2012•2013
FACULTY OF SCIENCES
Master of Statistics: Biostatistics

Masterproef

Investigating Migration as an important risk determinant of HIV infection among public school educators in South Africa

Promotor :
Prof. dr. Ziv SHKEDY

Promotor :
Prof. KHANGELANI ZUMA

Geraldine Agiraembabazi

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

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

universiteit hasselt
KNOWLEDGE IN ACTION
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

universiteit hasselt | Maastricht University

2012•2013
FACULTY OF SCIENCES
Master of Statistics: Biostatistics

Masterproef

Investigating Migration as an important risk determinant of HIV infection among public school educators in South Africa

Promotor :
Prof. dr. Ziv SHKEDY

Promotor :
Prof. KHANGELANI ZUMA

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

universiteit hasselt | Maastricht University

Acknowledgements

First and foremost I would like to thank the Almighty God for His grace that has sustained me. '*It is God who arms me with strength and keeps my way secure. He makes my feet like the feet of a deer; he causes me to stand on the heights. He trains my hands for battle; my arms can bend a bow of bronze. You make your saving help my shield, and your right hand sustains me; your help has made me great. You provide a broad path for my feet, so that my ankles do not give way*' Psalm 18:32-36.

I would like to thank VLIR for all the financial support you provided that made this possible. My supervisors Prof. Dr. Ziv Shkedy and Dr. Khangelani Zuma, thank you for all the support and the guidance you gave me in writing this Thesis and to all my professors for the knowledge imparted throughout the two years of this Master program. My appreciation also goes to HSRC who allowed me to use their dataset.

I am also very grateful to my parents, siblings and friends for all the prayers, the phone calls and all the support that you provided. I would not have made it without your support. To all my classmates and colleagues at UHasselt for the knowledge and friendship that we shared, thank you!

Geraldine Agiraembabazi

September, 2013

ABSTRACT

Objective: To investigate the risk determinants of HIV, with focus on migration status of the educator, among South African public school educators.

Methodology: It was imperative to investigate the individual relationship of each of the potential risk factors with the response first. Survey analysis techniques for multilevel clustered data including Survey design based logistic regression, GEE and GLMM were fitted to the data. Comparisons were made to explore how the inferences differ under the different methods of analysis. Finally, multiple imputation technique was employed to deal with data missingness.

Results: The univariate results showed that mobility/migration is an important risk determinant of HIV. When adjusted for other potential risk factors, the different methods applied showed contradicting results. The estimates obtained when missingness was ignored and when it was taken into account are close.

Conclusion: Most of the methods that were applied in this Thesis did not find a statistically significant effect of migration/ mobility on HIV status. The reason for the contradiction is because the predictors are related.

Recommendations: It is recommended that the variable selection be done using data mining techniques like the lasso, ridge and elastic-net (Hastie et al, 2008) which deal with the correct model variable selection for such correlated predictors. The Department of Education in South Africa should also make an effort to put in place a systematic deployment structure where educators will get posted near their homes in order to reduce on their mobility and time spent away from their families.

Keywords: *Clustered data, Generalized Estimating Equations, Generalized Linear Mixed models, HIV, Multilevel, Survey data.*

Contents

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

1.1 Background

Human mobility has always been a major driving force in epidemics of infectious diseases. The Human Immunodeficiency Virus (HIV) like any other infectious disease follows the movement of people (Hunt, 1989; Caldwell et al., 1997). Several studies in the literature have shown a significant relation between mobility and the risk of HIV and other infectious diseases (Pison et al, 1993, Nunn et al, 1995, Lydie et al, 2004). A study in Tanzania demonstrated an association between people's mobility behaviour, their sexual risk behaviour and their HIV status (Kishamawe et al, 2006). This association holds for both men (Jochelson et al, 1999, Lurie et al, 2003) and women (Zuma et al, 2003) in South Africa.

Sub-Saharan Africa is one of the worst affected regions in the whole World (UNAIDS 2004). In a mobility study conducted in Cameroon over a period of one year, HIV prevalence was found to be highest among men who had been away from home for more than a month (7.6%), been away for a period less than a month (3.4%) and lowest among those who had never been away the previous year (1.4%) (Lydie et al, 2004). The significant association was related to high risky sexual behaviour among migrant men whilst away from their stable sexual partners. The association between migration and HIV is more likely to be a consequence of the conditions and structure of the migration process than the actual dissemination of the virus along corridors of migration (Decosas et al., 1995). It is the combination of migration and high risk behaviour with people who are already carriers that is central to the topic (Skeldon, 2000). A longitudinal cohort study conducted in a rural Ugandan population found that change of residence was strongly associated with an increased risk of HIV-1 infection in this rural population (Nunn et al, 1995). This study found that the sero-prevalence rates were 5.5% for 2,129 adults who had not changed address since the previous survey, 8.2% for 336 who moved within the village, 12.4% for 128 who moved to a neighbouring village, 11.5% for 1,130 who had left the area and 16.3% for 541 who had joined the study area during the previous 3 years (P-value < 0.001).

In dealing with the topic of migration and HIV, one issue that arises is that; in the movement of people who engage in high-risk behaviour are groups that might not normally be classified as "migrants" (Skeldon, 2000); for instance, public school educators that are deployed to places far from their own homes who are the main focus of this study. The very life situation of migrant workers in a range of contexts (*including migrant educators*) renders them particularly vulnerable to HIV (Hunt, 1989). In South Africa, a study conducted in rural KwaZulu-Natal confirmed that migration does play an important role in spreading HIV (Lurie et al., 2003). The study examined couples in which the male partner was either a migrant worker or not. The study found that in HIV discordant couples, female partners of migrant men were as likely to be infected with HIV as their male partners. This indicates that migration also puts female partners of migrant workers at high risk of HIV infection due to the possibility of acquiring extra sexual partners whilst their male partners are away. People are not only vulnerable to HIV infection by the risk behaviour of their partners but also by their own risk behaviour when left behind (Kishamawe et al, 2006).

1.2 Research Objective

Research has identified specific sectors as playing a critical role in the risk and vulnerability of migrants to HIV infection. However, no study to date has explored the effects of migration among educators even though educators are often deployed to work in areas away from their families. This study aims at investigating the risk determinants of HIV, with focus on migration status of the educator, among South African public school educators. This was done by using the HIV status (positive or negative) as the outcome of interest. The data used is from a national [1]second-generation surveillance survey among educators in South African public schools in 2004.

[1] It involves regular systematic collection, analysis and interpretation of information used in tracking and describing changes in the HIV/AIDS epidemic over time. It also collects information on risk behaviors using them to warn or explain changes in levels of infection (WHO, 2003).

CHAPTER 2: THE DATA

2.1 Data description

The data used in this study is from a second-generation surveillance cross-sectional survey conducted among educators in public schools in South Africa in March 2004. The South African Department of Education provided the list of public schools to draw the sample from. The sampling frame contained 26 713 schools with an estimated total of 356 749 educators. A sample of 1 766 public schools was drawn and is considered in this Thesis. The selected 1 766 public schools formed the primary sampling units and are shown in Figure 2.1.

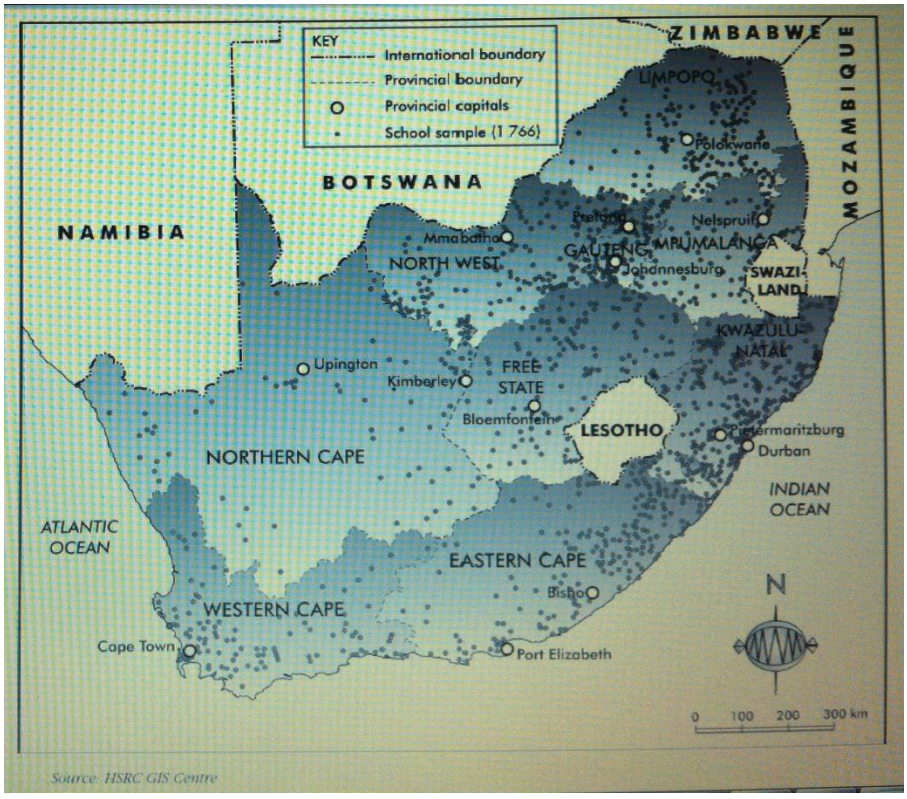


Figure 2.1: *Distribution of the sampled schools*

The sample design was stratified into 52 educational districts. The schools were sampled proportional to size, with the number of educators in each school used as a measure of size. At each school, all educators (ultimate sampling units) present on the day of the survey were invited to participate in the study.

2.2 Data collection

A detailed questionnaire was administered that elicited information related to bio- and socio-demographic information (Shisana, et al 2004). Recently retired nurses conducted the interviews (data collection) after undergoing rigorous training.

2.3 Biological tests

The study participants were given a choice to provide blood specimen or oral fluid for HIV testing. This was done in order to maximise participation. Blood specimens were tested for HIV antibodies on the Abbott AXSYM third generation HIV 1 / 2 g0 testing system (Abbott Laboratories, Abbot Park, IL). Oral fluid specimens were obtained by using the 'OraSure' oral fluid collection device (OraSure Technologies, Inc.). The oral fluid specimens were tested using the Vironostika HIV Uni-Form II Oral Fluid testing system. Only a single test was conducted per specimen. In the case of blood specimens that gave borderline readings, the result was confirmed by using the Biorad HIV test. Specimens that remained inconclusive on the Biorad system were reported as 'indeterminate'. Oral fluid specimens that were inconclusive were not repeat-tested and were reported as indeterminate.

