Modelling the non-linear force of infection for HIV in Malawi in 2015

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

By

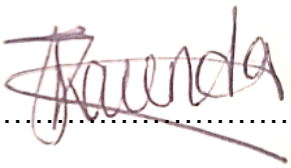
Justina Kaunda

A Research Report Submitted to the Faculty of Health Sciences, University of the Witwatersrand in partial fulfilment of the requirements for the Degree of Masters in Epidemiology in the field Biostatistics

August, 2019

Declaration

I, Justina Kaunda, a student of the Department of Epidemiology and Biostatistics, School of Public Health hereby declare that the project titled "**Modelling the non-linear force of infection of HIV in Malawi in 2015**" is my original work and has not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This research report is a partial fulfilment of Master of Science in Biostatistics Degree done at University of Witwatersrand under the guidance and supervision of Professor Eustasius Musenge, an Associate Professor in the Division of Epidemiology and Biostatistics.



……

Justina Kaunda

23rd August, 2019

Dedication

I would like to dedicate this work to the almighty God for His entire grace and mercies upon my life. He has been my source of wisdom and courage. May His Name be praised!

Abstract

Background

HIV/AIDS is one of the leading causes of deaths in Sub-Saharan Africa among individuals in the reproductive ages. Malawi is one of the countries in this hardest-hit region. This study aimed to model the age-specific Force of infection of HIV infection for Malawi in 2015.

Methods

The report is a secondary analysis of cross-sectional study data from the 2015 Malawi Demographic and Health Survey. The outcome variable was HIV status which was a binary variable coded, positive and negative. Statistical analysis was done using survey logistic and multilevel logistic models, survival models. Modelling the age-specific force of infection was done using Farrington models, log-logistic model, non-linear model using maximum likelihood estimation and Susceptible Infected model.

Results

There were 14,010 male and female individuals aged 15-49 years in our analysis, where 9.1% were HIV positive. Place of residence, age in years, gender, marital status and total number sexual partners showed a significant association with HIV infection. Individuals living in urban areas were more likely to have HIV infection than those living in the rural areas. The risk of having HIV infection was less likely among males than females. Furthermore, the risk of having HIV infection increased as the age in years increased. This was also observed with the total number of sexual partners. Individuals who reported to have not been married before were less likely to have the infection unlike those who reported living together, married, divorced and widowed. The Farrington's two parameter model was the best fit for the Force of Infection. The Force of infection for HIV decreased as the ages increased.

Conclusion

Modelling the age-specific forces of infections of HIV is very important as its estimates help in targeting the specific groups that need more interventions. Modelling also helps in predictive models such as those used on the 90:90:90 for HIV by year 2020. It also helps in evaluating disease control programs, comparing prevention and therapy programs and informing policies.

Keywords

Force of Infection (FOI), Modelling, Human Immuno-deficiency Virus (HIV), Malawi Demographic and Health Survey (MDHS)

Acknowledgements

I would like to express my sincere gratitude towards my husband, Justin for his guidance, patience, prayers and moral support towards my report. I would also like to extend my gratitude to my supervisor, Professor Eustasius Musenge for his entire guidance, criticism and academic support towards the research report.

I am also thankful for the prayers and support from my pastors and fellow church members and my family.

Financial Support

This research study was supported through the DELTAS Africa Initiative SSACAB (Grant No. 107754/Z/15/Z). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS) Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (Grant No.107754/Z/15/Z) and the UK government. The views expressed in this research report are those of the authors and not necessarily those of the AAS, NEPAD Agency, Wellcome Trust or the UK government.

Contents

List of Figures

List of Tables

Abbreviations

AAS	African Academy of Sciences
AESA	Alliance for Accelerating Excellence in Science in Africa
AIC	Akaike Information Criterion
AIDS	Acquired Immune Deficiency Syndrome
BIC	Bayesian Information Criterion
DIC	Deviance Information Criterion
ELISA	Enzyme-Linked Immunosorbent Assay
FOI	Force of Infection
HIV	Human Immuno-deficiency Virus
MDHS	Malawi Demographic and Health Survey
NAC	National Aids Commission
NEPAD Agency	New Partnership for Africa's Development Planning and Coordinating Agency
NSO	National Statistical Office
UN	United Nation
UNFPA	United Nations Population Fund
UNICEF	United Nations Children's Fund
USAID	United States Agency for International Development
WHO	World Health Organisation
UNAIDS	Joint United Nations Programme on HIV/AIDS

Chapter 1 Introduction

1.1 Background

The Human Immuno-deficiency Virus (HIV) is an infectious virus that causes Acquired Immuno-deficiency Syndrome (AIDS), a weakened immune condition now most commonly referred to as HIV/AIDS. HIV/AIDS is one of the leading causes of death worldwide among women and men in the reproductive age groups in middle and low-income countries (1).

