are presented in Table 2. The variable X_1 is measured at provincial level, X_{22} to X_{24} are at region and residence zone level respectively while the rest corresponds to couple-specific variables.

Table 2: Descriptions of Covariates

Variable	Description	Categories	Indicator Var.
Prevalence	HIV prevalence of province	1:[< 5%]	Reference
		2:[5% − 15%]	x_1
		3:[> 15%]	x_2
STID	Sexual transmission infectious disease	1:None had STID	Reference
		2:1 member had STID	x_3
HIV Test	HIV Test of each couple	1:None tested	Reference
		2:1 member tested	x_4
Union Number	Number of union of couple	1:Both 1 union	Reference
		2:Woman> 1;Man= 1	x_5
		3:Woman= 1;Man> 1	x_6
		4:Both> 1	x_7
Marital Duration	Marital Duration of couple	1:[0 − 4] years	x_8
		2:[5 − 9] Years	x_9
		3:[10 − 14] Years	x_{10}
		4:[15 − 19] Years	x_{11}
		5:[20 − 24] Years	x_{12}
		6:[25 − 29] Years	Reference
Education level	Education level of couple	1:Woman more educated	x_{13}
		2:Both no education	Reference
		3:Both primary level	x_{14}
		4:Both sec/higher level	x_{15}
		5:Man more educated	x_{16}
Wealth Index	Wealth Index of couple	1:Poorer	x_{17}
		2:Middle	x_{18}
		3:Richer	Reference
Condom used	Condom used with non-spouse/husband	1:Not used	x_{19}
		2:Used	Reference
Age Difference	Age Difference within a couple	1:Woman older	Reference
		2:Man older [0 − 5]Years	x_{20}
		3:Man older > 5 Years	x_{21}
Residence zone	Residence zone of of couple	1:Rural	Reference
		2:Urban	x_{22}
Region	Geographical region of Mozambique	1:North	Reference
		2:Central	x_{23}
		3:South	x_{24}

3 Methodology

This section present methods used to analyse the data in order to answer the research questions. Statistical models for clustered univariate binary and for multicategory responses are concisely described in Section 3.1 and 3.2 respectively while for bivariate binary responses they are described in Section 3.3. Additional hierarchical structures are presented in Section 3.4 and Section 3.5 describes how to account for survey design.

3.1 Models for Clustered Univariate Binary Responses

3.1.1 Alternating Logistic Regression (ALR)

Alternating logistic regression (ALR) was proposed by Carey, Zeger, and Diggle in 1993 [3]. The method is different from other Generalized Estimating Equations (GEE) methods but has similarity with both GEE1 and GEE2 based on odds ratios [17]. The ALR extends beyond classical GEE in sense that the precision of estimates follow for both regression parameters β and the association parameters α. Moreover, with ALR inferences can be made, not only about marginal parameters but about pairwise associations between subjects as well [17]. The odds ratio $OR(Y_{ij}, Y_{ik})$ between the j^{th} and k^{th} observation for the i^{th} cluster is expressed as:

$$OR(Y_{ij}, Y_{ik}) = \frac{P[(Y_{ij} = 1, Y_{ik} = 1)P(Y_{ij} = 0, Y_{ik} = 0)]}{P[(Y_{ij} = 1, Y_{ik} = 0)P(Y_{ij} = 0, Y_{ik} = 1)]}. \tag{1}$$

Let $Y_i = [(Y_{i1}, ..., Y_{im})]'$ be the vector of response on the i^{th} cluster and X a matrix containing covariates associated with the response values. The model that expresses the response as function of covariates can be formulated as in a generalized linear models (GLM):

$$E[Y_i] = \mu_i, \eta(\mu_i) = X_i\beta, \tag{2}$$

where: μ_i is the mean response vector for i^{th} cluster; β is a vector of unknown regression coefficients and $\eta(.)$ the link function. The ALR model with Y_3 as outcome was applied to account for two-level hierarchical structure, couples nested within EAs and EAs within a provinces.

3.1.2 Random Effects Model

A generalized linear mixed model (GLMM) is one of random effects model that is alternative modelling approach for multivariate data that captures individual heterogeneity by conditioning on subject-specific random effects in the model. Hence, this model can be viewed as a conditional model. In general, using the same notation as ALR, GLMM is defined as [17]:

$$E[Y_i|b_i] = \mu_i, \eta(\mu_i) = X_i\beta + Z_ib_i, \tag{3}$$

where: μ_i is the mean response vector for i^{th} cluster conditional on the random effect b_i, X and Z are matrixes of covariates and for the random effects, respectively. β is a vector of unknown regression coefficients. The random effects b_i are assumed to independent and normal

distributed with mean zero and covariance matrix $D, [b_i \sim N(0; D)]$. Equation (3) can be extended to handle two-level random-effects leading to the following model:

$$E[Y_{ij}|b_i, b_{ij}] = \mu_{ij}, \eta(\mu_{ij}) = X_{ij}\beta + Z_{ij}b_i + Z_{ij}b_{ij}, \tag{4}$$

where: the random effects b_i and b_{ij} are assumed to be independently and normal distributed as $b_i \sim N(0; D)$ and $b_{ij} \sim N(0; D)$, respectively, where D can assume different covariance structure, from simple to unstructured. The variance component for b_i refers to the between-cluster and for b_{ij} to the within-cluster variability. As in ALR, two-level hierarchical structure (couples nested within EAs and EAs within provinces) was considered.

To investigate the need for random-effects, likelihood ratio (LR) test was performed in hierarchical way and the corresponding p-value was obtained as follow:

$$p - value = P(\chi^2_{k_1:k_1+1} > -2ln\lambda_N) = 1/2[P(\chi^2_{k_1} > -2ln\lambda_N)] + 1/2[P(\chi^2_{k_1+1} > -2ln\lambda_N)], \tag{5}$$

where: $\chi^2_{k_1:k_1+1}$ is a mixture of two χ^2 distributions with k_1 and $k_1 + 1$ degrees of freedom with equal weight for both distributions, and $-2ln\lambda_N$ is the LR statistic. The use of mixture of χ^2 distributions is due to the fact that the classical likelihood-based inference cannot be applied because the corresponding hypothesis is on the boundary of the parameter space [17].

3.2 Models for Multicategory Response

3.2.1 Generalized Estimating Equations (GEE)

GEE developed by Liang and Zeger (1986) are an extension of GLM, in which a specific type of correlation structure is incorporated into the variance function. The model variance is adjusted in such a manner as to minimize bias resulting from extra correlation in the data [10]. GEE requires only the correct specification of the univariate marginal distributions provided one is willing to adopt 'working' assumptions about the correlation structure. These include independence, exchangeable, autoregressive [AR(1)] and unstructured [17]. However, for multinomial response, independence is currently the only working correlation matrix in SAS and it indicate that the data are not correlated [20].

Let $Y_i = [(Y_{i1}, ..., Y_{im})]'$ be the vector of responses on the i^{th} cluster and X a matrix contains covariates associated with the response values. The model that express the response as function of covariates can be formulated as in a GLM:

$$E[Y_i] = \mu_i, \eta(\mu_i) = X_i\beta, \tag{6}$$

where: μ_i is the mean response vector for i^{th} cluster; β is a vector of unknown regression coefficients and $\eta(.)$ is the link function. The score equation that is used to estimate the

marginal regression parameters while accounting for the correlation structure is given by:

$$S(\beta) = \sum_{i=1}^{N} [\frac{\partial \mu_i}{\partial \beta}](A_i^{1/2} R_i A_i^{1/2})^{-1}(Y_i - \mu_i) = 0, \tag{7}$$

where the marginal covariance matrix V_i has been decomposed in the $A_i^{1/2} R_i A_i^{1/2}$ with A_i the matrix containing the marginal variances on the main diagonal and zeros elsewhere, and with R_i equal to the marginal correlation matrix [17]. Y_i Is multivariate vector of binary responses variables. For nominal responses Y_i, baseline-category logits with L categories, the models that describe the odds of each category relative to a baseline is expressed as follow [1]:

$$log[\frac{P(Y_i = k)}{P(Y_i = L)}] = \alpha_k + X\beta_k, k = 1, ..., L-1. \tag{8}$$

One can then compare marginal distributions at particular settings of X or evaluate effects of X on the response Y_i. In the analysis of this report, BCL models with Y_4 and Y_5 separately as response was fitted. Later, model (8) was extended to allow random effects at provincial as well as at EA level by fitting a GLMM as is formulated by model (4), in order to handle two-level hierarchical structure (couples clustered within EAs and EAs within provinces).