2.4 Anonymity and Quality control

Informed consent from the participant was obtained separately for agreeing to participate in the interview and for providing a specimen for HIV testing. All specimens were linked to the questionnaire by means of a bar code. This enabled an HIV result to be linked to data from the questionnaire, but no HIV result could be linked back to any individual, thus ensuring anonymity and confidentiality.

Quality control started with the meticulous process of questionnaire design to the extensive training of fieldworkers. Field supervisors ensured that the interviewer team visited the correct school, assisted in setting up the interviewing process and checked the completed questionnaires for obvious errors. A team of editors in the office went through the questionnaires, coded open-ended questions and ensured that the geographic and other details were correct. Finally, voluntary counselling and testing (VCT) for HIV testing was not provided as part of the study. Instead, those

interested in finding out about their HIV status were given a referral card to go to the nearest primary health care centre that provided VCT services free of charge.

2.5 Weighting of Samples

Weighting procedures were done in order to take into account the realised samples and the non-responses (Shisana et al., 2004). The objective in applying nonresponse factors in survey weights is to attenuate bias due to differential nonresponse across sample elements (Heeringa et al, 2010). More so, given the design of the study, the probabilities of selection were unequal and this had to be taken into account by using weights in order to avoid possible selection bias. Four steps were required to weigh the data (see Shisana et al., 2004 for details of this procedure). As noted by Molenberghs 2012, including weights that properly reflect stratification is a first and very important step towards a correct analysis.

2.6 The Response Variable

Binary response data is often encountered in epidemiological, sociological and medical research. The type of outcome is a key characteristic in the choice of the appropriate models to use in the analysis. For discrete data, interest is usually in the association between the response and the independent predictor variables. This association gives an idea of the important risk factors for the outcome or disease. In this study, the HIV status was used as the outcome of interest and was defined as;

$$\text{Response} = \begin{cases} 1, & \text{HIV+} \\ 0, & \text{HIV-} \end{cases}$$

2.7 Potential Risk Determinants

Table 2.1 presents the description of the predictor variables/potential risk factors that were considered in the study. These variables/risk factors were investigated for their relation to the response of interest in an attempt to answer the research question. In addition to the predictor variables, a weight variable, a clustering variable and a stratification variable were also used in the analysis.

Table 2.1: *Description of the demographic and mobility factors*

Variable	Description (Coding and categories)	Total	%
Sex of the respondent:			
	1 Male	6731	31.5
	2 Female	14215	66.6
Race:			
	1 African	14643	68.6
	2 White	2840	13.3
	3 Coloured	2748	12.9
	4 Asian	630	2.9
Age in years:			
	1 18-24	274	1.3
	2 25-34	5200	24.3
	3 35-44	9078	42.5
	4 45-54	5251	24.6
	5 55 and above	1065	5.0
Marital status:			
	1 Married or cohabiting	15016	70.3
	2 Single	5559	26.0
Location of the institution:			
	1 Urban formal	9037	42.3
	2 Urban informal	1436	6.7
	3 Non-urban	10775	50.4
Away:	*In the last 12 months, have you been away from home for more than a month?*		
	1 Yes	225	10.6
	2 No	18196	85.2
Post-near:	*After completing teacher training did you work closer to your family or posted to a different place?*		
		348	16.3
	1 Stayed in the same area	2505	11.7
	2 Moved to a different area		
Move:			
	When posted, did your family move with you?	1425	6.7
	1 Moved with me	1274	6.0
	2 Stayed behind	794	3.7
	3 No family of my own at the time		

CHAPTER 3: STATISTICAL METHODOLOGY

Hierarchical or clustered data structures are usually encountered in many applications/studies, such as medical and sociological research. These studies often involve the analysis of data with complex patterns of variability, such as multilevel, nested sources of variability (Dai et al, 2006) and in many of these studies the outcome of interest is usually binary. The standard method of analysis of data with binary outcomes is a logistic regression model which is a generalized linear model (GLM, Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989; Agresti, 2002) with the logit link. However, in general, especially when dealing with survey data, this model is too simplistic but there is a multitude of models which do consider clustering (Aerts et al, 2002) including marginal and subject-specific models.

In multilevel designs, subjects are observed nested within larger units, for example, individuals (level 1) nested in families (level 2) and families nested in communities (level 3). In this study, we have teachers/educators (level 1) nested in schools (level 2) and schools nested in districts (level 3), Figure 3.1.

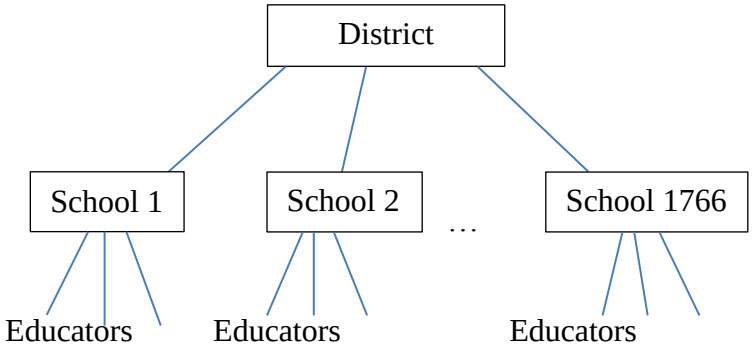


Figure 3.1: *Multilevel structure of the data*

Appropriate scientific methodology for the analysis of such data should account for the potential clustering between observations from the same group/cluster, brought about by the unobserved heterogeneity at each level of the hierarchy. Many characteristics measured on sample elements within naturally occurring clusters, such as children in a school classroom or adults living in the same neighbourhood, are correlated (Heeringa et al, 2010). For example, the HIV prevalence among educators from the same school could be related because they share an unobserved "school effects" such as access to information on HIV, access to health care, socioeconomic status and peer

influence regarding sexual behavior. The models applied here include the Design-based survey logistic regression model; Generalized Estimating Equations (GEE) in the marginal model family and the Generalized Linear Mixed Model (GLMM) for binary outcomes in the cluster specific model family.

3.1 Survey Design-based logistic regression

As mentioned above, the standard approach for the analysis of the relationship between a binary dependent variable and a set of explanatory variables is the logistic regression model (Agresti, 2002, Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). Like ordinary linear regression, logistic regression extends to models with multiple explanatory variables (Agresti, 2002). A multiple logistic regression model is of the form

$$\text{Logit } [\pi(\mathbf{x})] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad \text{(Model 1)}$$

Where $\pi(\mathbf{x})$ denotes the probability of being HIV+ (i.e., $\pi(\mathbf{x})$ =P [Y =1]) at the values $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ of the p predictors or risk factors. The regression coefficient/parameter β_i refers to the effect of *x* on the log odds that the response Y = 1 controlling for the other *x* in the model and Y ~ Bernoulli [$\pi(\mathbf{x})$]. Unlike the linear regression model for normally distributed Y, there is no direct solution such as the method of least squares to estimate the regression coefficients in the logit model. An iterative estimation procedure such as the Newton–Raphson (Agresti, 2002) is used instead.

The survey-design-based logistic regression is an extension of the standard logistic regression which enables incorporation of complex survey sample design information including stratification, clustering and unequal weighting. For simple random samples, the logistic regression model parameters and standard errors can be estimated using the method of maximum likelihood where the likelihood function

$$\mathrm{L}(\boldsymbol{\beta}|\mathbf{x}) = \prod_{i=1}^{n} \pi(x_i)^{yi} \, [1 - \pi(x_i)]^{1-yi}$$

is based on the binomial distribution with $\pi(x_i)$ linked to the regression model coefficients through the logistic cumulative distribution function:

$$\pi(x_i) = \frac{\exp(x_i\boldsymbol{\beta})}{1 + \exp(x_i\boldsymbol{\beta})}$$

When a complex survey sample design has been used, application of maximum likelihood estimation (MLE) procedures is no longer straightforward for reasons such as unequal selection probabilities (and responding) of sample observations and requiring the incorporation of sample weights. More so, the stratification and clustering of observations violates the assumption of independence of observations that is crucial to the standard MLE approach to estimating the sampling variances of the model parameters (Heeringa et al, 2010).The pseudo-maximum likelihood estimation (PMLE) technique proposed by Binder (1981) is now the standard method for logistic regression modeling in all of the major software systems that support analysis of complex sample survey data. In this Thesis, this method was applied using the SAS procedure SURVEYLOGISTIC (SAS Institute, 2011) which is designed to handle sample survey data, and thus incorporates the sample design information into the analysis. It is imperative to incorporate the sample design in the data analysis in order to make statistically valid inferences for the population. Therefore, the survey-design-based logistic regression was adopted in this Thesis as one of the main analyses and the standard logistic analysis in which the design is assumed to be simple random sampling (SRS) was included for comparison purposes.

3.2 Marginal Model

In the clustered setting, measurements are taken on subjects that share a common category or characteristics that lead to correlation (SAS Institute, 2002). In some cases, ordering of the subjects within a group may be of interest leading to the so-called longitudinal studies while in other cases, the order of subjects within a group may not be of interest. No matter the case, the assumption of independent outcomes is implausible since measurements on members of the same group are likely to be more related than those from members of different groups. Understanding the disease clustering is important for providing insights into the risk factors operating within different levels of clusters. If one is interested in the overall population-averaged (PA) effects, then PA/marginal models are most appropriate. These models relate the covariates directly to the marginal

probabilities (Aerts et al, 2002) and the marginal model considered here is the Generalized Estimating Equations (GEE) described in the following section.

3.2.1 Generalized Estimating Equations (GEE)

Even though a variety of flexible models exist for the analysis of uncorrelated binary data, when dealing with correlated or hierarchical data, the standard methods are too simplistic and don't reflect the data generating mechanism. So much progress has been made in the analysis of correlated data (see for example, Molenberghs and Verbeke, 2005). One of the methods used to estimate the unknown regression parameters is Maximum likelihood (ML); but it can be unattractive due to excessive computational requirements, especially when high dimensional vectors of correlated data arise (Aerts et al, 2002). The use of classical maximum likelihood methodology necessitates the full specification of the joint distribution for the response vector **Y**. In the context of discrete data, one needs to specify the first-order moments as well as all higher-order moments (Molenberghs and Verbeke, 2005) which are often computationally restrictive for high-dimensional vectors of correlated data.