Approximately, about 35 million people globally have died from AIDS-related illness since the start of the epidemic in the early 1980s to 2015 (1). Sub-Saharan Africa is the region with the highest number of people living with HIV worldwide with an estimated 36.7 million individuals who lived with HIV in 2015 (1). About 19 million people from the 36.7 million living with HIV are from Eastern and Southern Africa and these two regions bare the greatest burden of the disease (1).

Malawi is one of the countries in this highly HIV-burdened region. Reports from previous Malawi Demographic Health Survey (MDHS) and surveillance surveys in Malawi showed a steady decline in HIV prevalence since 1991 to 2010 (from 16.4% to 10.6%) (2). The 2015 MDHS found that HIV prevalence in the country was 8.8% (3). It also showed that prevalence was higher among females aged 15-49 years (10.8%) than males in the same age group (6.4%) which also varied with socio-demographic characteristics.

Furthermore, identifying the most affected groups provides information to policy makers to spread awareness of the impacts of HIV and allocate other interventions to these groups (4). This in turn helps in controlling the HIV burden.

Parametric, non-parametric, semi-parametric and Bayesian-based models can be used to model force of infection (FOI) for various infectious diseases. Bayesian-based models are critical as they can estimate the FOI directly from data, which can in turn give a real representation of a population (5). Estimating the FOI per age group is one way which can be modelled using the Bayesian models

1.2 Literature review

1.2.1 HIV-related Force of Infection in Malawi

Malawi is one of the hardest hit countries by the HIV/AIDS epidemic, but information on the Force of Infection (FOI) is insufficient (6). The force of infection which is the measure of risk of becoming infected is defined as the rate at which susceptible individuals become infected with a particular disease. It is closely related to incidence, which is hard to measure directly (5, 7). The force of infection is also known as the hazard function (5). The FOI consists of the number of contacts between susceptible and infected individuals, the probability of transmission and the probability of a partner being infectious. The FOI estimates the risk of contracting a disease, unlike prevalence which can lead to inaccurate conclusions as it indicates how widespread the disease is. Force of infection models are important in public health as they can help to show if a disease is pandemic across different ages, since

they are often reported as age-specific rates. Models also help in making or informing policies, comparing prevention and therapy programs and evaluating disease control programs (8, 9).

1.2.2 Approaches to modelling FOI

Ross (10) established a modeling framework of infectious disease transmission in the early 90's. He formulated a mathematical model for FOI known as a priori model. The priori model assumes knowledge of the cause, constructing differential equations on the assumptions, testing calculated results and comparing with the observed statistics. However, the main problem with priori models is that some of the parameters in the differential equations are unknown (5). Ross (10) proposed another statistical model known as a posteriori model to solve the problem encountered with the priori model. The posteriori model starts with the observed statistics (data), tries to fit models onto the data and check the underlying cause. Comparing the results of the two models is very important as the two models are complementary and this can further improve model building within each model (10). However, the posteriori model is very important as it uses methods derived from the data to estimate FOI (5). The priori and posteriori are shown in figure 1.1 below.