3.3 Models for Bivariate Binary Responses

3.3.1 Alternating logistic regressions (ALR) and Bivariate Dale Model (BDM)

Let $Y_i = (Y_{i1}, Y_{i2})$, be a vector of indicator variables representing the HIV status of woman and man (within a couple) from i^{th} cluster. Given bivariate binary responses, from the same subject (couple), the ALR model (introduced in Section 3.2) and BDM were found to be plausible to model the joint probability $\Pi_{11} = P(Y_{i1} = 1, Y_{i2} = 1)$ for both woman and man to be HIV positive. Using BDM, the marginal structure is flexibly modelled, i.e., the cumulative marginal probabilities can be fitted in the GLM framework. Marginal parameters are orthogonal onto the association parameters in the sense that the corresponding elements in the expected covariance matrix are identically zero. BDM does not require marginal scores for the responses and is essentially invariant under any monotonic transformation of the marginal response variables [15]. The model considered in this analysis consists of the following three models which are modelled simultaneously:

$$\begin{aligned} logit(\Pi_{1+}) &= X_i\beta_1 \\ logit(\Pi_{+1}) &= X_i\beta_2 \\ log(OR) &= X_i\beta_3, \end{aligned} \tag{9}$$

where: β_1, β_2 and β_3 are the vector of unknown regression coefficients to be estimated; X is vector of covariates associated with the marginal probability of being HIV positive for woman and man, respectively, and the OR denotes the odd ratios $\frac{\Pi_{11}\Pi_{00}}{\Pi_{10}\Pi_{01}}$; Π_{1+}, Π_{+1} are marginal probabilities while Π_{00} refers to couple where both man and woman are HIV negative; Π_{10} refers

to couple where the man is HIV positive and woman is HIV negative; Π_{01} refers to couple where man is HIV negative and woman HIV positive; Π_{11} refers to couple where both man and woman are HIV positive; Using (9), when $OR \neq 1$, $\Pi_{11} = 1 + (\Pi_{1+} + \Pi_{+1})(OR - 1) - [1 + (\Pi_{1+} + \Pi_{+1})(OR - 1)]^2 \Pi_{1+}\Pi_{+1}{}^{1/2}/[2(OR - 1)]$ and when $OR = 1, \Pi_{11} = \Pi_{1+} + \Pi_{+1}$ the multinomial (log) likelihood can be expressed straightforward.

3.3.2 Random Effects Model

When interest is in the marginal population-averaged models to describe the relationships of the covariates to the dependent variable for an entire population, marginal models as discussed in Section 3.3.1 are preferred. However, subject-specific inference may be of interest, hence random effect models are the optimal choice. These models differ from marginal models by the inclusion of parameters that are specific to the subject. In the analysis of bivariate binary responses, model (9) was extended to model (10) in order to allow two-level random effects, at province as well as at EA level.

$$\begin{aligned} logit[P(Y_{i1} = 1|b_i, b_{ij})] &= X_i\beta_1 + Z_{ij}b_i + Z_{ij}b_{ij} \\ logit[P(Y_{i2} = 1|b_i, b_{ij})] &= X_i\beta_2 + Z_{ij}b_i + Z_{ij}b_{ij} \\ log(OR) &= X_i\beta_3, \end{aligned} \tag{10}$$

where: β_1, β_2 and β_3 are the vectors of unknown regression coefficients to be estimated; X is vector of covariates, respectively. The random effects b_i and b_{ij} are assumed to be independent and normally distributed as $b_i \sim N(0; D)$ and $b_{ij} \sim N(0; D)$, respectively. The D matrix can assume different covariance structures. The variance component for b_i refers to the between-cluster and for b_{ij} to the within-cluster variability. A two-level hierarchical structure, couples nested within EAs and EAs within provinces was also considered.

3.3.3 Shared Random Effects Model

As introduced in Section 3.1.2, a GLMM is a conditional model that can capture individual heterogeneity by conditioning on subject-specific random effects in the model. For case of two binary responses $Y_i = (Y_{i1}, Y_{i2})$ (as defined in Section 3.3.1) from the same subject, there are two ways in which the correlations between the two responses can be incorporated: through induced shared random effects or model the dependency directly [20]. Let a_i and b_i be a subject-specific random effects from j^{th} subject in i^{th} cluster assumed to be normally distributed, $a_i \sim N(0; D)$ $b_i \sim N(0; D)$. Different covariance structures for D matrix can be assumed. These include independent with equal variance, independent with different variance, shared, unstructured and Toeplitz matrix. More details on different covariance structure can be found in [20]. Conditional on a_i and b_i the model showing the probability of woman and man from i^{th} cluster being HIV positive is given as follows:

$$g[P(Y_{i1} = 1|a_i)] = X_i\beta + a_i$$

$$g[P(Y_{i2} = 1|b_i)] = X_i\beta + b_i, \tag{11}$$

where: X is vector of covariates and β vector of unknown regression coefficients to be estimated; Choosing the function $g(.)$ to be the logit link, parameters estimates β can interpreted as the log odds ratios. In model (11), if the null hypothesis of variance for random effects a_i and b_i is not rejected, this implies that the two HIV statues of man and woman are independent, otherwise HIV status of man is associated with the HIV status of woman or vice versa.

3.4 Additional Hierarchical Structures

In the analyses of this report clustering resulting from polygamy as well as clustering within household was ignored or assumed to independent since few couples had more than one wife and few households had more one couple(See Section 2.2). These clustering could be modeled by including additional random effects at household level in a hierarchical way, as was done at provincial and EA level.

3.5 Accounting for Survey Design

Most of the sample designs for household surveys such as in INSIDA are complex and involve stratification, multistage sampling, and unequal sampling rates. Such survey designs are sometimes important since they better cover the entire region of interest (stratification) and there are efficient in interviewing subjects [21]. One of the gains in accounting for a complex sample design is the precision of survey estimates. Ignoring the design structure, for instance assuming simple random sampling (SRS), could result in underestimated standard errors, possibly leading to results that are seem to be statistically significant, when in fact, they are not. These happen when there are certain subpopulations that have been oversampled. The difference in point estimates and standard errors obtained using non-survey and survey procedure varies from dataset to dataset and even between variables within the same data set.

There are two approaches that can be used to take into account the survey design: designed-based and model-based approaches. Under design-based approach, a single level analysis can be maintained after adjustments that are made for sample design effects including unequal subject selection probabilities (sample weights) and non-independence of observations resulting from clustered designs [18]. In Model-based approaches (i.e., multilevel) directly incorporate the clustered sample design into the analytical models. The variation at each level can then be explained simultaneously by sets of covariates at each level of the data hierarchy [18]. In the analyses of this report, the designed-based approach was applied to account for the INSIDA survey design by using sample weights.

4 Application to INSIDA 2009 Data

4.1 Exploratory Data Analysis (EDA)

Out of 2466 couples, 83.01% were concordant negative, 5.56% female discordant, 5.76% male discordant and the rest 5.68% were concordant positive. Figure 3 shows that the percentage of HIV positive women and men in south region of the country is low compared to other regions. The percentage of HIV positive women and men is high for those with primary level of education as well as when the man is more educated than the woman.

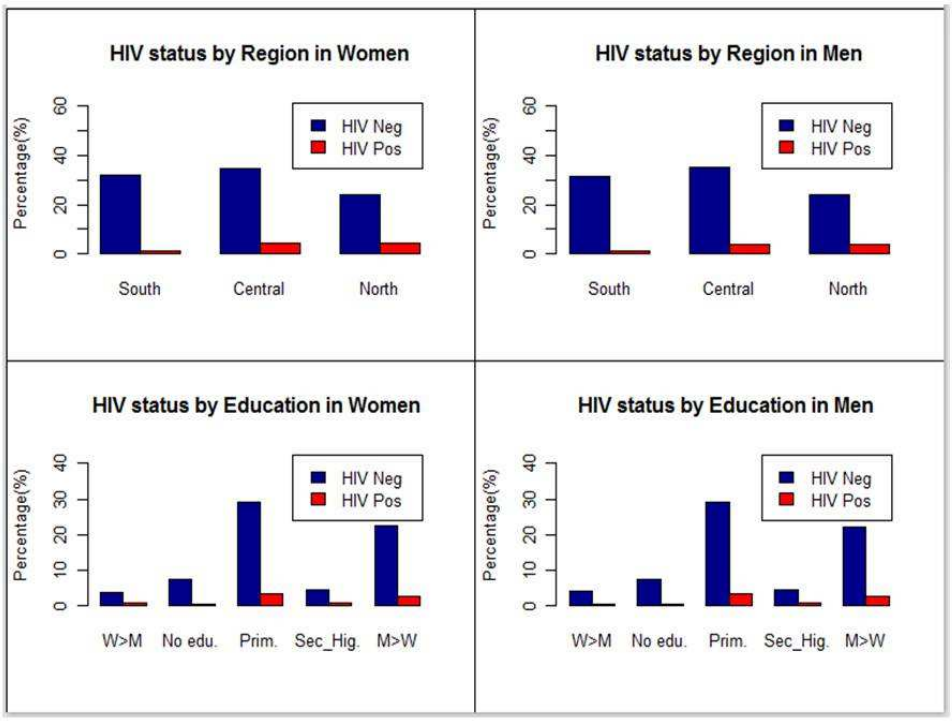

Figure 3: HIV Status by Region and Education Level

From figure 4 it seems that there is a high percentage of women and men who did not test for HIV. With regard to residence zone, the percentage of men who are HIV positive seems to be higher than for women in urban as well as in rural area.