Alternative methods like GEE (Liang and Zeger, 1986) and pseudo-likelihood (PL, Molenberghs and Verbeke 2005, Zhao and Joe 2005, Arnold and Strauss 1991, Geys, Molenberghs and Ryan, 1997) have been proposed and implemented in most statistical software. These methods allow for within-cluster dependence and are more practical computationally compared to full/classical likelihood. The GEE method estimates the variances and covariances in the random part of the multilevel model directly from the residuals, which makes them faster to compute than full ML estimates (Hox, 2010). It can also be used for logistic modeling of complex survey data (Lehtonen and Pahkinen, 2004) and for binary data, one can use a GEE approach to account for the correlation between responses of interest for subjects from the same cluster (Diggle et al., 1994). GEEs estimate the parameters associated with the binary responses and phrase the working assumptions about the association between pairs of outcomes in terms of marginal correlations (Molenberghs and Verbeke, 2005). When adopting GEE1, one does not use information of the association structure to estimate the main effect parameters (Geys, Molenberghs and Ryan, 2002) and therefore, GEE1 yields consistent main effect estimators, even when the association structure is mis-specified. However, severe misspecification (Aerts, Declerck and Molenberghs, 1997) may bias your

estimates and may also lead to a breakdown in the iterative procedure if the correlation matrix is not positive-definite (Sun, Shults, and Leonard, 2009). More so, correct estimation of the correlation also improves efficiency of the estimated regression parameters (Wang and Carey, 2004).

The GEE methodology is based on solving score equations and to account for the correlation structure, the score function/estimating equation that is solved is of the form

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \mu i}{\partial \beta'} (A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}})^{-1} (\mathbf{y}_i - \mu_i) = \mathbf{0}$$

Where $R_i$the correlation matrix is often referred to as the *working* correlation matrix (that could be wrong) and A_i is the diagonal matrix with the marginal variances on the main diagonal. The correlation between measurements can be assumed as, for example, Corr (Y_{ij}, Y_{ik}) = 0 for independence, Corr (Y_{ij}, Y_{ik}) = ρ for exchangeability or Corr (Y_{ij}, Y_{ik}) = ρ_{jk} for unstructured; (j ≠k) working assumptions. The variance function of the observations within clusters is modeled by

$$V_i = \left(A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}} \right)^{-1}$$

And $\mathbf{R_i}(\boldsymbol{\alpha})$ is again the working correlation matrix to model the dependence between within-cluster observations expressed in terms of $\boldsymbol{\alpha}$ a vector of unknown parameters. In this Thesis, the GEE methodology was applied to estimate the average effects at the school level (population average effects) and the results were compared to those from the other methods applied.

3.3 Cluster-Specific Model

Cluster-specific models are differentiated from population-averaged models by the inclusion of parameters that are specific to the cluster (Aerts et al, 2002). For correlated binary data, the parameters of the cluster-specific and of the population-averaged models describe different types of effects of the covariates on the response probabilities (Neuhaus, 1992). With the cluster specific approach, the response probabilities are modeled as a function of covariates and parameters specific to a cluster (random effects). Interpretation of fixed effect parameters in these models is conditional on a constant level of the cluster-specific parameter. The next section gives a description of the cluster-specific model (GLMM) that was applied here.

3.3.1 Generalized Linear Mixed Model (GLMM) for binary response data

Breslow and Clayton (1993) and Wolfinger and O'Connell (1993) extended the generalized linear modeling (GLM) framework to the so-called generalized linear mixed model (GLMM) in which the correlation is accounted for by use of random effects. The GLMM is the most frequently used random-effects model for discrete outcomes (Molenberghs and Verbeke, 2005; Aerts et al, 2002). GLMMs for data having a hierarchical grouping as described in Figure 3.1 above are called multi-level models (Goldstein, 2003; Raudenbush and Bryk, 2002) and the random effects enter the model at each level of the hierarchy (Agresti, 2002). The random effects incorporate correlation between the repeated observations within each cluster and variation between clusters (Wu, 2010).

The general formulation of the GLMM is as follows. Let y_{ij} be the outcome measured for the ith subject (educator), in cluster (school) j and $\boldsymbol{Y}_i$ is the *ni*-dimensional vector of all measurements available for cluster j. It is assumed that, conditionally on *q*-dimensional random-effects $\boldsymbol{b_j}$, assumed to be drawn independently from the *N* (**0**, ***D***), the outcomes y_{ij} are independent with densities of the form

$$f_j\,(y_{ij}\,|\boldsymbol{b}_j, \boldsymbol{\beta}, \varphi) \;=\; exp\,(\varphi - 1[y_{ij}\theta_{ij} \;-\; \psi(\theta_{ij}\,)] \;+\; c(y_{ij}, \varphi))$$

with $\eta(\mu_{ij}) \;=\; \eta[E(Y_{ij}\,|\boldsymbol{b}_j)] \;=\; \boldsymbol{x}'_{ij}\boldsymbol{\beta} \;+\; \boldsymbol{z}'_{ij}\boldsymbol{b}_j$ for a known link function $\eta(\cdot)$, with $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ *p*-dimensional and *q*-dimensional vectors of known covariate values, with $\boldsymbol{\beta}$ a *p*-dimensional vector of unknown fixed regression coefficients, and with φ a scale parameter.

In this Thesis, a 2-level model for binary outcomes was considered with random intercepts at the school level and fixed effects for the explanatory variables. This model is of the form

$$\text{Logit}\,(\pi_{ij}) = \mathbf{X}'_{ij}\boldsymbol{\beta} + \mu_j \qquad \text{(Model 2)}$$

Where u_j is the random intercept at the school level and is assumed to be normally distributed with mean 0 and variance, σ_u^2 i.e. $u_j \sim N(0, \sigma_u^2)$; the average probability of being HIV+ is assumed to vary randomly across schools and it is assumed that $Y_{ij} \sim$ Bernoulli(π_{ij}).

Model 2 was further extended to a 3-level model including random intercepts at both the school and district levels.

$$\text{Logit}(\pi_{ijk}) = \mathbf{X}_{ijk}^{'}\boldsymbol{\beta} + \mu_{jk} + \nu_k \qquad \text{(Model 3)}$$

Where, $\pi_{ijk} = P[Y_{ijk}=1 | \nu_k, \mu_{jk}]$; $\mu_{jk} \sim N(0, \sigma_u^2)$ is school-level intercept term, and $\nu_k \sim N(0, \sigma_v^2)$ is a district-level intercept term and it is assumed that $Y_{ijk} \sim$ Bernoulli(π_{ijk}). Again the average probability of being HIV+, is assumed to vary randomly distributed across schools and across districts.

Models 2 and 3 were fitted using the GLIMMIX Procedure in SAS (SAS Institute, 2005) taking into account the design aspects of the study i.e. weights and clustering.

3.4 Multiple Imputation

Weighting adjustments for nonresponse do not compensate for the otherwise complete cases that are lost in the analysis due to item-missingness on one or more of the analysis variables. Complete case analysis requires the assumption that the missing data are missing completely at random (MCAR). However, from the standpoint of being able to effectively and practically address item-missing data in a survey data set, a missing at random (MAR) mechanism is the more reasonable assumption (Heeringa et al, 2010). Multiple imputation technique (Rubin, 1987; Little and Rubin, 2002) was used to address the problem of missingness. Missing values are filled in *m* times to generate *m* complete data sets that are generated from a plausible model. A further step involves combining the results from the analyses giving valid statistical inferences that properly reflect the uncertainty due to missingness.

3.5 Statistical software

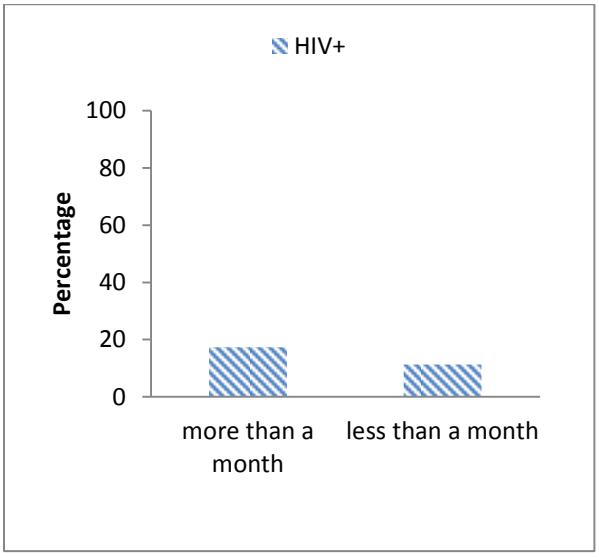
All analyses were done using SAS 9.3 and R 3.0.1. The data management was done in STATA 11 and Microsoft Excel 2010 and statistical significance was taken at 5% level.

CHAPTER 4: RESULTS

4.1 Exploratory Data Analysis

A total of 21,358 educators participated in the study. The majority (66.6%) of educators were females and most were Africans (68.6%) followed by Whites (13.3%), Coloureds (12.9%) and Asians (2.9%). This is reflective of the racial composition of the South African population. Majority of the educators were either married or cohabiting and were between 25 and 54 years of age. 10.6% of the educators had been away from home for more than a month in the last 1 year (see Table 2.1).

Figure 4.1 presents the descriptive statistics for the mobility factors in relation to HIV prevalence. It is observed that those who had been away from home for more than a month in the last 1 year had a higher HIV prevalence (*Fig.4.1 a*). Similarly, the percentage of HIV+ educators was higher for those who moved away from home after completing teacher training (*Fig.4.1 b*) and also for those whose families stayed behind (*Fig.4.1 c*).