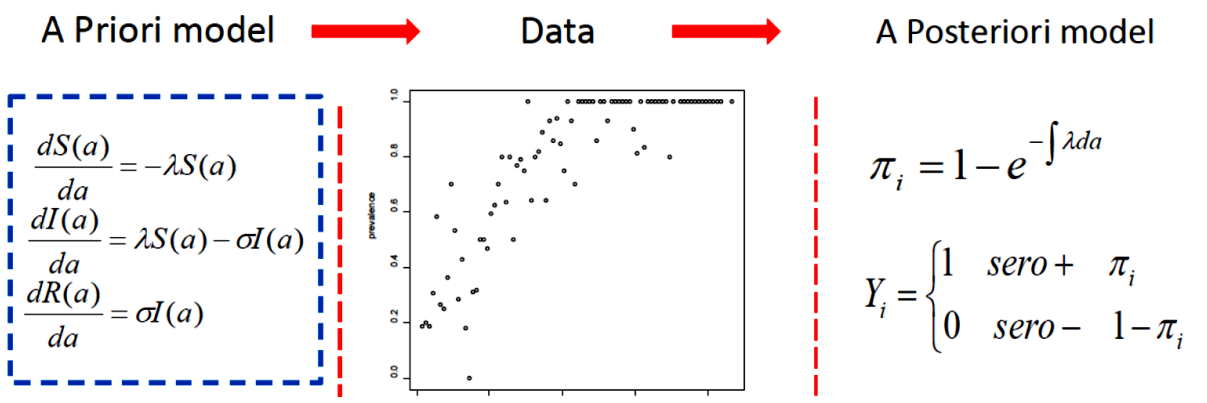


Figure 1.1: Priori and Posteriori models

A study done by Misiri, Edriss (6) estimated the HIV force of infection in Malawi using recurrence relation. However, modelling force of infection using parametric models only could lead to inaccurate estimations since parametric models can estimate negative FOI if estimated probability is an increasing function (5). This can be controlled for by carefully adjusting the range of data observations. Bayesian approaches used to model the FOI are important since they can estimate secondary peaks for FOI, unlike parametric models which estimate a single peak model for FOI. Bayesian models can also estimate a non-decreasing FOI (5). This study, therefore, sought to estimate the HIV force of infection in Malawi in order to track and monitor the HIV burden by using three modelling approaches namely, parametric, semi-parametric and Bayesian-based approaches (11, 12).

A study done in Brazil used a catalytic model to estimate force of infection (11). Three models under the catalytic approach were fitted. The first model assumed that the force of infection was the same for all the ages, and the other two models adjusted for different covariates. Another study which was done by Sutton et.al, (2008) used maximum likelihood methods to estimate the force of infection of blood-borne viruses in injecting drug users. The prevalence of the function was assumed to depend on the FOI as a function of time and injecting career length. The FOI contained a parameter which was assumed to follow a gamma distribution. They found that force of infection for the viruses varied across European countries (7).

1.2.3 Estimating FOI adjusting for socio-demographic and biological characteristics

Force of infection estimates can vary as a consequence of different behavioural patterns, biological variance (e.g viral load) and other transmission processes (e.g., injection-related) since most models focus specifically on sexually-transmitted HIV (13). Johnson (2004) suggested that allowing for heterogeneity in sexual behaviours, condom use and viral load levels in the HIV transmission model is very important as this helps to estimate the force of infection accurately. When modelling a heterosexual epidemic, dividing the population by gender is very important. Women tend to have a higher-risk of being infected than men and HIV prevalence is higher among women than males (14).

Furthermore, considering the effect of age is very important when modelling the force of infection since different age groups have different transmission rates and different behavioural and sexual patterns. Various studies have shown an association between age and HIV infection. Some studies found that the risk of having HIV in Malawi increased as age increased (6, 15). Most of the studies described above used polynomial models to estimate the force of infection.

Authors agree that reducing the force of infection for HIV is critical for the development of better approaches and strategies for controlling HIV infection (6, 8, 15) . Thus in order to effectively reduce the force of infection for HIV, we have to first of all estimate what the force of infection for the disease is in a given setting. This emphasises the need for continuous monitoring of the infection among populations at risk, continual surveillance of the risk of infection as well as the force of infection in order to better protect people from the disease.

1.3 Problem statement

Most mathematical models use estimated force of infection not necessarily derived from data. This estimated force of infection can sometimes lead to inaccurate conclusions of the HIV epidemic in a country. Posteriori models are very useful as they use methods derived from data which can provide better estimates of the FOI. Furthermore, little has been done to estimate the measure of risk of being infected with HIV among women and men aged 15-49 years in Malawi using data and the comparison of different approaches to estimate the force of infection.

1.4 Justification

Estimating the force of infection for HIV can be useful in controlling the spread of the HIV epidemic in real strategies and can also help in detecting the infection and planning national strategic responses. In this study, estimating the FOI using real data will provide a better estimate of the FOI than a common approach where mathematical models estimate the FOI without using data.