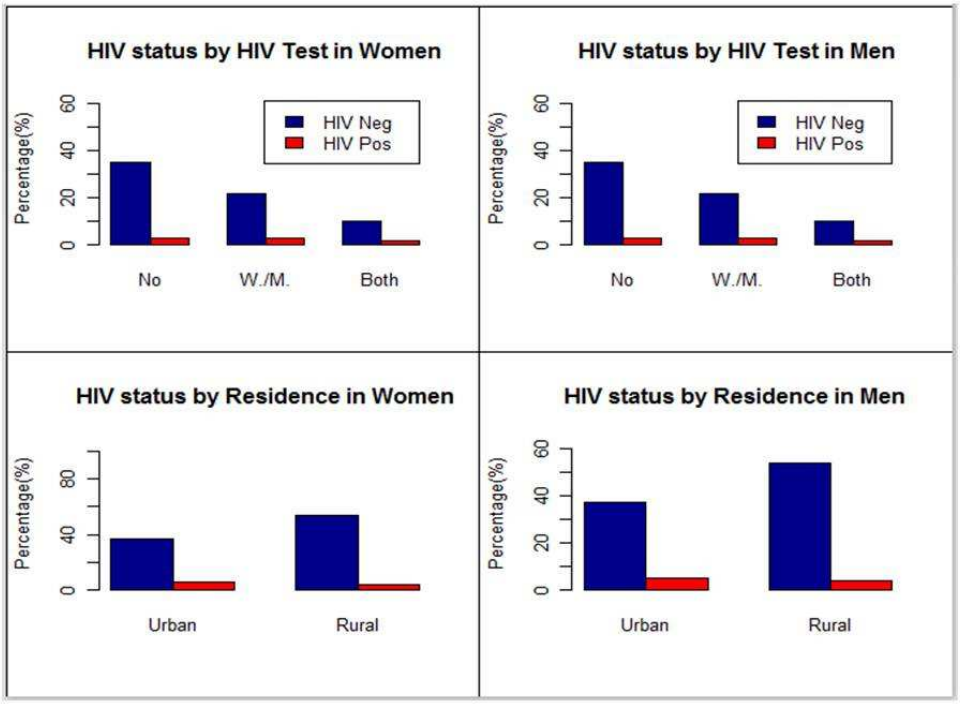

Figure 4: HIV Status by HIV test and Residence zone

Figure 5, shows that the percentage of HIV positive women and men is slightly higher when both (woman and man) are married once and when both are married more than once. With regard to marital duration, it can be observed that as the duration increases the percentage of HIV positive is tending to decrease, either in women or men.

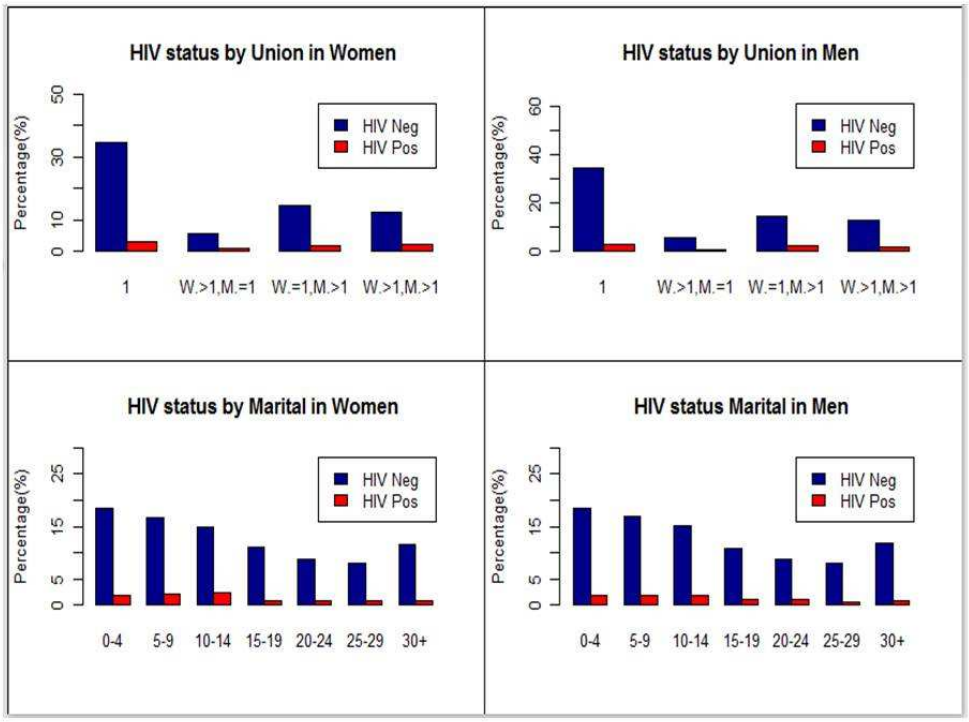

Figure 5: HIV Status by Union Number and Marital Duration

The percentage of female discordant, male discordant as well as concordant positive couples is slightly higher when the marital duration is less than 14 years. Also the percentage of female discordant, male discordant as well as concordant positive couples is slightly higher for those couples that didn't use condoms with non-spouse/husband in last 12 months (Figure 6).

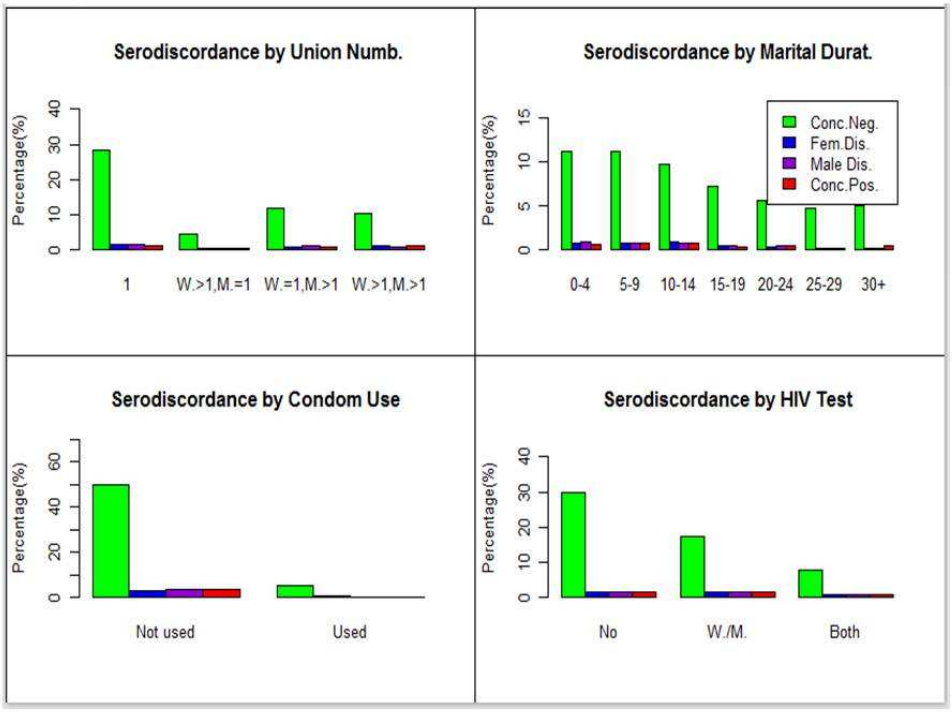

Figure 6: Serodiscordance by Union Number, Marital Duration, Condom and HIV Test

In central as well as in the northern region, the percentage of discordant (female or male) and concordant positive couples is slightly higher than in the southern region. Richer couples and those with primary level of education showed a high rate of HIV positive, either in female or male discordant and in concordant positive (Figure 7).

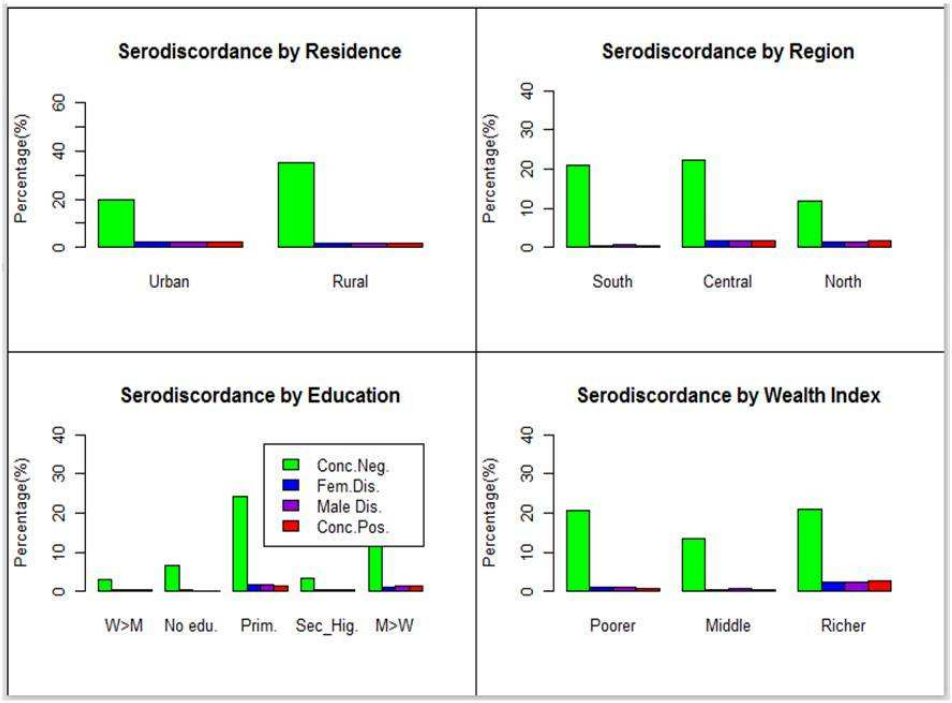

Figure 7: Serodiscordance by Region, Residence zone, Education level and Wealth Index

4.2 Bivariate Associations

Table 3 shows p-values of chi-square test of bivariate association of each HIV statuses (woman and man) by each risk factor. Both HIV statuses are highly associated with prevalence, wealth index, union number and region. This gives an indication that these covariates may be important risk factor to include in the models. Education level and STID had highest percentage of missingness, 36.9% and 6.81% respectively.