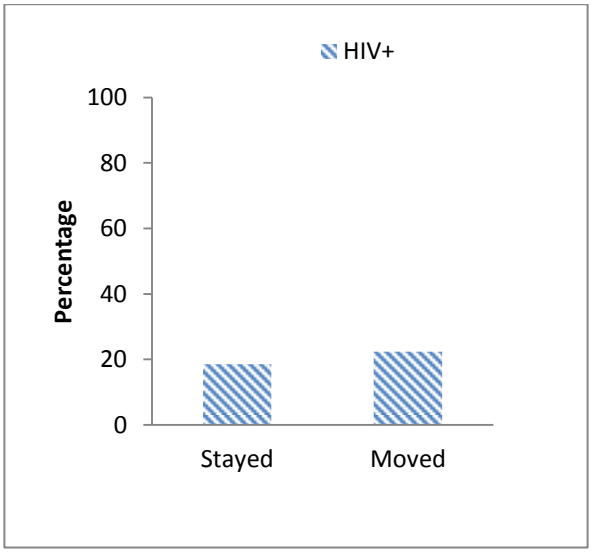
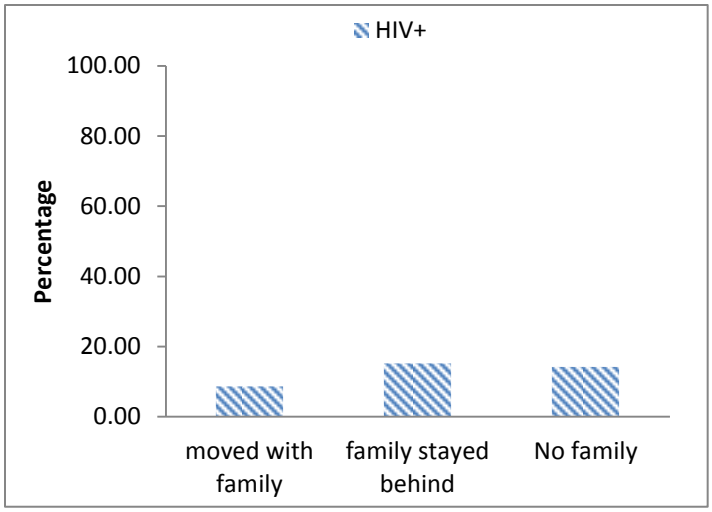


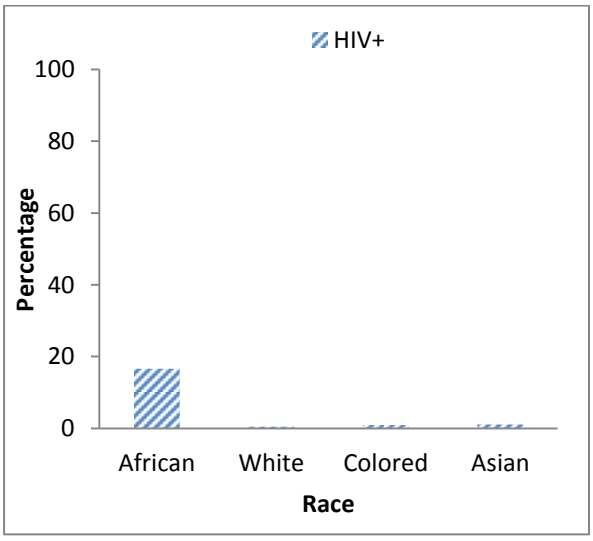
Figure 4.1 **a)** *In the last 12 months, have you been away from home for more than a month?*

b) *After completing teacher training did you work closer to your family or posted to a different place?*



c) *When posted, did your family move with you?*

Figure 4.2, 4.3 and 4.4 present the descriptive statistics for the socio-demographic factors. These statistics indicate that the HIV prevalence varies among the different factors. HIV prevalence was highest among Africans (16.7%) compared to other races (*Fig. 4.2 a*). Marital status was grouped into two categories including the married or cohabiting and single. The percentage of the HIV+ educators was found to be higher for those who were single (22.3% vs. 8.2%).



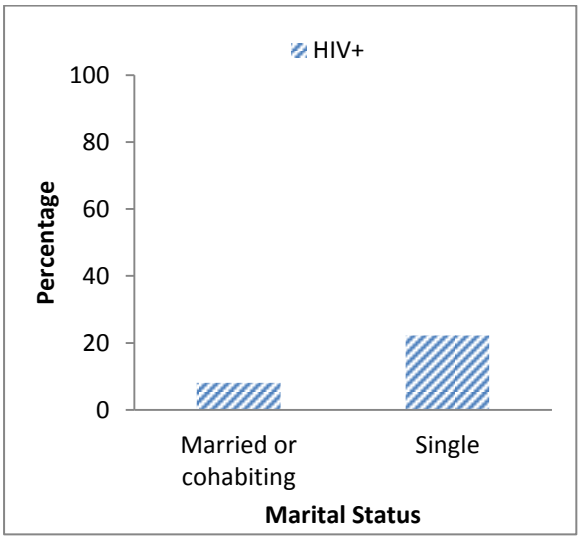
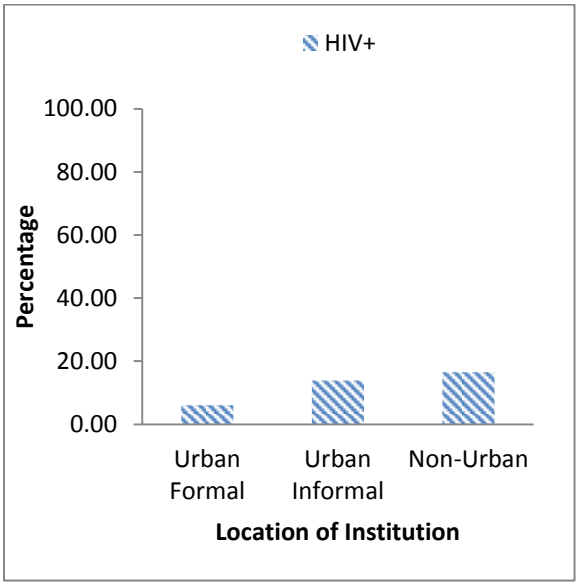


Figure 4.2: **a)** *Race of the participant* **b)** *Marital status in two categories*

According to the location of the institution where the educators worked, those in the formal urban areas had the least HIV prevalence (6%). For the non-urban and urban-informal, the prevalence was

higher and almost the same (13.8% vs. 16.2%). According to the living arrangement, educators that were living with a partner (husband or wife) had the least prevalence (*Fig. 4.3 b*). Age of the educators was also grouped into 5 categories as indicated in Figure 4.4 (a). HIV prevalence is highest among educators between the ages of 25 to 34 years (20.1%) and is lowest for those below 25 years and above 55 years. It is however not so different between males and females (*Fig. 4.4 b*).



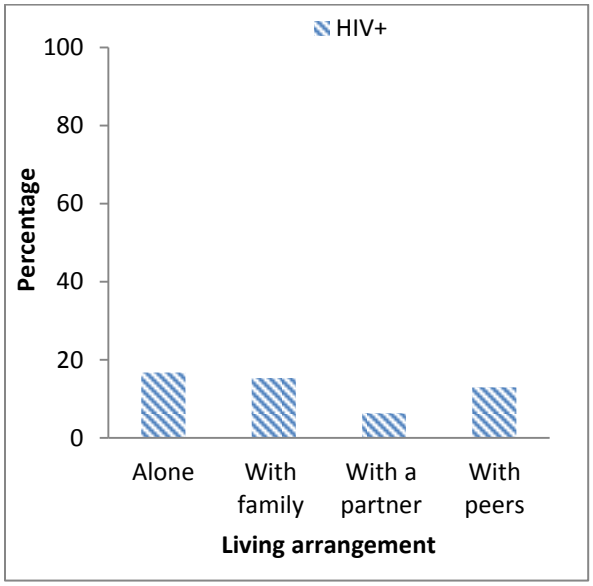
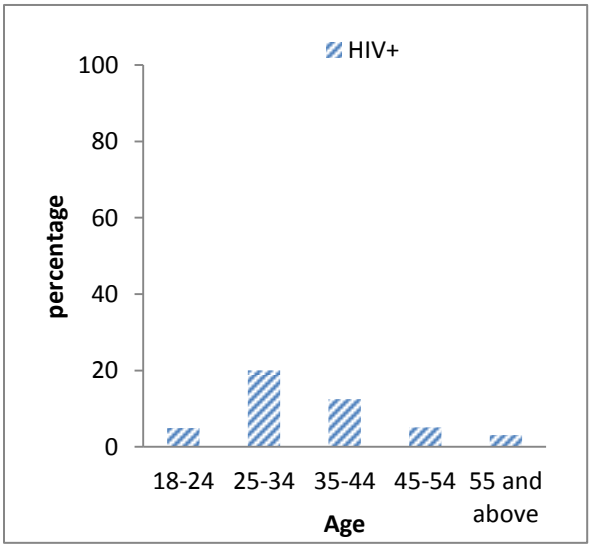


Figure 4.3 **a)** *Location of the Institution* **b)** *Present living arrangement*



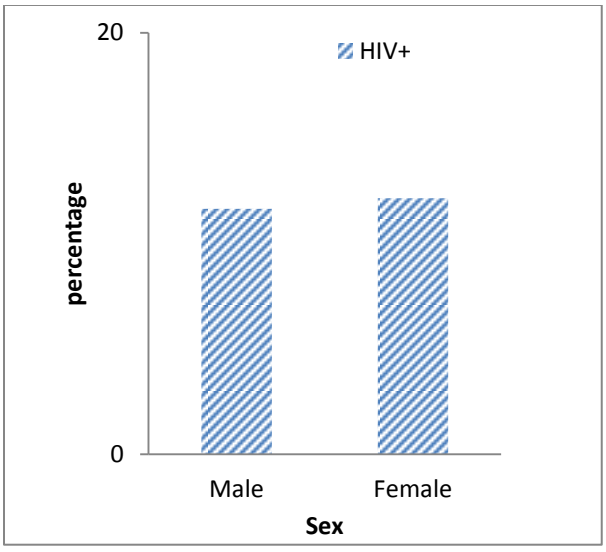


Figure 4.4 **a)** *Age grouped in five categories* **b)** *Sex of the educator*

4.2 Statistical Results

4.2.1 Univariate Analysis

Univariate (single predictor) categorical data analyses and reporting are important in their own right, and they are also important as exploratory tools in the development of more complex multivariate models (Heeringa et al, 2010). The relationship between each of the independent risk factors and HIV was first examined in a logistic regression. This was done using the SAS procedures SURVEYFREQ and SURVEYLOGISTIC taking into account all the design aspects of the study i.e. weights, stratification by district and clustering of educators within schools. Tables 4.1 and 4.2 present the results including the odds ratios and the HIV prevalences for the demographic factors and the mobility factors respectively. These odds ratios are labeled as an *unadjusted* since they are estimated with no additional controls for other factors.