Comparison of different approaches to estimate the force of infection is important because it helps to choose or check which approach is the best to estimate the force of infection. Also, adjusting for other covariates and risk groups helps to determine the adjusted forces of infection controlling for those covariates. This can help in targeting a specific group with higher force of infection with HIV control strategies and monitoring interventions in these groups.

1.5 Research question, aim and objectives

The study sought to answer the following question; what were the age group specific forces of infection for HIV infection in Malawi in 2015?

The main aim of this study was therefore to determine the age group specific forces of infection for HIV infection in Malawi using 2015 Demographic Health survey data.

Specific objectives of the study were as follows:

1. Describe the socio-demographic characteristics of the population stratified and compared by HIV status for the year 2015
2. Estimate the age group specific forces of infection for the given population using non-linear approaches
3. Model the force of infection for HIV using Susceptible Infected (SI) model

Chapter 2 Methodology

2.1 Introduction

This section covers the methodology and data analysis for this research project. It also explains the statistical methods used to model the non-linear force of infection.

2.2 Description of Original Study

2.2.1 Survey design and study description area

Malawi is a landlocked country in the sub-Saharan Africa region as shown in figure 2.1. It has a surface area of 118,484 square kilometers and shares its boundaries with Tanzania to the North and North-East, Zambia to the North-West and Mozambique to the East, South and South-Western part. Malawi is divided into three regions, Northern, Central and Southern which are further divided into 28 districts (16). The districts are subdivided into Traditional Authorities (TA) which are composed of villages. The 2015 Malawi Demographic and Health Survey (MDHS) was a cross-sectional, nationally represented survey. The sample in the survey was stratified and selected in two stages and districts were stratified to urban and rural areas. The first stage of sampling involved the sampling of enumeration areas (clusters) and the second stage involved the sampling of households within enumeration areas(3).

Figure 2.1: Map of Malawi ((17))

2.2.2 HIV testing procedures

The 2015 MDHS was the third national survey to include HIV testing. Blood samples were collected from individuals (men and women) who consented to HIV testing. All samples collected were tested using three test assays (Enzyme-linked immunoassay (ELISA I), ELISA II and Murex HIV Ag/Ab combination (DiaSorin)) (3). The three test assays are described below:

- All samples that tested positive on ELISA I were subjected to ELISA II and Diasorin, 5% of those that tested negative were also subjected to ELISA II while the rest were recorded negative. Results that showed negative on both ELISA I and II were recorded negative and if results were discordant, the two assays were repeated in parallel (3).
- Results that remained discordant were subjected to a third assay, the InnoLia HIV I/II Score (Fujirebio) line immunoassay. Positive results on both the ELISA I and II were also subjected to the third assay, which was rendered positive if InnoLia was positive and inconclusive if InnoLia was negative or indeterminate. Discordant results on both ELISA I and II were rendered inconclusive if InnoLia was positive, negative if Innolia was negative and indeterminate if InnoLia was indeterminate.

The HIV results were linked to the individual interview data using unique identifiers (3).

2.2.3 Data collection and Objective of Primary study

Questionnaires were used to collect individual information from men aged 15-54 and women aged 15-49 years. A total of 27,516 households were selected, of which 26,361 were successfully interviewed from 850 clusters. Of the interviewed households, a total of 24,562 women and 7,478 men were successfully interviewed (3).

The primary objective of the primary study was to provide recent estimates of demographic and health indicators which would help policy makers and program managers in evaluating and implementing health strategies and programmes(3).

The survey was conducted by the National Statistical Office (NSO) and Ministry of Health of Malawi. It was funded by the United States Agency for International Development (USAID), National Aids Commission (NAC), United Nations Children's Fund (UNICEF), United Nations Population Fund (UNFPA) and UN Women (3).

2.3 Study design, population and inclusion criteria

This was a secondary data analysis of a cross-sectional, nationally represented survey of Malawi conducted in the year 2015. The study utilized data of men and women in Malawi who were tested for HIV in 2015. A total of 850 enumeration areas and from a total 27,516 households were selected for the 2015 MDHS survey and a third were tested for HIV (3). A total of 26,361 households were interviewed where,

24,562 women aged 15-49 years and 7,478 men 15-54 years were interviewed. A total of 15,125 individuals were tested for HIV and this was merged with the demographic dataset which resulted to 14,779 individuals. Individuals who had indeterminate or inconclusive HIV results were excluded from the sample which resulted into 14,010 individuals whose data were used in the analysis as shown in figure 2.2 below.