Table 3: P-value for Chi-Square association and (%) of Missingness in covariates

Factor	HIV Status of Women	Status of Men	(%) of Missingness
Prevalence	$< 2.2e-16$	$< 2.2e-16$	0%
STID	0.013	0.095	6.81%
HIV Test	0.001	0.004	0%
Condom Used	0.411	0.894	0%
Union Number	2.7e-17	8.1e-5	0%
Marital Duration	0.219	0.305	0.67%
Age Difference	0.013	0.146	0%
Education Level	0.001	0.001	36.9%
Wealth Index	1.5e-15	2.2e-15	0%
Residence zone	1.4e-10	2.2e-08	0%
Region	<2.2e-16	1.8e-15	0%

4.3 Models for Clustered Univariate Binary Responses

4.3.1 Alternating Logistic Regression (ALR)

As introduced in Section 3.3.1 ALR was found to be a plausible approach to model the probability of a couple being discordant while accounting for the two levels of clustering. A stepwise approach was used to select variables into the model. After selecting significant covariates, all pair-wise interactions of those covariates were added into the model. Non-significant interactions were removed one at the time starting with those highly insignificant. The final model expressing the probability of a couple being discordant from the j^{th} EA in the i^{th} province is given below (QIC= 399.431) and parameter estimates are shown in Table 4.

$$Logit[P(Y_{3ij}=1)] = \beta_0 + \beta_1 x_{1ij} + ... + \beta_3 x_{3ij} + \beta_8 x_{8ij} + ... + \beta_{12} x_{12ij} + \beta_{13} x_{22ij} + \beta_{14}$$

$$x_{1ij}x_{17ij} + \beta_{15} x_{1ij}x_{18ij} + \beta_{16} x_{2ij}x_{17ij} + \beta_{17} x_{2ij}x_{18ij} + \beta_{18} x_{1ij}x_{22ij} + \beta_{19} x_{2ij}x_{22ij}. \quad (12)$$

Interaction between prevalence and residence zone was found to be significant, meaning that the effect of the prevalence on the probability of a couple being discordant differs in urban and rural areas. For instance, a couple within province where the HIV prevalence is higher than 15% and in urban area was $2.305[e^{3.401-2.566}]$ times more likely to be discordant versus concordant positive compared to a couple within the same province but in rural area, controlling all other factors.

A couple where at least one member had STID or symptoms of STID in the past 12 months was $0.328[e^{-1.115}]$ times less likely to be discordant versus concordant positive compared to a couple where none had have STID or symptoms of STID, controlling all other factors. On the other hand, a couple with marital duration between $[0-4]$ years was $2.717[e^{0.997}]$ times more likely to be discordant versus concordant positive compared to a couple with marital duration between $[25-29]$ years, controlling all other factors. Alpha 2 was found to be significant, implying that a couple was $0.922[e^{-0.082}]$ times less likely to be discordant when another couple from different EA, within the same province was discordant.

Table 4: Estimates(SE) of ALR and GLMM(Binary Response)

Effect	ALR Estimates(SE)	GLMM Estimates(SE)
Intercept	-1.532(0.286)*	-1.105(1.605)
Prevalence		
$[5\% - 15\%]$	1.316(0.309)*	0.955(1.546)
$[> 15\%]$	1.830(0.417)*	1.672(1.523)
STID		
1 had STID	-1.115(0.314)*	-1.306(0.503)*
Marital Duration (Years)		
$[0 - 4]$	0.997(0.299)*	1.117(0.711)
$[5 - 9]$	-0.103(0.287)	-0.170(0.689)
$[10 - 14]$	0.219(0.457)	0.127(0.689)
$[15 - 19]$	-0.059(0.653)	0.137(0.733)
$[20 - 24]$	0.217(0.341)	0.246(0.760)
Wealth Index		
Poorer	2.549(1.345)	3.054(1.865)
Middle	3.293(0.260)*	2.274(1.792)
Residence zone		
Urban	3.401(0.518)*	3.122(1.799)
Prevalence x Residence zone		
$[5\% - 15\%]$ x Urban	-3.400(0.604)*	-3.225(1.868)
$[> 15\%]$ x Urban	-2.566(0.624)*	-2.428(1.892)
Prevalence x Wealth Index		
$[5\% - 15\%]$ x Poorer	-2.525(1.381)	-2.178(1.935)
$[5\% - 15\%]$ x Middle	-2.305(0.456)*	-2.302(1.881)
$[> 15\%]$ x Poorer	-1.727(1.392)	-1.915(2.099)
$[> 15\%]$ x Middle	-2.445(0.413)*	-1.239(1.948)
Association		
Alpha1	0.092(0.321)	-
Alpha2	-0.082(0.026)*	-
Variance of Component(EA)		
$\sigma^2_{b_{ij}}$	-	0.715(0.419)*a

* Significant at 5% level (Wald)
*a Significant at 5% level (mixtures of chi square)
SE Standard error

4.3.2 Random Effects Model

A GLMM with a simple variance-covariance matrix was fitted with the same covariates as in ALR. The AIC and BIC of this model is 407.99 and 468.12, respectively. Results of LR tests indicate that the random intercepts at province level could be deleted from the model $[\chi^2_{1:2}(0.001), p - value = 0.987]$, while the random intercepts at EA level cannot be simplified $[\chi^2_{0:1}(746.86), p - value = 0.0001]$. Hence, conditional on EA-specific random intercept b_{ij}, the model showing the probability of a couple being discordant from the j^{th} EA in the i^{th} province is given below and parameter estimates are shown in Table 4 above:

$$Logit[P(Y_{3ij} = 1|b_{ij})] = \beta_0 + \beta_1 x_{1ij} + ... + \beta_3 x_{3ij} + \beta_8 x_{8ij} + ... + \beta_{12} x_{12ij} + \beta_{13} x_{22ij} + \beta_{14}$$

$$x_{1ij} x_{17ij} + \beta_{15} x_{1ij} x_{18ij} + \beta_{16} x_{2ij} x_{17ij} + \beta_{17} x_{2ij} x_{18ij} + \beta_{18} x_{1ij} x_{22ij} + \beta_{19} x_{2ij} x_{22ij} + b_{ij}. \quad (13)$$

The variance of the EA random intercepts was found to be significant, indicating that there is heterogeneity among EA-specific intercepts, resulting in variation from one EA to another in probability a couple being discordant. The only fixed effect that was found to be significant is STID, meaning that for a given EA, a couple where at least one member had STID or symptom of STID was $0.266[e^{-1.326}]$ times less likely to be discordant versus concordant positive compared to a couple where none had have STID or symptom of STID.

4.4 Models for Multicategory Response

4.4.1 Baseline Category Logit (BCL) Model

Since the response Y_4 and Y_5 of serodiscordance was assumed to nominal, a BCL model was estimated and again a stepwise procedure was used to select covariates in the model. The logit models are given below [model (14) including all couples while model (15) was restricted to positive couples only] and parameter estimates are presented in Table 5.

$$
\begin{aligned}
log(\frac{\Pi_{00}}{\Pi_{10}}) &= \beta_{10} + \beta_{11}x_{1i} + ... + \beta_{17}x_{7i} + \beta_{117}x_{17i} + \beta_{118}x_{18i} \\
log(\frac{\Pi_{01}}{\Pi_{10}}) &= \beta_{20} + \beta_{21}x_{1i} + ... + \beta_{27}x_{7i} + \beta_{217}x_{17i} + \beta_{218}x_{18i}, \qquad (14) \\
log(\frac{\Pi_{11}}{\Pi_{10}}) &= \beta_{30} + \beta_{31}x_{1i} + ... + \beta_{37}x_{7i} + \beta_{317}x_{17i} + \beta_{318}x_{18i}
\end{aligned}
$$

and:

$$
\begin{aligned}
log(\frac{\Pi_{01}}{\Pi_{11}}) &= \beta_{10} + \beta_{11}x_{1ij} + ... + \beta_{17}x_{7ij} + \beta_{117}x_{17ij} + \beta_{118}x_{18ij} \\
log(\frac{\Pi_{10}}{\Pi_{11}}) &= \beta_{20} + \beta_{21}x_{1ij} + ... + \beta_{27}x_{7ij} + \beta_{217}x_{17ij} + \beta_{218}x_{18ij}, \qquad (15)
\end{aligned}
$$

where: Π_{00} refers to a concordant negative couple; Π_{01} to a female discordant couple; Π_{10} to a male discordant; Π_{11} refers to a concordant positive couple. The AIC and BIC for model (14) is 2733.64 and 2888.61 while for model (15) 797.91 and 869.48, respectively. A couple where both members (woman and man) had married more than once was $0.251[e^{-1.382}]$ times less likely to be female versus male discordant compared to a couple where both members had married only once (Model 14), controlling for all other factors. This means that prior marriage for both members within couples appears to be a risk factor associated with lower probability of HIV infection in the women than in men.