The prevalence of HIV was 12.8% [95%CI: 12.1% – 13.5%] among female educators and this was not significantly difference from the prevalence among males (12.7 [95%CI: 12.1% - 13.3%]). However, the prevalence of HIV was significantly higher among African educators (16.3% [95%CI: 15.9%- 16.7]) compared to any other race. This was also reflected in the EDA in Figure 4.2 (a).

HIV infection was more pronounced among educators aged between 25 and 34 years followed by educators aged between 35 and 44. The prevalence of HIV was about three times higher among educators who were single compared to those who were married or cohabiting (Table 4.1). Similarly, educators who were living alone, with family or peers had a considerably high HIV prevalence than educators living with their partners.

Table 4.1: *Demographic profiles of educators and HIV prevalence*

DEMOGRAPHICS	Total	HIV+%[95%CI]	Unadjusted OR [95% CI]
Sex of the respondent			
Male	5456	12.7[12.1 –13.5]	1
Female	11621	12.8[12.1 –13.3]	1.00 [0.90 – 1.13]
Race			
African	12022	16.3[15.9 – 16.7]	46.20 [24.77–88.16]
White	2165	0.4[0.3 – 0.5]	1
Coloured	2309	0.7[0.5 – 0.9]	1.70 [0.72 – 4.00]
Asian	533	1.0[0.6 - 0.14]	2.35 [0.87 – 6.35]
Age in years			
18-24	240	6.5[4.6 – 8.4]	2.16 [0.97 – 4.81]
25-34	4282	21.4[20.6 – 22.2]	8.49 [5.54 –12.98]
35-44	7444	12.8[12.3 – 13.3]	4.56 [2.99 – 6.96]
45-54	4274	5.8[5.4 – 6.2]	1.93 [1.26 – 2.95]
55 and above	842	3.1[2.5 – 3.7]	1
Marital status			
Married or cohabiting	12440	8.8[8.5 – 9.2]	1
Single	4589	22.9[22.2 – 23.7]	3.09 [2.76 – 3.46]
Location of the institution			
Urban formal	7032	6.3[5.8 - 6.8]	1
Urban informal	1120	13.9[12.8 – 15.0]	2.39 [1.82 – 3.14]
Non-urban	8860	16.8[16.3 - 17.3]	3.00 [2.50 – 3.61]
Present living arrangement			
Alone	1794	17.6[16.5 – 18.7]	2.89 [2.39 – 3.49]
With family or relatives	8454	15.8[15.3 – 16.3]	2.55 [2.23 – 2.92]
With a partner (husband or wife)	6528	6.9[16.5 – 7.3]	1
With peers/friends/co-workers	192	15.8[13.0 – 18.6]	2.55 [1.55 – 4.17]

Migrant educators are at a higher risk of HIV than educators who are not migrants (Table 4.2). Educators whom after completing teacher training had to move and take up a position in a different place from where their family was located were at a higher risk of HIV (23.5%) compared to those who stayed in the same area (18.9%). The risk of being infected with HIV was significantly higher (32%) among these educators (OR=1.32 [95%CI: 1.13 – 1.55]).

Among the educators who moved, those who moved without family or had no family were more likely to be infected with HIV than those who moved with their families. The risk was not different between the educators whose family stayed behind and those who had no family at the time. Educators who had been away for at least one month in the last 1 year had a 57% higher chance of being infected with HIV (OR=1.57 [95%CI:1.34 – 1.83]) compared to their counterparts who had been away for less time.

Table 4.2: *Mobility and HIV prevalence*

MOBILITY	Total	HIV+%[95%CI]	OR [95% CI]
After completing teacher training did you work closer to your family or posted to a different place	2939	18.9[18.1 – 19.7]	1
Stayed in the same area	2077	23.5[22.4 – 24.6]	1.32 [1.13 – 1.55]
Moved to a different area			
When posted, did your family move with you?			
Moved with me	1186	10.4[9.3 – 11.5]	1
Stayed behind	1063	15.2[13.9 – 16.5]	1.54 [1.13 – 2.09]
No family of my own at the time	657	15.2[13.6 – 16.8]	1.55 [1.11 – 2.16]
*Last 12 months, have you been **away** from home for more than a month?*			
Yes	1892	17.8[16.8 – 18.8]	1.57 [1.35 – 1.83]
No	15088	12.1[11.7 – 12.5]	1

4.2.2 Survey Design-based Logistic regression

A logistic regression model is usually built by entering predictors into the model using subject-matter criteria or significance measures of potential predictors (Lehtonen and Pahkinen, 2004). Based on the initial tests of association, all of the predictor variables appear to have significant associations with HIV, except for Sex of the respondent. These predictors appear to be good candidates for inclusion in the initial multivariate logistic regression model. However, a further analysis of associations (Table A.6) between variables revealed that some were confounders. Forward, backward and stepwise model selection techniques were used and the model including the variables Race, Age, Marital status, Location, Living Arrangement and Away (*In the last 12 months, have you been **away** from home for more than a month?*) was chosen as the final model and

was fitted. The design-based logistic regression was fitted using the SURVEYLOGISTIC procedure in SAS where the district was used as the stratifying variable and clustering of educators within schools and weights were also taken into account (Table 4.3). For comparison purposes, a naïve model; simple random sampling (SRS) model was also fitted, which does not take the study designs into account. SRS provides a comparative benchmark that can be used to evaluate the relative efficiency of the more complex designs that are common in survey practice (Heeringa et al, 2010).
Due to missing values for either the response or the explanatory variables, 16 668 out of 21 358 (78%) observations were used in the analysis. Table 4.3 presents the results from the SRS model and the design-based logistic regression model for HIV and the potential risk factors including the estimates, standard errors and odds ratios.

Table 4.3: *Design-based logistic regression model and the Naïve (SRS) model*

	Design based		**Naïve (SRS)**	
Variable	**Estimate(se)**	**OR**	**Estimate(se)**	**OR**
Intercept	-6.98(0.394)		-6.60(0.355)	
Race4(ref: White)				
African	3.24(0.323)	25.601	2.91(0.294)	18.367
Colored	0.32(0.442)	1.376	0.28(0.359)	1.317
Asian	0.73(0.504)	2.076	0.52(0.504)	1.688
Age(ref: ≥ 55 years)				
18-24	0.85(0.420)	2.337	0.56(0.374)	1.742
25-34	1.70(0.227)	5.498	1.58(0.212)	4.836
35-44	1.28(0.223)	3.611	1.21(0.210)	3.344
45-54	0.60(0.224)	1.825	0.49(0.218)	1.632
Marital status(ref: married/coh)				
Single	0.58(0.069)	1.781	0.61(0.059)	1.840
Location (ref: urban formal)				
Urban Informal	0.25(0.133)	1.287	0.28(0.105)	1.321
Non-urban	0.39(0.078)	1.477	0.51 (0.062)	1.669
Living Arr(ref: with partner)				
Alone	0.38(0.114)	1.464	0.38(0.094)	1.457
With family	0.42(0.076)	1.529	0.42(0.067)	1.520
With peers	0.52(0.268)	1.674	0.37(0.243)	1.442
Away(ref: no)				
Yes	0.04(0.088)	1.038	0.07(0.073)	1.071

The primary research question of analytical interest is whether AWAY is related to HIV prevalence after adjusting for the effects of the above listed factors. For a risk factor with more than two categories, the computation of odds ratios depends on how the risk factor is parameterized. The reference cell parameterization scheme (PARAM=REF) was used for all the risk factors and the parameter estimate for a given category represents the log odds ratio of that category versus the reference category, adjusted for the other factors in the model.

The predictor variable that depicts the movement of the educators i.e. *Away*; which is the main focus of this thesis was found to be non-significant (p-values=0.6663, 0.3417 for design-based and SRS respectively). However, under the design-based model, the odds of being HIV positive were 3.8% (odds ratio=1.038) higher for the educators who had been away from home for more than a month in the last 1 year than for those who had been away for less time, keeping other factors constant.

We also note a very big difference between African and White educators. The odds of being HIV positive for Africans is about 25 times that of whites (p-value = <.0001). This extremely high difference was depicted in the descriptive results (Figure 4.2a) where the percentage of Africans with HIV infection was higher than all the other races and in the univariate analysis (Table 4.1). Further, the educators in the age group of 25-34 had the highest odds of being HIV positive. These odds are about 5 times higher than those who are 55 years and above. Similarly, relative to married educators, the single educators had about twice the odds of being HIV positive after adjusting for the other covariates. The odds of being HIV positive were similar among educators who were living alone, with family and with peers compared to those who were living with a partner. And those who worked in institutions located in the non-urban areas had about 3 times the odds of being infected with HIV compared to those in urban formal institutions.

The naïve estimates are close to the design-based ones, but failure to account for the survey design might declare effects significant when they are in fact not.

4.2.3 Marginal model: GEE

If interest lies in the overall HIV prevalence, then population average models are most appropriate. The robust and model-based standard errors did not differ much under the three working assumptions (exchangeable, independence, unstructured). The independence working correlation

structure assumes that there is no clustering at all and the model-based standard errors assume the weights are replications. It is imperative therefore, to use the empirically corrected standard errors, since the purely model based ones do not properly deal with the weights (Molenberghs, 2012). The GEE was applied here with an exchangeable correlation assumed between pairs of observations within a cluster. Table 4.4 gives the parameter estimates together with the empirically corrected (robust) standard errors and the odds ratios for GEE together with the results from the SRS model. The observed values of the test statistics (not shown) from the SRS model are somewhat larger and thus more liberal test results are attained. Again, the SRS results are included as a tool to demonstrate the importance of accounting for the design in the analysis.