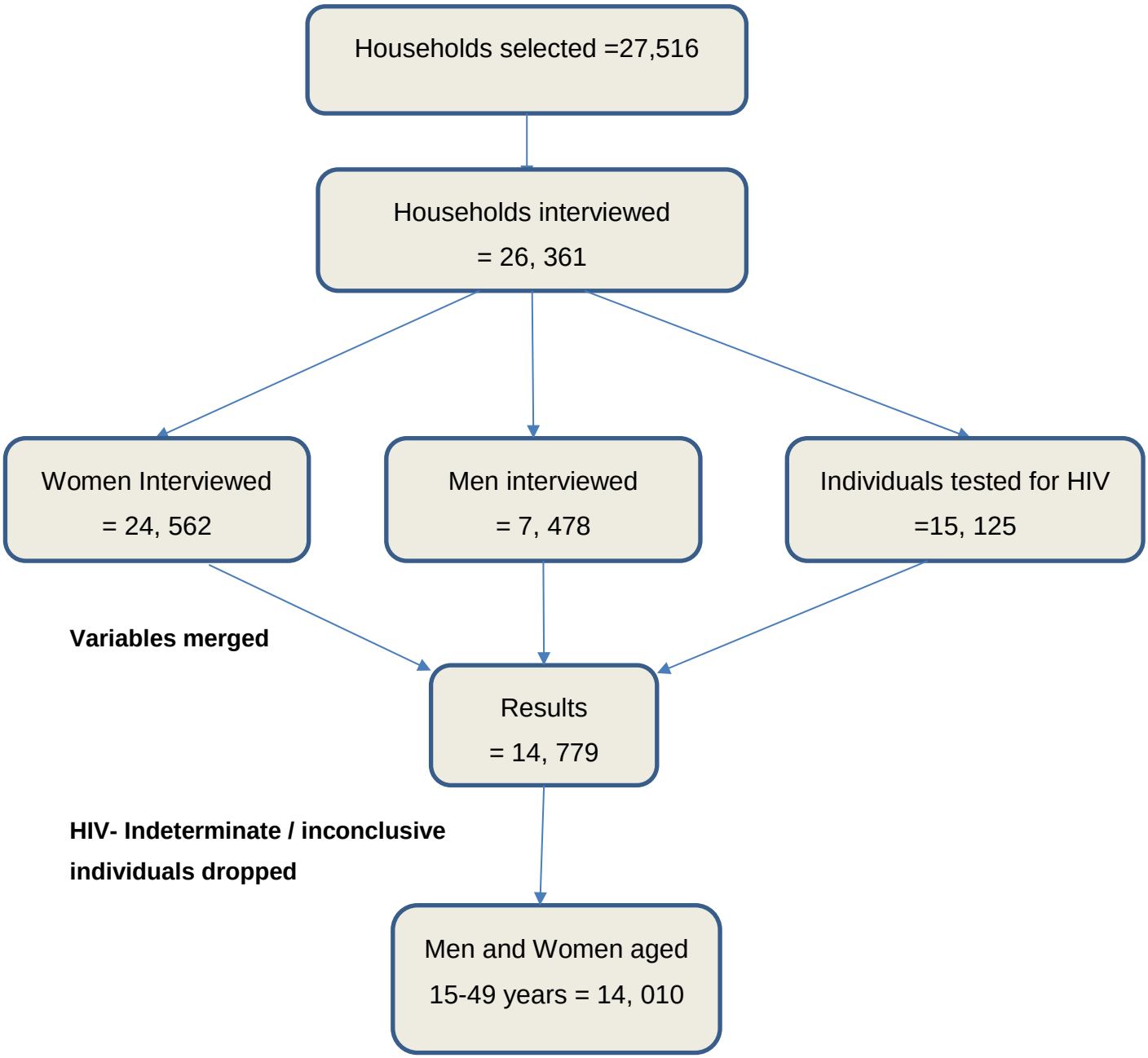


Figure 2.2: Flow chart of participants who were used in this study

2.4 Power computation

The study utilized all individuals who were tested for HIV. Power was calculated using Stata clustersampsi command, based on a Stata Ado file (18). At an alpha level of 0.05, a prevalence of 8.45% was assumed for the population and a prevalence of 10.2% was observed for the sample. An inter-cluster correlation (ICC) of 0.15 and a design effect of 3.70 with 847 clusters and 19 households per cluster, we obtained an 80% power as shown in the appendix A.1.