On the other hand, a couple where the woman had married only once and man had married more than once was $2.085[e^{0.735}]$ times more likely to be female discordant versus concordant positive or $3.068[e^{1.121}]$ times more likely to be male discordant versus concordant positive compared to a couple where both had married only once (Model 15), controlling for other factors. A couple where both members had married more than once was $3.062[e^{1.119}]$ times more likely to be male discordant versus concordant positive compared to a couple where both had married only once (Model 15), holding other factors constant. These findings show that prior marriage for men or both members within a couple seems to be a risk factor associated with higher probability of HIV infection in men than in women.

A couple where at least one member had STID or symptoms of STID was $3.271[e^{1.185}]$ times more likely to be male discordant versus concordant positive, compared to a couple where none had STID or symptoms of STID on the past 12 months (Model 15), controlling for all other

factors. This means that couples where at least one member had STID or symptoms of STID was associated with a higher risk of HIV infection in men. A poorer couple was $0.343[e^{-1.069}]$ times less likely to be female discordant or $0.415[e^{-0.880}]$ times less likely to be male discordant relative of being concordant positive, compared to a richer couple (Model 15). This means that poorer couples were more likely to be positive concordant.

Table 5: Estimates(SE) of BCL Models

	Model 14			Model 15	
Effect	$log(\frac{\Pi_{00}}{\Pi_{10}})$	$log(\frac{\Pi_{01}}{\Pi_{10}})$	$log(\frac{\Pi_{11}}{\Pi_{10}})$	$log(\frac{\Pi_{01}}{\Pi_{11}})$	$log(\frac{\Pi_{10}}{\Pi_{11}})$
Intercept	2.692(0.564)*	0.928(0.746)	1.657(0.679)*	-0.445(0.643)	-1.399(0.673)*
Prevalence					
$[5\% - 15\%]$	0.368(0.227)	-0.338(0.315)	-0.365(.307)	0.031(0.312)	0.343(0.305)
$[> 15\%]$	1.836(0.350)*	0.230(0.458)	-0.723(0.519)	0.657(0.498)	0.447(0.528)
STID					
1 had STID	-0.134(0.4673)	-0.148(0.620)	-1.390(0.522)*	0.506(0.506)	1.185(0.526)*
Union Number					
Woman>1;Man=1	0.135(0.259)	-0.407(0.330)	-1.250(0.346)*	0.735(0.328)*	1.121(0.351)*
Woman=1;Man>1	-0.373(0.349)	-0.557(0.464)	-0.734(0.456)	0.249(0.460)	0.797(0.467)
Both>1	-0.260(0.281)	-1.382(0.417)*	-1.129(0.376)*	-0.228(0.421)	1.119(0.389)*
Wealth Index					
Poorer	-0.631(0.227)*	-0.251(0.312)	0.871(0.349)*	-1.069(0.355)*	-0.880(0.357)*
Middle	-0.094(0.264)	-0.308(0.374)	0.657(0.403)	-0.813(0.430)	-0.596(0.424)

* Significant at 5% level (Wald)
SE Standard error

4.4.2 Random Effects Model

Similar covariates as in the BCL were used in a GLMM with a simple variance-covariance matrix. Results of LR tests indicate that the covariance structure cannot be simplified by deleting the EA random intercepts from the model $[\chi^2_{0:1}(67.1)$, p-value$<0.0001]$ while for positive couples it can be simplified by deleting the random intercepts from the model $[\chi^2_{0:1}(0.55)$, p-value$=0.458]$. An attempt was made to test the need of random effects at province level however due to convergence issues this was not possible. Hence model (14) was extended to allow EA random intercepts only (model 16) and parameter estimates are given in Table 6.

$$
\begin{aligned}
log(\frac{\Pi_{00}}{\Pi_{10}}|b_{1j}) &= \beta_{10} + \beta_{11}x_{1i} + ... + \beta_{17}x_{7i} + \beta_{117}x_{17i} + \beta_{118}x_{18i} + b_{1j} \\
log(\frac{\Pi_{01}}{\Pi_{10}}|b_{2j}) &= \beta_{20} + \beta_{21}x_{1i} + ... + \beta_{27}x_{7i} + \beta_{217}x_{17i} + \beta_{218}x_{18i} + b_{2j} \\
log(\frac{\Pi_{11}}{\Pi_{10}}|b_{3j}) &= \beta_{30} + \beta_{31}x_{1i} + ... + \beta_{37}x_{7i} + \beta_{317}x_{17i} + \beta_{318}x_{18i} + b_{3j}.
\end{aligned} \tag{16}
$$

The AIC and BIC of this model is 2672.54 and 2780, respectively. The variance of the EA random intercepts was found to be significant for female discordant and concordant positive.

This shows that there was heterogeneity between EAs on the probability of couples being female discordant as well as of being positive concordant. The only fixed effect significant for female discordant is the union number where both members had married more than once. Conditional to specific EA, a couple where both members had married more than once was $0.242[e^{-1.420}]$ times less likely to be female versus male discordant compared to a couple where both had married only once. This indicates that prior marriage for both members within a couple is associated with a lower risk of HIV infection for women than for men.

Table 6: Estimates(SE) of GLMM(Multicategory Response)

	Model 14		
Effect	$log(\frac{\Pi_{00}}{\Pi_{10}})$	$log(\frac{\Pi_{01}}{\Pi_{10}})$	$log(\frac{\Pi_{11}}{\Pi_{10}})$
Intercept	2.749(0.586)*	0.832(0.760)	1.461(0.723)*
Prevalence			
$[5\% - 15\%]$	0.555(0.264)*	-0.394(0.338)	-0.463(0.371)
$[> 15\%]$	1.893(0.386)*	0.191(0.481)	-0.946(0.609)*
STID			
1 had STID	-0.224(0.478)	-0.124(0.625)	-1.475(0.543)*
Union Number			
Woman>1;Man=1	0.213(0.268)	-0.432(0.336)	-1.338(0.368)*
Woman=1;Man>1	-0.389(0.361)	-0.530(0.470)	-0.773(0.480)
Both>1	-0.182(0.291)	-1.420(0.422)*	-1.209(0.397)*
Wealth Index			
Poorer	-0.578(0.251)*	-0.255(0.332)	0.971(0.387)*
Middle	-0.110(0.278)	-0.287(0.386)	0.630(0.432)
Variance Component(EA)			
$\sigma^2_{b_j}$	0.236(0.223)	0.452(0.139)*a	0.930(0.374)*a

* Significant at 5% level (Wald)
*a Significant based on mixture of Chi square
SE Standard error

4.5 Models for Bivariate Binary Responses

4.5.1 Alternating Logistic Regression (ALR) and Bivariate Dale Model (BDM)

Several BDM were fitted and stepwise procedure was used to select covariates into the model. Marital duration is highly associated with union number (p-value<2.2e-16), two separate models were fitted and AIC as well as BIC were compared. The one with marital duration showed lowest AIC and BIC (AIC=1557.879 and BIC=1524.879) compared to the one with union number (AIC=1558.919 and BIC=1531.919). Due to the fact the both models were not fit well the data, both variables were retained in the model (AIC=1544.754 and BIC=1515.754).