Table 4.4: *Marginal model-GEE and the Naïve model (SRS). Parameter estimates with standard errors in parentheses and Odds ratios.*

	GEE		**Naïve (SRS)**	
Variable	**Estimate(se)**	**OR**	**Estimate(se)**	**OR**
Intercept	-7.09(0.393)		-6.60(0.355)	
Race4(ref: White)				
African	3.35(0.328)	28.446	2.91(0.294)	18.37
Colored	0.44(0.448)	1.547	0.28(0.359)	1.317
Asian	0.82(0.515)	2.261	0.52 (0.504)	1.688
Age(ref: ≥ 55 years)				
18-24	0.85(0.431)	2.347	0.56(0.374)	1.742
25-34	1.70(0.231)	5.490	1.58(0.212)	4.836
35-44	1.29(0.226)	3.621	1.21(0.210)	3.344
45-54	0.61(0.228)	1.835	0.49(0.218)	1.632
Marital status(ref: married/coh)				
Single	0.57(0.069)	1.763	0.61(0.059)	1.840
Location (ref: urban formal)				
Urban Informal	0.26(0.133)	1.296	0.28(0.105)	1.321
Non-urban	0.40(0.080)	1.486	0.51(0.062)	1.669
Living Arr (ref: with partner)				
Alone	0.37(0.113)	1.449	0.38(0.094)	1.457
With family	0.41(0.076)	1.511	0.42(0.067)	1.520
With peers	0.53(0.264)	1.690	0.37(0.243)	1.442
Away(ref: no)				
Yes	0.04(0.088)	1.040	0.07(0.073)	1.071

Under this model, the parameter estimates represent the average effect (log odds) of a covariate on the probability of being HIV positive. This interpretation is also dependent on the level of the categorical covariate.

On average, the educators who had been away from home for more than a month in the last 1 year were 4% [OR=exp (0.039) = 1.04] more likely to be infected with HIV than those who had been away for less time , other factors being constant. This percentage was about 7% (OR=1.071) when the study design was ignored. However, this was found be statistically non-significant (p-values=0.651, 0.381 under GEE and SRS respectively).

The working correlation structure is allowed to be wrong and hence should not be over interpreted. Nevertheless, we obtain a good indication about the average correlation between educators in the same school in terms of HIV prevalence; the estimated correlation was 0.017.

The same final model (mean structure) as the design- based analysis was also fitted using GEE. A comparison of the results from the design-based logistic model (Table 4.3) and those from the GEE model above indicates that the estimates are slightly different but nevertheless close. No matter the method that is chosen, the inferential conclusions are the same.

4.2.4 Multilevel logistic regression: GLMM

Table 4.5 presents the analysis results from the 2-level and the 3-level model including the parameter estimates, the standard errors, as well as the odds ratios. In the 2-level model, the random intercept variance at the school level was estimated to be 6.817 (SE = 0.394). It was found to be statistically significant with a p-value = < 0.0001. The mobility factor *Away*, was also found to have a significant effect on HIV status (p-value=0.003). However, this was not the case for the 3-level model (p-value=0.5983).

The interpretation of the parameter estimates here is conditional on the cluster-specific intercepts. For instance, for the 2-level model, keeping other factors constant, the odds of being HIV positive versus being negative for an individual educator were increased by 6.6% for the educators that had been away from home for more than a month in the last one year, given that the school intercept equals zero($\boldsymbol{b_j}$=0).

When the analysis extended to include the random intercepts at the district level, the estimated variance components revealed that the clustering effect at the district level (σ_v^2=0.905) was small. However, a likelihood ratio test based on a mixture of chi-squares (Snijders and Bosker, 2012, p. 99) to test the need for the district level effects resulted in a p-value <.0001. From this, it is concluded that the model with random intercepts at both the school and district level should be used to account for the heterogeneity within and between clusters. A scatter plot of the Empirical Bayes estimates plotted in Figure A.1 in the Appendix shows that a few schools have higher intercepts compared to the majority meaning a higher HIV prevalence in these schools.

Table 4.5: *2-level and 3-level models. Parameter estimates with standard errors in parentheses*

	2 level GLMM		**3-level GLMM**	
Variable	**Estimate(se)**	**OR**	**Estimate(se)**	**OR**
Intercept	-9.54(0.214)		-9.45(0.786)	
Race4(ref: White)				
African	3.96(0.186)	52.31	3.84(0.650)	46.68
Colored	0.44(0.219)	1.552	0.54(0.735)	1.708
Asian	0.62(0.238)	1.855	0.49(1.171)	1.627
Age(ref: ≥ 55 years)				
18-24	1.14(0.106)	3.128	1.14(0.744)	3.137
25-34	1.98(0.060)	7.229	1.98(0.386)	7.239
35-44	1.53(0.059)	4.607	1.53(0.335)	4.615
45-54	0.75(0.061)	2.119	0.75(0.327)	2.124
Marital status(ref: married/coh)				
Single	0.57(0.017)	1.759	0.56(0.095)	1.758
Location (ref: urban formal)				
Urban Informal	0.17(0.098)	1.188	0.15(0.258)	1.165
Non-urban	0.97(0.102)	2.631	0.94(0.357)	2.558
Living Arr(ref: with partner)				
Alone	0.35(0.028)	1.424	0.35(0.114)	1.421
With family	0.35(0.019)	1.416	0.35(0.106)	1.411
With peers	0.67(0.074)	1.950	0.67(0.306)	1.948
Away(ref: no)				
Yes	0.06(0.021)	1.066	0.06(0.120)	1.065
σ_u^2	6.82(0.394)		2.39(0.165)	
σ_v^2			0.91(0.144)	

4.2.5 Data Missingness

Table 4.6 presents an analysis of the missing values by count and percentage for each of the risk factors included in the models above.

Table 4.6: *Missingness*

Univariate Statistics		
Variable	**Missing Values**	
	Count	**Percent**
Race	497	2.33
Age	490	2.29
Marital status	783	3.67
Living Arr	913	4.27
Location	110	0.52
Away	905	4.24

It is observed that the missing value percentages are generally small. Multiple imputation technique was applied to the generated datasets but the results did not differ much from those where missingness was ignored. Table A.5 in the appendix shows the results from fitting GEE under multiple imputation (MI-GEE).

CHAPTER 5: DISCUSSION AND CONCLUSION

5.1 Discussion

This Thesis aimed at investigating the factors associated with HIV among public school educators in in South Africa with a focus on migration as an important risk factor. The data used is from a national second-generation surveillance survey conducted in 2004 among educators in South African public schools. The main outcome of interest studied was HIV status of the individual. The majority of the educators were female, of ages 34 and above, married and of the African race.

The analysis involved the application of three methods where the design of the study was accounted for. However before these methods were applied, a univariate analysis was carried to assess the relationship between each of the predictor variables and the response. The word 'univariate' is sometimes used to mean a single response but here it was used to mean a single predictor and the same goes for 'multivariate'. As noted by Lehtonen and Pahkinen, 2004, the simple random sampling method can be used as a reference for the design-based option when quantifying the effects of the design complexities on analysis results. Therefore, for comparison purposes, a naïve; SRS model was also fitted, which does not take the study designs into account.

Large biomedical data sets often confront investigators with the need to address multiple levels of 'clustering' that arise from the organizational structure of the health care delivery system (Miglioretti and Heagerty, 2004). In clustered settings, the vector of responses measured for each unit results into a number of methods that extend the classical univariate analysis. The first analysis was a survey-design-based logistic regression where the district was used as a stratifying variable and the school was used as a clustering variable. The other two approaches were GEE and GLMM (cluster-specific model) that take into account the weights and clustering in the data.

The unweighted analysis (i.e. SRS) implicitly assumes (incorrectly) that the district populations are roughly equal, members of the same school have roughly the same selection probability and that other components of weights are relatively unimportant. Although the direct modeling of clustered data is statistically efficient, it will generally be important to incorporate weights in the analysis which reflect the sample design so that robust population estimates can be obtained and so that there will be some protection against serious model mis-specification (Goldstein, 2010).

The characteristics shared by individuals in a sample cluster present group similarities. This means that the amount of "statistical information" contained in a clustered sample of *n* persons is less than in an independently selected simple random sample of the same size. Hence, clustered sampling increases the standard errors of estimates relative to a SRS of equivalent size (Heeringa et al, 2010). Furthermore, as noted by Cochran 1977, relative to a SRS of equal size, stratified samples that employ proportional allocation or optimal allocation of the sample size to the individual strata have smaller standard errors for sample estimates. Real survey designs result in a "tug of war" between the variance inflation due to clustering and the variance reduction due to stratification (Heeringa et al, 2010) but in most cases, the net effect is an increase in variance as observed in Table 4.3. Along with sample stratification and clustering, weighted estimation contributes to the final design effect for a survey estimate. It is generally found in survey practice that the net effect of weighted estimation is inflation in the standard errors of estimates. When clustering is ignored, the parameter estimates are still consistent but this is not the case for standard errors. In case of a "positive" clustering effect (i.e., units within a cluster are more alike than between clusters), then ignoring this aspect of the data, just as ignoring over dispersion, overestimates precision and hence underestimates standard errors and lengths of confidence intervals (Aerts et al, 2002).

Generalized estimating equations play an important role in the analysis of repeated or clustered outcomes of a non-normally distributed type (Geys, Molenberghs and Ryan, 2002). In this Thesis, it was used as a marginal logistic regression tool for the analysis of clustered binary data. GEE captures the association among the components of the outcome vector **Y** by means of correlation. Group comparisons for example districts with high HIV prevalence versus those with low prevalence are of main interest. Within-cluster associations (educators in the same school) are accounted for to correct the standard errors but they are not of main interest. On the other hand, in the cluster-specific model, random or cluster specific effects are also included in addition to the predictor variables. The components of the response vector are assumed to be independent conditional on these random effects. The GEE and the design-based survey logistic approaches are population-averaged (or marginal) modeling techniques and they provide comparable estimates of robust standard errors for the logistic regression coefficients of the covariates. They however, do not separately estimate the variances of the random effects or their contributions to the total sampling variability and this is the key distinction between these alternative approaches and the mixed-model based logistic regression for hierarchical data (cluster-specific model/GLMMs). What

method is used to fit the model should not only depend on the assumptions the investigator is willing to make, but also (to some extent) on the availability of computational algorithms (Aerts et al, 2002). For instance the cluster-specific model involves integration of the likelihood function over the random effects distribution which, in general, does not have a closed form solution. Parameter estimates are therefore obtained through approximations or numerical integration techniques such as Gaussian quadrature.