2.5 Data management and processing

The data were cleaned and checked for discrepancies. The study only utilised variables needed to answer the research question and the other variables were dropped. Variables were renamed for easier identification. Exploratory data analyses were used to summarise the characteristics of the data and check for outliers. The analyses were done in STATA versions 14 and 15, WINBUGS and R software packages (19-22).

2.6 Data Analysis

2.6.1 Explanatory Variables

The explanatory variables were grouped into two, categorical and continuous variables. The categorical variables were:

- **Region**: This variable categorised the country into three regions; northern, central and southern region.
- **Place of residence**: This variable was categorised into 2 levels; urban or rural.
- **Education level**: This variable explained the highest education one attained. It was categorised into 4 levels; no education, primary, secondary and higher **education.**
- **Wealth Index**: This variable was categorised into 5 levels; poorest, poorer, middle, richer and richest
- **Gender**: This was categorised into 2, whether male or female.
- **Marital status**: This variable was categorised into 4 levels; never married, married, living together and one was divorced or widowed.
- **Religion**: This was categorised into 3 levels; Catholic, Protestants or Muslim and any other

The continuous explanatory variables were

- **Age:** This variable described the age of the individual
- **Age at first sex**: This variable described the age at which one had sex for the first time.

2.6.2 Outcome variable

The outcome variable was HIV status which was coded 0 (negative) and 1 (positive) as defined in section 2.2.2. The force of infection was a function of the prevalence of

the HIV positive group. The equations showing the relationship are shown in section 2.6.6.1

2.6.3 Survey setting of the data

Survey setting was done in STATA 14.0 using the cluster number (which was the primary sampling unit), the sample weight and the strata. The strata variable was generated using district variable (with 28 districts) and place of residence (urban and rural).

2.6.4 Descriptive and Bivariate analysis

To describe socio-demographic characteristics of the residents, descriptive statistics were used. Survey-adjusted summary measures stratified by HIV status were used for both categorical and continuous variable. For the continuous variables, mean and standard errors were obtained. For the categorical variables, weighted percentages, graphs and bar charts were displayed. Spatial maps showing prevalence and force of infection per district were drawn using Spmap Ado files in Stata (23).

The Pearson's Chi-square test adjusting for survey design effect was used for the bivariate analysis. This adjusted Chi-square test was used to establish the relationship between the outcome variable and categorical explanatory variables.

2.6.5 Multivariable Analysis

2.6.5.1 Survey logistic and multilevel models

Survey adjusted logistic models and mixed-effects (multilevel) logistic models were fitted on our data. Mixed-effects logistic models are logistic models containing both fixed and random effects. Models with random effects are very useful in modelling the intra-cluster correlation which measures how observations are correlated. The outcome variable for these models was a binary outcome which was coded negative (0) and positive (1). A two-level binomial model was fitted with cluster as random effect. Odds ratios, 95 percent confidence intervals, and p-values were reported for each variable and all variables with $p<0.05$ were considered significant. Adjusted Wald tests and seemingly unrelated estimation (suest) procedures were used to test the significance of the parameters and whether the model was a good fit (24).

The survey logistic (one-level) model can be shown as follows (25);

$$\Pr(y_{ij}=1|x_{ij}) = H(x_{ij}\beta) \quad \ldots\ldots \quad \text{(eqn 1)}$$

Where the function H(.) is the logit form i.e,

$$\log it\left(p\right)=\log\left(\frac{p}{1-p}\right)=\eta$$

$$\frac{p}{1-p}=e^{\eta}$$

$$p=\frac{e^{\eta}}{1+e^{\eta}}$$

$$H(\eta)=\exp(\eta)/\left\{1+\exp(\eta)\right\}\ \ldots\ldots \qquad \text{(eqn 2)}$$

Where $\eta = x_{ij}\beta$

for j = 1,2,…, K clusters, with cluster j consisting of i = 1,…, n_j observations. Y_j are the binary responses which are 1 if HIV positive and 0 if HIV negative. The x_{ij} are the covariates and β are regression coefficients in the model. The *H(.)* is the logistic cumulative distribution which links the linear predictor ($\eta_{ij} = x_{ij}\beta$) to the probability of successes (y_{ij}=1) using equation 2 above and exp(β) are odds ratios.