Models with constant OR and non-constant OR (OR depending on covariates) with three different link functions were fitted and comparison of these models is shown in Table 7. The model with non-constant OR and cloglog link function seems to be the "best", since it has the lowest AIC (1532.197) and BIC (1490.197) respectively. The residual deviance [$\chi^2(3780)$] of 1448.197 indicates that this model does provide a very good fit to the data. The final model considered

Table 7: BDM with different link functions

	Constant OR		Non-constant OR	
Link Function	**AIC**	**BIC**	**AIC**	**BIC**
Logit	1544.754	1515.754	1532.901	1490.901
Probit	1545.022	1516.022	1532.97	1490.97
Cloglog	1544.064	1515.064	1532.197	1490.197

for BDM is given below. This model was also fitted for the ALR model (QIC=2908.99) and parameter estimates of both models (ALR model and BDM) are presented in Table 8.

$$
\begin{aligned}
logit(\Pi_{1+}) &= \beta_{10} + \beta_{11}x_{i1} + ... + \beta_{13}x_{i3} + \beta_{15}x_{i5} + ... + \beta_{112}x_{i12} + \beta_{117}x_{i17} + \beta_{118}x_{i18} \\
logit(\Pi_{+1}) &= \beta_{20} + \beta_{21}x_{i1} + ... + \beta_{23}x_{i3} + \beta_{25}x_{i5} + ... + \beta_{212}x_{i12} + \beta_{217}x_{i17} + \beta_{218}x_{i18} \qquad (17) \\
logit(OR) &= \beta_{30} + \beta_{31}x_{i1} + ... + \beta_{33}x_{i3} + \beta_{35}x_{i5} + ... + \beta_{312}x_{i12} + \beta_{317}x_{i17} + \beta_{318}x_{i18}.
\end{aligned}
$$

The estimates for association parameters [log(OR)] between the two HIV statuses (woman and man) were found to be not significant, indicating that the associations between the two HIV statuses was the same in different levels of each variable. Prevalence, marital duration and wealth index were found to statistical significant in both women and men. The estimates of prevalence, STID and marital duration are all positive meaning that within a couple, both woman and man were more likely to HIV positive compared a couple in reference category of each variable. For instance, a couple within a province where the HIV prevalence was higher than 15%, the woman was 4.792[$e^{1.567}$] times more likely to be HIV positive and the man was 5.073[$e^{1.624}$] times more likely to be HIV positive compared to a couple within a province where the HIV prevalence was lower than 5%, controlling for all other factors. A couple with marital

duration between $[0-4]$ years, the woman was $7.121[e^{1.963}]$ times more likely to be HIV positive and the man was $3.717[e^{1.313}]$ times less likely to be HIV positive compared to a couple with marital duration between [25-29] years, controlling for all other factors.

Within poorer as well as middle couples, women and men were less likely to be HIV positive compared to a richer couples, since the estimates are negative. For example, within a poorer couple, the woman was $0.377[e^{-0.975}]$ times less likely to be HIV positive and the man was $0.390[e^{-0.942}]$ times less likely to be HIV positive compared to a richer couple, controlling for all other factors. On the other hand, this shows that richer couples were at higher risk of being HIV positive.

With regards to the union number, women are consistently at lower risk of being HIV positive than men. This indicates that within couples, women were less likely to be HIV positive than men, independent of whether the woman had or not prior marriage. For instance, a couple where the woman had married more than once and the man only once, the woman was $0.265[e^{-1.329}]$ times less likely to be HIV positive while the man was $0.538[e^{-0.620}]$ times less likely to be HIV positive compared to a couple where both members had married only once and controlling for all other factors. STID is only significant in men indicating that men were $2.075[e^{0.730}]$ more likely to be HIV positive within couples where at least one member had STID or symptoms of STID in the past 12 months, controlling for all other factors.

Estimates from the ALR model are close to those from the BDM, since both are marginal model through a ALR model is quasi likelihood and was fitted with constant OR. Some estimates are significant in women, not in men or vice versa. Similar interpretation as in the BDM can be done. Alpha 1 was found to be significant, meaning that a couple was $1.492[e^{0.400}]$ times more likely to be HIV positive when another couple within the same EA was also HIV positive. This results shows that within EAs there is higher rate of HIV infection between members of different couples, once one member of couple is infected.

Table 8: Estimates(SE) of ALR and BDM

	ALR		BDM		
Effect	**Women Estimates(SE)**	**Men Estimates(SE)**	**Women Estimates(SE)**	**Men Estimates(SE)**	**log(OR) Estimates(SE)**
Intercept	-4.391(0.270)*	-4.550(0.271)*	-3.109(0.625)*	-3.339(0.615)*	6.764(1.932)*
Prevalence					
$[5\% - 15\%]$	1.401(0.247)*	1.584(0.191)*	1.093(0.255)*	1.410(0.271)*	-1.582(0.845)
$[> 15\%]$	2.073(0.140)*	2.045(0.182)*	1.567(0.284)*	1.624(0.296)*	-1.490(0.944)
STID					
1 had STID	0.507(0.282)	0.525(0.455)	0.404(0.293)	0.730(0.273)*	-1.939(1.069)
Union Number					
Woman>1;Man=1	0.777(0.253)*	0.751(0.148)*	-1.329(0.218)*	-0.620(0.222)*	-0.688(0.709)
Woman=1;Man>1	0.120(0.260)	0.651(0.167)*	-0.656(0.300)*	0.010(0.288)	-1.247(0.916)
Both>1	1.154(0.182)*	0.803(0.170)*	-1.175(0.270)*	-0.235(0.249)	0.157(0.838)
Marital Duration[1]					
$[0 - 4]$	1.071(0.455)*	0.870(0.230)*	1.963(0.559)*	1.313(0.548)*	-3.004(1.561)
$[5 - 9]$	0.710(0.375)	0.964(0.276)*	1.627(0.554)*	1.725(0.532)*	0.954(1.561)
$[10 - 14]$	0.836(0.379)*	0.805(0.225)*	1.813(0.556)*	1.443(0.543)*	-0.967(1.523)
$[15 - 19]$	0.200(0.410)	0.581(0.292)*	1.168(0.578)*	1.254(0.557)*	0.368(1.601)
$[20 - 24]$	0.486(0.394)	0.761(0.322)*	1.510(0.577)*	1.456(0.559)*	-1.796(1.574)
Wealth Index					
Poorer	-0.726(0.158)*	-0.930(0.254)*	-0.975(0.212)*	-0.942(0.207)*	0.129(0.666)
Middle	-0.402(0.192)*	-0.429(0.235)	-1.007(0.264)*	-0.900(0.254)*	-0.142(0.836)
Association					
Alpha1	0.400(0.144)*		-		
Alpha2	0.030(0.019)		-		

* Significant at 5% level
SE Standard error
OR Odds radio
1 In Years

4.5.2 Random Effects Model

A BDM with simple variance-covariance matrix for independent random effects and different variances was fitted. Results of LR tests indicate that the covariance structure cannot be simplified by deleting the province random intercepts from the model $[\chi^2_{0:1}(12.95), p - value < 0.0003]$. EA random intercepts were not possible to include, due to non-convergence. Hence model (17) was extended to allow province random intercepts only (model 18). Conditional on random intercept a_i and b_i of the j^{th} couple from the i^{th} province, the model showing the probability of couple being HIV positive is given below and parameter estimates are shown in Table 9:

$$logit[P(Y_{ij1} = 1|a_i)] = \beta_{10} + \beta_{11}x_{i1} + ... + \beta_{13}x_{i3} + \beta_{15}x_{i5} + ... + \beta_{112}x_{i12} + \beta_{117}x_{i17} +$$

$$\beta_{118}x_{i18} + a_i$$

$$logit[P(Y_{ij2} = 1|b_i)] = \beta_{20} + \beta_{21}x_{i1} + ... + \beta_{23}x_{i3} + \beta_{25}x_{i5} + ... + \beta_{212}x_{i12} + \beta_{217}x_{i17} +$$

$$\beta_{218}X_{i18} + b_i \tag{18}$$

$$log(OR) = \beta_{30}.$$

The AIC and BIC of this model is 878.19 and 889.33. Overall, parameter estimates are higher than those observed in the marginal models and most of them were found to be significant. The association between the HIV status of woman and man was estimated to be $1.176[e^{0.162}]$ and significant. This implies that when one member within a couple was HIV positive another was about 1.2 times more likely to be also HIV positive. The variance of random intercepts for women is higher than for men, both was found to be statistical significant. This means that there was heterogeneity or difference between provinces in the probability of being HIV positive. Note that the interpretation of these estimates can be done as in marginal models but conditional on province specific.

Table 9: Estimates(SE) of Random Effects Model (Bivariate Responses)

Effect	**Women Estimates(SE)**	**Men Estimates(SE)**
Intercept	-4.399(0.754)*	-4.473(0.507)*
Prevalence		
$[5\% - 15\%]$	1.278(0.782)	1.441(0.401)*
$[> 15\%]$	2.113(0.810)*	1.973(0.420)*
STID		
1 had STID	0.527(0.250)*	0.539(0.250)*
Union Number		
Woman>1;Man=1	0.765(0.255)*	0.719(0.254)*
Woman=1;Man>1	0.109(0.203)	0.649(0.184)*
Both>1	1.195(0.193)*	0.805(0.201)*
Marital Duration (Years)		
$[0 - 4]$	1.079(0.350)*	0.905(0.370)*
$[5 - 9]$	0.726(0.346)*	1.023(0.359)*
$[10 - 14]$	0.863(0.347)*	0.845(0.368)*
$[15 - 19]$	0.154(0.376)	0.598(0.383)
$[20 - 24]$	0.462(0.383)	0.799(0.391)*
Wealth Index		
Poorer	-0.781(0.195)*	-1.031(0.193)*
Middle	-0.391(0.216)	-0.517(0.206)*
Variance Component(Province)		
σ^2	0.082(0.022)*a	0.059(0.018)*a
Association parameter		
Association	0.162(0.102)*	

* Significant at 5% level
*a Significant based on mixture of Chi square
SE Standard error