When a nonlinear model is estimated, the GEE estimates are different from the ML estimates. For example, in an intercept-only logistic regression model the average probability of being HIV positive can be calculated from the population-average estimate of the intercept. The cluster-specific intercept can in general not be used to calculate this probability. In other words, the parameters β in the GLMM have no marginal interpretation. They however show a strong relation to their marginal counterparts for this particular model fitted here as demonstrated by Molenberghs, 2012; the marginal mean $\mu_i = E(y_{ij})$ satisfies $h(\mu_i) \approx X_i\beta^*$ with $\beta^* = [c^2 \text{Var}(b_i) + 1]^{-1/2}\beta$, in which c equals $16\sqrt{3}/15\pi$. This relation is the reason why the parameter estimates obtained from the cluster specific model (Table 4.5) are larger than those in the GEE model (Table 4.4). In scientific research where the variation in behavior of individuals within groups is studied, and the focus is on explaining how individual and group-level variables affect the behavior, unit-specific models are appropriate (Hox, 2010). However, PA models are applied where the research problem concerns interest in an entire population when one of the group-level variables is manipulated.

The significant risk factors were identified as; Race, Age, Marital status, Location and Living Arrangement and among the mobility factors, only *Away* was retained. The 2-level GLMM model showed a significant effect of the mobility variable *Away*; adjusted for other factors. It was clear that the educators who had been away from home for more than a month were at a higher risk of being HIV positive compared to their counterparts who had been away for less than a month in the last 12 months. Increased vulnerability of mobile populations is due to factors such as; the obligation to travel regularly and live away from spouses, separation from socio-cultural norms that regulate behaviour in stable communities, easy access to commercial sex workers and a sense of anonymity which allows for more sexual freedom (Shisana et al, 2005). It was found that educators in the rural areas were more like to be HIV positive. This is usually the opposite in the general population but here it may be because the educators posted in rural areas attract higher incomes

giving them the financial freedom that facilitates increased risky behaviour. HIV prevalence was highest among Africans compared to other races. This was also observed from several other studies (Shisana *et al*., 2005; Zuma *et al*., 2010; Mzolo, 2009). The results also showed a higher HIV prevalence among educators between the ages of 24-34 years. This may be explained by the high risky sexual behaviour usually prevalence among individuals in this age group and the fact that majority enter serious sexual relationships at this age.

5.2 Conclusion and Recommendation

Most of the methods that were applied in this Thesis did not find a statistically significant effect of migration/ mobility on HIV status. This comes as a surprise since most of the studies that have been conducted in this area as mentioned before did conclude that mobility is a significant risk determinant of HIV. The reason for the contradiction here is because some of the predictors are related. The effect of mobility could be masked in the other predictors for example Age or marital status. It is possible that the educators in the younger age groups or the unmarried ones are more likely to move after completing their teacher training. It is therefore recommended that the variable selection be done using Data mining techniques like the lasso, ridge and elastic-net (Hastie et al, 2008) which deal with the correct model variable selection for such correlated predictors.

It is also recommended that the Department of Education in South Africa should make an effort to put in place a systematic deployment structure where educators will get posted near their homes in order to reduce on their mobility and time spent away from their families. Efforts should also be concentrated in the districts that have a high prevalence and encourage the translation of knowledge into behavior change.

REFERENCES

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L.M. (2002) *Topics in Modelling of Clustered Data*. London: Chapman & Hall/CRC.

Aerts, M., Declerck, L., and Molenberghs, G. (1997) Likelihood misspecification and safe dose determination for clustered binary data. *Environmetrics*, **8**, 613–627.

Agresti, A. (2002) *Categorical Data Analysis*(2nd Edition). New York: John Wiley & Sons.

Appleton, S., Morgan, W.J., and Sives, A. (2006a) Should teachers stay at home? .The impact of international teacher mobility. *Journal of International Development*, **18**, 771-786.

Appleton, S., Sives, A., and Morgan, W.J. (2006b) The impact of international teacher migration on schooling in developing countries—the case of Southern Africa. *Globalization, Societies and Education*, **4**(1), 121-142.

Arnold, B. and Strauss, D. (1991) Pseudo-likelihood estimation: some examples. *Sankhya: the Indian Journal of Statistics, Series B* **53**, 233–243.

Binder, D.A. (1981) On the variances of asymptotically normal estimators from complex surveys, *Survey Methodology*, **7**, 157–170.

Breslow, N. and Clayton, D. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 813–820.

Caldwell, J., Caldwell, P., and Quiggin, P. (1997) 'Mobility, Migration, Sex, STDs, and AIDS: An Essay on Sub-Saharan Africa with Other Parallels' In G. Herdt, ed., *Sexual Cultures and Migration in the Era of AIDS* Oxford: Claredon Press.

Campbell, C. (2000) Selling Sex in the Time of AIDS: The Psycho-Social Context of Condom-use by Southern African Sex Workers. *Social Science and Medicine* **50**: 479-494.

Carriere, I. and Bouyer, J. (2002) Choosing marginal or random-effects models for longitudinal binary responses: application to self-reported disability among older persons. *BMC Medical Research Methodology* **2**,15.

Cochran, W.G. (1977) *Sampling Techniques,* 3d ed., John Wiley & Sons, New York.

Dai, J., Li, Z., and Rocke, D. (2006) *Hierarchical Logistic Regression Modelling with SAS GLIMMIX*. University of California, Davis, CA

Decosas, J., Kane, F., Anarfi, J., Sodji., K and Wagner, W.(1995) 'Migration and AIDS'. *The Lancet* **346**: 826-8.

Diggle, P.J., Liang, K.-Y., and Zeger, S.L. (1994) *Analysis of Longitudinal Data*. Oxford Science Publications, Oxford: Clarendon Press.

Geys, H., Molenberghs, G., and Ryan, L.M. (1997) Pseudo-likelihood inference for clustered binary data. *Communications in Statistics: Theory and Methods*, **26**, 2743–2767.

Geys, H., Molenberghs, G., and Ryan, L.M. (2002) Generalized Estimating Equations. In: *Topics in Modelling of Clustered Data*, Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (eds.), London: Chapman &Hall.

Giedler-Brown, B. (1998) 'Eastern and Southern Africa' in *Special Issue on Migration and HIV/AIDS of International Migration* **36:** 513-51.

Goldstein, H. (2010) *Multilevel Statistical Models* (4th Edition). New York: John Wiley & Sons.

Hastie, T., Tibshirani, R., and Friedman, J. (2008) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition. Springer-Verlag, New York.

Hox, J.P. (2010) *Multilevel Analysis: Techniques and Applications* (2nd Edition). Routledge: New York

Hunt, C. (1989) 'Migrant Labour and Sexually Transmitted Diseases: AIDS in Africa' *Journal of Health and Social Behaviour* **30**: 353-73.

Hubbard, A. E., Ahern, J., Fleischer, N.L., Van der Laan, M., Lippman, S.A., Jewell, N., Bruckner, T., and Satarianob, W.A.(20101) To GEE or Not to GEE: Comparing Population Average and Mixed Models for Estimating the Associations Between Neighborhood Risk Factors and Health. *Epidemiology* **21**: 467– 474

Jochelson, K., Mothibeli, M., and Leger, J.P. (1991) Human immunodeficiency virus and migrant labour in South Africa. *International Journal of Health Services* **21**: 157-173.

Kevin, D., Justine, O., and Deborah, J. (Dec 2010) Linking Migration, Mobility and HIV. *Tropical Medicine and international Health* **Vol.15** No. 12 pp 1458-1468

Kishamawe, C., Vissers, D., Urassa, M., Isingo, R., Mwaluko, G., Borsboom, G., Voeten, H., Zaba, B., Habbema, D. and Vlas, S. (2006) Mobility and HIV Tanzanian couples: both mobile persons and their partners show increased risk. *AIDS* **20**, 601-608.

Kok, P., Gelderblom, D., Oucho, J.O. and Van Zyl, J. (eds). (2006) *Migration in South and southern Africa: dynamics and determinants.* Cape Town: HSRC Press.

Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data* (2nd Edition). New York: John Wiley & Sons.

Lurie, M., Williams, B., Zuma, K., Mkaya-Mwamburi, D., Garnette, G., Sweat, M., Gittelson, J., and Abdool-Karim, S. (2003) Who infects whom? HIV-1 concordance and discordance among migrant and non-migrant couples in South Africa. *AIDS*, **17**:2245–2252

Lydie, N., Robinson, N.J., Ferry, B., Akam, E., De Loenzien, M., and Abega, S. (2004) Mobility, sexual behavior, and HIV infection in an urban population in Cameroon. *Journal of Acquired Immune Deficiency Syndrome*; **35**:67-74.

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

Miglioretti, D.L. Heagerty, P.J. (2004) Marginal modeling of multilevel binary data with time-varying covariates *Biostatistics*; **5**, 3, pp. 381–398

Molenberghs, G.(2012) *Survey Methods and Sampling Techniques Couse Notes* Master of Statistics Universiteit Hasselt.

Morris, M., Wawer, M.J., Makumbi, F., Zavisca, J.R., and Sewankambo, N. (2000) Condom acceptance is higher among travelers in Uganda. *AIDS*; **14**:733-741.

Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized linear models. *Journal of the Royal Statistical Society, Series B*, **135**, 370-384.

Neuhaus, J.M. (1992) Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research*, **1**, 249-273.

Neuhaus, J. M., Kalbfleisch, J. D. and Hauck, W. W. (1991) A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* **59**: 25–35

Nunn, A., Wagner, H., Kamali, A., Kengeya-Kayondo, J., and Mulder, D. (1995) Migration and HIV-1 sero-prevalence in a rural Ugandan population. Medical Research Council (UK) Programme on AIDS, Uganda Virus Research Institute, *AIDS*.**9**(5):503-6.

Peberdy, S. (2000) Mobile Entrepreneurship: Informal Sector Cross-border Trade and Street Trade in South Africa. *Development Southern Africa*, **17**(2): 201-219.

Pison, G., Le Guenno, B., Lagarde, E., Enel C., and Seck, C.(1993) Seasonal migration: a risk factor for HIV infection in rural Senegal. *Journal of Acquired Immune Deficiency Syndrome*; **6**:196-200.