The two-level logistic model (adjusting for random effects) fitted above for a series of K independent clusters can be shown in the following equations (26);

$$\Pr(y_{ij}=1\left|x_{ij},\boldsymbol{u}_j\right.)=H(x_{ij}\beta+z_{ij}\boldsymbol{u}_j) \qquad \text{(eqn 3)}$$

$\boldsymbol{u_j}$ are the random effects with z_{ij} as the covariates of the random effects.

The binomial distribution follows the probability density distribution as follows

$$\Pr\left(y_{ij}\right)=\binom{r_{ij}}{y_{ij}}\left\{H(x_{ij}\beta+z_{ij}\boldsymbol{u}_j)\right\}^{y_{ij}}\left\{1-H(x_{ij}\beta+z_{ij}\boldsymbol{u}_j)\right\}^{r_{ij}-y_{ij}} \quad \ldots \quad \text{(eqn 4)}$$

The conditional distribution of $\mathbf{y}_j$ for cluster j , with cluster-level random effects $\mathbf{u}_j$ and where y_{ij} are number of successes in the series of r_{ij} Bernoulli trials is,

$$f\left(y_j\left|\boldsymbol{u}_j\right.\right)=\prod_{i=1}^{n_j}\Pr(y_{ij}=1\left|x_{ij},\boldsymbol{u}_j\right.) \quad \ldots \quad \text{(eqn 5)}$$

$$f(y_j \mid \boldsymbol{u}_j) = \prod_{i=1}^{n_j} \left[\binom{r_{ij}}{y_{ij}} \left\{ H(\eta_{ij}) \right\}^{y_{ij}} \left\{ 1 - H(\eta_{ij}) \right\}^{r_{ij} - y_{ij}} \right]$$

$$H(\eta) = \exp(\eta) / \left\{ 1 + \exp(\eta) \right\}$$

$$f(y_j \mid \boldsymbol{u}_j) = \prod_{i=1}^{n_j} \left[\binom{r_{ij}}{y_{ij}} \left\{ \frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})} \right\}^{y_{ij}} \left\{ 1 - \frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})} \right\}^{r_{ij} - y_{ij}} \right]$$

$$= \prod_{i=1}^{n_j} \left[\binom{r_{ij}}{y_{ij}} \left\{ \frac{\exp(\eta_{ij} y_{ij})}{\left(1+\exp(\eta_{ij})\right)^{y_{ij}}} \right\} \left\{ \frac{1}{\left\{ 1+\exp(\eta_{ij}) \right\}} \right\}^{r_{ij} - y_{ij}} \right]$$

$$= \prod_{i=1}^{n_j} \left[\binom{r_{ij}}{y_{ij}} \exp(\eta_{ij} y_{ij}) \left(1+\exp(\eta_{ij})\right)^{-y_{ij}} \left\{ 1+\exp(\eta_{ij}) \right\}^{-(r_{ij} - y_{ij})} \right]$$

$$= \prod_{i=1}^{n_j} \left[\binom{r_{ij}}{y_{ij}} \exp(\eta_{ij} y_{ij}) \left(1+\exp(\eta_{ij})\right)^{-y_{ij}} \left(1+\exp(\eta_{ij})\right)^{-r_{ij}} \left(1+\exp(\eta_{ij})\right)^{y_{ij}} \right]$$

$$= \prod_{i=1}^{n_j} \left[\binom{r_{ij}}{y_{ij}} \exp(\eta_{ij} y_{ij}) \left(1+\exp(\eta_{ij})\right)^{-r_{ij}} \right] \ldots \qquad \text{(eqn 6)}$$

expressing equation 6 into exponential family form

$$= \exp\left\{ \log\left\{ \prod_{i=1}^{n_j} \left[\binom{r_{ij}}{y_{ij}} \exp(\eta_{ij} y_{ij}) \left(1+\exp(\eta_{ij})\right)^{-r_{ij}} \right] \right\} \right\}$$

$$= \exp\left(\sum_{i=1}^{n_j} \left[y_{ij} \eta_{ij} - r_{ij} \log\left\{ 1+\exp(\eta_{ij}) \right\} + \log \binom{r_{ij}}{y_{ij}} \right] \right) \quad \ldots \qquad \text{(eqn 7)}$$