4.5.3 Shared Random Effects Model

Let b_i be a couple-specific random intercept of j^{th} couple from i^{th} province assumed to be normally distributed $b_i \sim N(0; \sigma^2_{b_i})$. Three GLMM were estimated, one with independent random effect and equal variances (AIC=2731.03), with independent random effect and different variances (AIC=2952.27) and shared random effects (AIC=2767.89). A GLMM with other covariance structures, such unstructured and toeplitz were not considered since did not converge. However the GLMM with correlated shared random effects did not show the lowest AIC, this model was found to be plausible, since could take into account the possible association between the two HIV statuses within a couple, hence was considered as final model. Conditional on b_{ij}, the model showing the probability of couple being HIV positive is given below and parameter estimates are presented in Table 10:

$$\begin{aligned} g[P(Y_{ij1} = 1|b_i)] &= \beta_{10} + \beta_{11}x_{1ij} + ... + \beta_{13}x_{3ij} + \beta_{15}x_{15ij} + ... + \beta_{112}x_{12ij} + \beta_{117}x_{17ij} + \\ &\beta_{118}x_{18ij} + b_i \\ g[P(Y_{ij2} = 1|b_i)] &= \beta_{20} + \beta_{21}x_{1ij} + ... + \beta_{23}x_{3ij} + \beta_{25}x_{15ij} + ... + \beta_{212}x_{12ij} + \beta_{217}x_{17ij} + \\ &\beta_{218}x_{18ij} + \gamma b_i. \end{aligned} \tag{19}$$

In model (19) γ is scalar parameter and is used to relax the assumption of common variance between the random intercepts, since $\sigma^2_{Y_{ij1}} = \gamma\sigma^2_{Y_{ij2}}$. Note that model (19) also implies that if one member within a couple is at high risk of being HIV positive another member is also at high risk, since γ was found to be positive (1.110). The null hypothesis of LR test for variance of shared random intercepts can be rejected $[\chi^2_{0:1}(128.01), p - value < 0.0001]$, implying that there is significant heterogeneity among couple-specific intercepts, resulting in variation from couple-to-couple in a probability of being HIV positive.

Most of the parameter estimate were found to be statistically significant in both women and men. For a given couple, within a province where the HIV prevalence was between $[5\% - 15\%]$, the man was $4.229[e^{1.4426}]$ times more likely to be HIV positive and the woman was $3.466[e^{1.243}]$ times more likely to be HIV positive compared to a couple within a province where the HIV prevalence was less than 5%. Given a poorer couple, the man was $0.359[e^{-1.025}]$ times less likely to be HIV positive and the woman was $0.450[e^{-0.799}]$ times less likely to be HIV positive compared to a richer couple, controlling for all other factors.

Table 10: Estimates(SE) of Shared Random Effects Model(Bivariate Responses)

Effect	Women Estimates(SE)	Men Estimates(SE)
Intercept	-4.340(0.578)*	-4.508(0.594)*
Prevalence		
$[5\% - 15\%]$	1.243(0.537)*	1.442(0.543)*
$[> 15\%]$	2.183(0.287)*	2.021(0.566)*
STID		
1 had STID	0.516(0.248)*	0.540(0.250)*
Union Number		
Woman>1;Man=1	0.755(0.253)*	0.731(0.254)*
Woman=1;Man>1	0.114(0.202)	0.650(0.183)*
Both>1	1.179(0.189)*	0.822(0.200)*
Marital Duration (Years)		
$[0 - 4]$	1.097(0.349)*	0.887(0.371)*
$[5 - 9]$	0.744(0.345)*	1.007(0.360)*
$[10 - 14]$	0.874(0.346)*	0.832(0.368)*
$[15 - 19]$	0.179(0.375)	0.571(0.384)
$[20 - 24]$	0.482(0.382)	0.777(0.392)*
Wealth Index		
Poorer	-0.810(0.187)*	-1.017(0.191)*
Middle	-0.443(0.209)*	-0.470(0.205)*
Variance Component(Province)		
$\sigma^2_{b_i}$	0.105(0.062)*	
Scale parameter		
γ	1.110(0.022)	

* Significant at 5% level
SE Standard error

5 Discussion and Conclusion

This paper aimed at investigating the risk factors associated with HIV serodiscordance among couples in Mozambique. Cross-sectional data based on national-representative sample of the INSIDA survey in Mozambique was used. Several statistical models were applied motived by the nature of the response and by the design of the study. ALR model was estimated, modelling the probability of a couple being discordant. Moreover, a random effects model (GLMM) was considered by allow random effects at EA level. With regard to a multicategory response, two marginal models (BCL model) were estimated, one including all couples while another restricted to positive couples only. Furthermore, the model including all couple was extended to allow random effects at EA level. For bivariate binary outcomes, two marginal models (ALR model and BDM) were fitted, with a HIV status of each member within a couple as outcome. Moreover, two GLMM's, one allowing random effects at provincial level and another at couple level were also estimated.

Results from ALR show that the effect of HIV prevalence on the probability of a couple being discordant differs by residence zone as well as by different level of wealth index. Couples within provinces with high level of HIV prevalence and in urban area were more likely to be discordant. On the other hand, these results show that couples in rural (within province with high of HIV prevalence) were more likely to be concordant positive. These could be attributed to different factors such as lack of knowledge of their HIV statuses, lack of knowledge about prevention once one partner is infected, etc. Poorer as well as middle couples living within provinces with high level of HIV prevalence were more likely to be discordant compared to richer couples. STID or symptoms of STID is associated with high probability of couples being discordant. Furthermore, the chance of a couple being discordant was low when another couple from different EA within the same province was also discordant. These findings are in line with those from INSIDA 2009 where they found the wealthiest (richer) couples were less likely to be discordant than poorest couples; however this association lost significance after controlling for the HIV prevalence and other factors [8] (however they used ordinal logistic regression accounting for survey design). Random effects model revels that there is significant variation between EAs in the probability of a couple being discordant.

BCL estimates including all couples showed that prior marriage for both members within couples is associated with low chance of a couple being female versus male discordant. Analysis of BCL restricted to positive couples only indicate that prior marriage for men or both members within a couple is associated with high probability of a couple being male discordant. STID or symptoms of STID is associated with higher probability of men being HIV positive than women. Poorer couples were more likely to be female discordant or male discordant relative of being concordant positive, compared to richer couples. The random effects model shows that

there is significant variation between EAs in the probability of couples being female discordant as well as of being concordant positive. These findings are in agreement with those from ALR (univariate response).

The use of ALR (for bivariate binary responses) and BDM was motived by the presence of two paired HIV statuses (man and woman) from the same couple, which is expected to be associated, in sense that when one member was HIV positive another was also more likely to be HIV positive. BDM reveals that within a couple, this association was not depending on any risk factor. For couples within provinces with higher levels of HIV prevalence, men were at higher risk of being HIV positive than women. STID or symptoms of STID is associated with high probability of men being HIV positive than women.

With regards to union number, women are consistently at lower risk of being HIV positive than men. Couples where the woman had married more than once and the man only once, the woman was less likely to be HIV positive than the man. A similar trend is also observed within poorer as well as middle couples where women are less likely to HIV positive than men. Results from random effects models (GLMM) are in agreement with those from marginal models (ALR and BDM) however are slightly higher. Futhermore, this model (GLMM) reveals that there are differences between provinces in the probability of being HIV positive.

Note that the ALR and BDM (bivariate binary responses) model fit the marginal joint distribution of the two HIV statuses (woman and man) and are computationally simpler. Moreover, odds ratios are preferred to correlation coefficients (in case of BPM) when describing the association between the two HIV statuses (woman and man). In addition to that, the associations is modelled in a flexible way including covariates. However they are based on complete case only, which decreased the sample size.

The findings are consistent, indicating that within a couple, the man is at higher risk of HIV infection than the woman. This could be attributed to the fact that the man could be infected outside of marital relationship. In fact in Zambia, retrospectively study of 65 couples to estimates the likely origin of HIV infection, found that at least one quarter of cases of HIV infection in recently married men were acquired from extramarital partnerships, and for both men and women, less than one half of cases of HIV infection were acquired from their spouse/husband [9]. In addition, they report that many infections in married men, even in those with HIV-infected wives, could be acquired from outside the marriage [9]. Furthermore, a study conducted in South Africa to investigate who was infecting whom among migrant and non-migrant within concordance as well as discordance couples, found that non-migrant men were 10 times more likely to be infected from outside their regular relationships than inside [14].

The GLMM with shared random effects at couple level shows that there is significant heterogeneity among couples or variation from couple-to-couple in probability of being HIV positive. Furthermore, this model also indicate that if man or woman within a couple was at high risk of being HIV positive another member was also at high risk. This shows that the chance of one member being HIV positive within a couple is not uniform.