Ramjee, G., and Gouws, E. (2002) Prevalence of HIV among Truck Drivers Visiting Sex Workers in KwaZulu-Natal, South Africa. *Sexually Transmitted Diseases*, **29**(1): 44-49.

Raudenbush, S.W. and Bryk, A.S., (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd Edition). Sage, Newbury Park, CA.

Rayfield, M., Downing, R., Baggs, J., and Hu, D. (1998) A Molecular Epidemiologic Survey of HIV in Uganda. *AIDS* **12**: 521-527

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

Sagguti, N., Nair, S., Malviya, A., Decker, M., Silverman, J., and Raj, A. (2012) Male Migration/ Mobility and HIV among married couples: Cross-sectional analysis of Nationally Representative data from India. *Pub Med.* **16**(6):1649-58

SAS Institute (2005) *SAS/STAT 9.1 Production GLIMMIX Procedure for Windows*. SAS Institute Inc., Cary, NC, USA.

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

Shisana, O., Peltzer, K., Zungu-Dirwayi, N., and Louw, J. (2005) The Health of Our Educators: A focus on HIV/AIDS in South African public schools. Cape Town: HSRC Press.

Snijders, T.A.B., and Bosker, R.J. (2012) Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. 2nd Ed. *SAGE Publications Limited.*

Soto-Ramirez, L., Renjifo, B., Mclane, M., and Essex. M. (1996) 'HIV-1 Langerhans Cell Trophism Associated with Heterosexual Transmission of HIV' *Science*. **271**:1291-3.

Sun, W., Shults, J., and Leonard, M. (2009) A note on the use of unbiased estimating equations to estimate correlation in analysis of longitudinal trials. *Biometrical Journal* **51**, 5–18.

Van Harmelen, J., Wood., R., Lambrick, M., Rybicki., E., Williamson, A., and Williamson, C. (1997) An association Between HIV-1 Subtypes in Different Risk Groups in South Africa and Mode of Transmission in Cape Town South Africa. *AIDS* **11**:81-87.

Verbeke, G. and Molenberghs, G. (2003) "The Use of Score Tests for Inference on Variance Components," *Biometrics*, **59**, 254–262.

Wang, Y. G. and Carey, V. J. (2004) Unbiased estimating equations from working correlation models for irregularly timed repeated measures. *Journal of the American Statistical Association* **99**, 845–852.

Wedderburn, R.W.M. (1974). "Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method". *Biometrika* **61** (3): 439–447

Williams, B., and Campbell, C. (1996) Mines, Migrancy and HIV in South Africa-managing the epidemic. *South African Medical Journal*. **86**: 1249-1251.

Wolfinger, R. and O'Connell, M.(1993) Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computing and Simulation* **48**, 233–243.

Wu, L. (2010) *Mixed Effects Models for Complex Data.* Chapman and Hall, New York.

Zhao, Y. and Joe, H. (2005) Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics,* **33**, 335–356.

APPENDIX

Table A.1: *Odds ratio(95% CI) for the design-based survey logistic model*

Effect		Point Estimate	95% Confidence Limits	
Away	**1 vs 2**	1.038	0.875	1.233
Race	**1 vs 2**	25.601	13.586	48.24
	3 vs 2	1.376	0.579	3.272
	4 vs 2	2.076	0.773	5.575
Age	**1 vs 5**	2.337	1.025	5.327
	2 vs 5	5.498	3.523	8.580
	3 vs 5	3.611	2.332	5.591
	4 vs 5	1.825	1.175	2.832
Marital status	**2 vs 1**	1.781	1.555	2.039
Location	**2 vs 1**	1.287	0.991	1.671
	3 vs 1	1.477	1.269	1.721
Living arr	**1 vs 3**	1.464	1.170	1.832
	2 vs 3	1.529	1.318	1.773
	4 vs 3	1.674	0.991	2.828

Table A.2: *GEE Parameter Estimates, robust standard errors and P-values*

Parameter		Estimate	SE	P-value
Intercept		-7.0867	0.3933	<.0001
Race	1	3.3476	0.3280	<.0001
	3	0.4360	0.4476	0.3301
	4	0.8164	0.5154	0.1132
Age	1	0.8529	0.4313	0.0480
	2	1.7034	0.2311	<.0001
	3	1.2869	0.2261	<.0001
	4	0.6073	0.2277	0.0077
Marital status	1	0.5665	0.0696	<.0001
Living arr	1	0.3709	0.1132	0.0011
	2	0.4126	0.0760	<.0001
	4	0.5254	0.2643	0.0468
location	2	0.2594	0.1332	0.0515
	3	0.3956	0.0801	<.0001
Away	1	0.0393	0.0884	0.6569

Table A.3: *2-level GLMM estimates, standard errors and P-values*

Effect		Estimate	SE	P-value
Intercept		-9.536	0.214	<.0001
Race	1	3.957	0.186	<.0001
	2	0.439	0.219	0.0444
	3	0.618	0.238	0.0093
Age	1	1.141	0.106	<.0001
	2	1.978	0.060	<.0001
	3	1.528	0.059	<.0001
	4	0.751	0.061	<.0001
Marital stat	1	0.565	0.017	<.0001
Location	2	0.172	0.098	0.0807
	3	0.967	0.102	<.0001
Living arr	1	0.354	0.028	<.0001
	2	0.348	0.019	<.0001
	4	0.668	0.074	<.0001
Away	1	0.063	0.021	0.0030

Table A.4: *3-level GLMM estimates, standard errors and P-values*

Effect		Estimate	SE	P-value
Intercept		-9.446	0.786	<.0001
Race	1	3.843	0.650	<.0001
	2	0.535	0.735	0.4665
	3	0.487	1.171	0.6775
Age	1	1.143	0.744	0.1244
	2	1.980	0.386	<.0001
	3	1.529	0.335	<.0001
	4	0.754	0.327	0.0213
Marital stat	1	0.564	0.095	<.0001
Location	2	0.153	0.258	0.5534
	3	0.940	0.357	0.0085
Living arr	1	0.351	0.114	0.0021
	2	0.345	0.106	0.0012
	4	0.667	0.306	0.0291
Away	1	0.063	0.120	0.5983

Table A.5: *MI-GEE with Empirical Standard Error Estimates and P-values*

Parameter		Estimate	Standard Error	P-value
Intercept		-6.4521	0.3555	<.0001
Race	1	2.8538	0.2872	<.0001
	3	0.1031	0.3799	0.7860
	4	0.0655	0.5277	0.9012
Age	2	1.5460	0.2049	<.0001
	3	1.2009	0.2021	<.0001
	4	0.5680	0.2321	0.0144
	1	0.4583	0.3823	0.2305
Marital stat	2	0.6089	0.0604	<.0001
Living arr	2	0.3377	0.0962	0.0005
	1	0.2410	0.1585	0.1283
	4	0.2784	0.2363	0.2388
Location	3	0.4977	0.0665	<.0001
	2	0.8212	0.4150	0.0478
Away	1	0.0465	0.0797	0.5598

Table A.6: Chi-square Test for association between factors

	Race	Age	Marital	Location	Living arr	Post near	With fam	Away	HIV status
Race	**1**								
Age	<.0001	**1**							
Marital	<.0001	<.0001	**1**						
Location	<.0001	<.0001	<.0001	**1**					
Living arr	<.0001	<.0001	<.0001	<.0001	**1**				
Post near	<.0001	<.0001	<.0001	0.0002	<.0001	**1**			
With fam	<.0001	<.0001	<.0001	<.0001	<.0001	0.0436	**1**		
Away	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	**1**	
HIV status	<.0001	<.0001	<.0001	<.0001	<.0001	0.0012	<.0001	<.0001	**1**

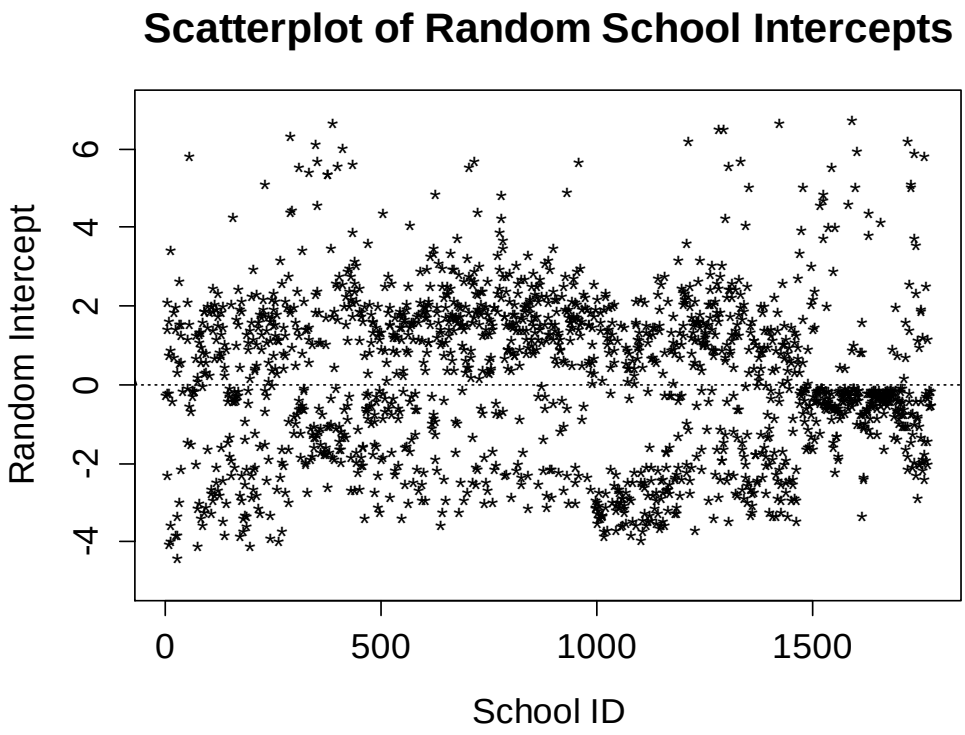


Figure A.1: *Empirical Bayes Estimates for the 2-level model*

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Investigating Migration as an important risk determinant of HIV infection among public school educators in South Africa

Richting: **Master of Statistics-Biostatistics**
Jaar: **2013**

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,

Agiraembabazi, Geraldine

Datum: **11/09/2013**