5.1 Conclusion

In this study, several statistical methods were applied to analyze HIV serodiscordance among couples in Mozambique. These methods showed that HIV prevalence, STID, union number and wealth index were risk factor associated HIV serodiscordance. The effect of HIV prevalence on probability of couples being discordant differs by residence zone as well as by different levels of wealth index. Couples within provinces with high level of HIV prevalence and in rural area are more likely to be concordant positive. Richer couples within provinces with higher level of HIV prevalence are more likely to be concordant positive. Prior marriage for men or both members within couples is associated with high probability of HIV infection in men than in women, consequently couples are more likely to be male discordant. STID or symptoms of STID is associated with higher probability of HIV infection in men than in women. In Mozambique there is heterogeneity or differences between provinces as well as between EAs in probability of couples being discordant and concordant positive. Within EAs there is high rate of HIV infection between members of different couples. In addition, positive and strong association between the HIV status of women and men was observed, in sense that when one member within a couple is HIV positive another member is also at high risk of HIV infection. Polices to reduce the HIV transmission within couples, EAs as well as within provinces once one member is HIV positive are required in Mozambique.

5.2 Limitations

One of the limitations is that with cross-sectional data it is impossible to know who was first infected within concordant positive couples. Longitudinal studies are the optimal choice to investigate risk factors associated with HIV seroconvertion within discordant couples to concordant positive couples. In addition, with cross-sectional data it is impossible to know whether the HIV positive members within couples were infected before or after marriage.

References

[1] Agresti, A. (2002).*Categorical Data Analysis*. Second Edition. New York: John Wiley and Sons Inc.

[2] Allison, P.D.(1999).*Logistic Regression Using the SAS System: Theory and Application.* Cary, North Carolina: SAS Institute Inc

[3] Carey, V., Zeger, S. L., and Diggle, P. Modelling multivariate binary data with alternating logistic regressions.*Biometrika.***Vol.80**, 517-526.1993

[4] Chromy, J. R. and S. Abeyasekera. Statistical analysis of survey data. Household Sample Surveys in Developing and Transition Countries. Unpublished.

[5] Del Fava, E., Shkedy, Z., Hens, N., Aerts, M., Suligoi, B., Camoni, L., Vallejo, F., Wiessing, L. and M. Kretzschmar. Joint Modelling of HCV and HIV Co-Infection among Injecting Drug Users in Italy and Spain Using Individual Cross-Sectional Data.*Statistical Communications in Infectious Diseases.***Vol.3:** Iss. 1, Article 3. 2011. Available at: http://www.bepress.com/scid/vol3/iss1/art3

[6] Dunkle, K.L., Stephenson, R., Karita, E., Chomba, E., Kayitenkore, K., Vwalika, C., Greenberg, L. and S.Allen. New heterosexually transmitted HIV infections in married or cohabiting couples in urban Zambia and Rwanda: an analysis of survey and clinical data. *The Lancet Infectious Diseases:* **Vol.371**: 2183-91. 2008

[7] Ewayo, O., de Walque, D., Ford, N., Gakii, G., Lester, R. T. and E.J. Mills. HIV status in discordant couples in sub-Saharan Africa: a systematic review and meta-analysis. *The Lancet Infectious Diseases.* **Vol.10**: 770-777. 2010

[8] Fishel, J. D., Bradley, S. EK., Young, P. W., Mbofana, F. and C. Botão. HIV among Couples in Mozambique: HIV Status, Knowledge of Status, and Factors Associated with HIV Serodiscordance. Further Analysis of the 2009 Inquérito Nacional de Prevalência, Riscos Comportamentais e Informação sobre o HIV e SIDA em Moçambique (INSIDA). *ICF International.* Calverton, Maryland, USA. 2011.

[9] Glynn JR, Carael M, Buve A, Musonda R.M. and M. Kahindo. Study Group on the Heterogeneity of HIVEiAC. HIV risk in relation to marriage in areas with high prevalence of HIV infection. *J Acquir Immune Defic Syndr.* **Vol.33**: 526-35. 2003.

[10] Hilbe, J.M.,(2009).*Logistic Regression Models.* New York: Chapman and Hall/CRC.

[11] Hens, N., Aerts, M., Shkedy, Z., Theeten, H., Van Damme, P. and Th. Beutels. Modelling multisera data: the estimation of new joint and conditional epidemiological parameters. *Statistical in Medicine.* **Vol.7**: 2651-2664. 2007.

[12] INSIDA. National Survey on Prevalence, Behavioral Risks and Information about HIV and AIDS. MOZAMBIQUE. Key Findings. 2009.

[13] Lingappa, J. R., Lambdin, B., Bukusi, E. A., Ngure, K., Kavuma, L., Inambao, M., Kanweka, W., Allen S., Kiarie, J. N., Makhema, J., Were, E., Manongi, R., Coetzee, D., Bruyn, G., Delany-Moretlwe, S., Magaret, A., Mugo, N,. Mujugira, A., Ndase, P. and C. Celum. Regional Differences in Prevalence of HIV-1 Discordance in Africa and Enrollment of HIV-1 Discordant Couples into an HIV-1 Prevention Trial. *journal.pone.0001411.* PLoS ONE **3(1)**: e1411. doi:10.1371. 2008.

[14] Lurie, M. N., Williams, B. G., Zuma, K., Mkaya-Mwamburi, D., Garnett, G. P., Sweat,M. D., Gittelsohn, J., and S. S. A. Karim. Who infects whom? HIV-1 concordance and discordance among migrant and non-migrant couples in South Africa. **Vol.17**:2245-2252. 2003.

[15] McMillan, G. and T. Hanson. SAS Macro BDM for Fitting the Dale Regression Model to Bivariate Ordinal Response Data. *Journal of Statistical Software.* **Vol.14**, Issue 2. 2005.

[16] McCullagh, P. and J.A. Nelder.(1989).*Generalized Linear Models.* Chapman & Hall. London.

[17] Molenberghs, G. and G. Verbeke.(2005).*Models for Discrete Longitudinal Data.* New York: Springer.

[18] Muthen, B. O., and A. Satorra.(1995).*Complex sample data in structural equation modelling. In P. Marsden (ed.). Sociological Methodology. American Sociological Association.* Washington, DC:

[19] Quinn, T.C., Wawer, M.J., Sewankambo, N., Serwadda, D., Li, C., Wabwire-Mangen, F., Meehan, M.O., Lutalo, T. and R.H. Gray. Viral load and heterosexual transmission of human immunodeficiency virus type 1. Rakai Project Study Group.*New England Journal of Medicine* **Vol.342**: 921-929. 2000.

[20] SAS Institute, Inc. (2008).*SAS/STAT®User's guide. The GLIMMIX Procedure (Book Excerpt).* SAS Online Doc®9.2. Cary, North Carolina: SAS Institute, Inc.

[21] Thomas, S. L. and R. H. Heck. Analysis of Large-Scale Secondary Data in Higher Education Research: Potential Perils Associated With Complex Sampling Designs.*Research in Higher Education.* **Vol.42**. No. 5. 2001

[22] World health organization.(2007).*HIV/AIDS epidemiological surveillance update for the African region.*

6 Appendix

Sample Weights

Since the allocation of the sample to the different provinces and to their urban-rural areas was not proportional, sampling weights were obtained to ensure the actual representativeness of the sample at the national level. These were calculated based on sampling probabilities separately for each sampling stage and for each Enumeration Area(cluster) as follows [8]:

P_{1hi}: First-stage sampling probability of the i^{th} EA in province(stratum) h
P_{2hi}: Second-stage sampling probability within the i^{th} EA

Let a h be the number of EA selected in province h, M_{hi} the number of households according to the sampling frame in the EA i^{th}, and $\sum M_{hi}$ the total number of households in the province h. The probability of selecting EA i^{th} is given as follows:

$$\frac{a_h M_{hi}}{\sum M_{hi}} \tag{20}$$

Let b_{hi} be the proportion of households in the selected segment compared to total number of households in EA i and in province h if the EA is segmented, otherwise $b_{hi} = 1$. Then the probability of selecting EA i in the sample is given by:

$$P_{1hi} = \frac{a_h M_{hi}}{\sum M_{hi}} b_{hi} \tag{21}$$

Let g_{hi} be the number of households selected in determined EA. The second stage's selection probability for each household in that EA is obtained by:

$$P_{2hi} = \frac{g_h}{M_{hi} b_{hi}} \tag{22}$$

The overall selection probability of each household in EA i of stratum h is therefore the product of the two selection probabilities:

$$P_{hi} = P_{1hi} \times P_{2hi} \tag{23}$$

The weight for each household in EA i of province h is the inverse of its selection probability:

$$W_{hi} = \frac{1}{P_{hi}} \tag{24}$$

Household as well as individual sampling weights were obtained by adjusting the above calculated design weight to compensate for household non-response and individual non-response, respectively. Individual sample weights were obtained for men and women based on the male and female response rates. These weights were further normalized at the national level to achieve the number of un-weighted cases equal to the number of weighted cases for both households and individuals at the national level. In addition to Household as well as individual

sampling weights, HIV weights were calculated since it was possible for a respondent to participate in the interview, but not in the HIV test. The HIV weights were calculated by using the individual sample weights with a further adjustment for non-response to the HIV test [8].

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Statistical Methodology for HIV Serodiscordance among Couples: The case of Mozambique

Richting: **Master of Statistics-Biostatistics**
Jaar: **2012**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Juga, Adelino Jose C

Datum: **14/09/2012**