Where $\eta_{ij} = x_{ij}\beta + z_{ij}\boldsymbol{u}_j + offset_{ij}$

The parameter estimates are obtained by maximising the log likelihood. The likelihood is obtained by integrating out the random effects $\mathbf{u}_j$ from equation 7. The random effects $\mathbf{u}_j$ are assumed to follow a multivariate normal distribution with mean **0** and variance $\boldsymbol{\Sigma}$. The likelihood function for cluster j is given as

$$L_j(\beta,\boldsymbol{\Sigma}) = \int_{-\infty}^{\infty} f(y_j \,|\, \boldsymbol{u}_j)\,\phi(\boldsymbol{u})d\boldsymbol{u}_j \qquad \text{(eqn 8)}$$

The log likelihood of the whole dataset is given as

$$\ell(\beta,\boldsymbol{\Sigma}) = \sum_{j=1}^{K} L_j(\beta,\boldsymbol{\Sigma}) \qquad \text{(eqn 9)}$$

The log likelihood adjusting for survey weights is given as

$$L(\beta,\boldsymbol{\Sigma}) = \sum_{j=1}^{K} w_j \log \int_{-\infty}^{\infty} \exp\left\{ \sum_{i=1}^{n_j} w_{i|j} \log f\left(y_{ij} \,|\, u_{ij}\right) \right\} \phi(\boldsymbol{u}_j)d\boldsymbol{u}_j \quad \ldots\ldots \qquad \text{(eqn 10)}$$

Where w_j is the inverse of probability of selection for cluster *j*, $w_{i|j}$ is the inverse of conditional probability of selection of individual *i* given selection of cluster *j.* The *f(.)* is as shown in equation 7 above and $\boldsymbol{\phi}(\mathbf{u}_j)$ is the multivariate normal density. $\boldsymbol{\phi}(\mathbf{u}_j)$ is given by $\phi(\boldsymbol{u}_j) = (2\pi)^{-\frac{q}{2}} \left|\Sigma\right|^{-\frac{1}{2}} \exp(-\frac{1}{2}\boldsymbol{u}_j'\Sigma^{-1}\boldsymbol{u}_j)$.

2.6.5.2 Survey Survival models

Survey Survival models were also fitted with the binary outcome (HIV status) and a time variable (Age). Hazard ratios, 95% confidence intervals and p-values were

reported. The Cox regression model fitted can be shown by the following equations (27),

$$h(t) = h_o(t)\exp(\beta_1 x_{11} + ... + \beta_p x_{pk}) \quad \text{(eqn 11)}$$

Where *h(t)* is the hazard

$H_o(t)$ is the baseline hazard estimate

β's are the parameter estimates

x_{ij} are the covariates adjusting for clusters

2.6.6 Non-linear Models

2.6.6.1 Non-linear mixed effects model

Non-linear mixed effect model was fitted using the Hazard rates obtained from the survival analysis. The age-specific hazard rates were used to fit the multilevel non-linear model and a specified likelihood for the hazard (force of infection) was maximized to estimate the parameters.

2.6.6.2 Farrington's model

To estimate the force of infection, Farrington's model was fitted to estimate the age-specific FOI under the frequentist approach. Farrington constrained the parameter space in order to have a non-negative FOI.

The number of HIV-infected individuals in age group a (y_a) was assumed to follow a Binomial distribution from a sequence of the individual level Bernoulli trials.

$$y_a = \begin{cases} 1 & HIV+, \text{ with prevalence } \pi(a) \\ 0 & HIV-, \quad 1-\pi(a) \end{cases} \quad \text{(eqn 12)}$$

$\pi(a)$ was known as the prevalence in age group a.

The age-specific prevalence $\pi(a)$ was given as the function of the age-specific FOI $\lambda(a)$ given by the following equation a

$$\lambda(a) = (\alpha a - \gamma)e^{-\beta a} + \gamma \quad \text{(eqn 13)}$$

Where $\lambda(a)$ - age-specific FOI

α - constant

β - Coefficient for age

a - age group

γ - long-term residual value, as $a \to \infty, \lambda(a) \to \gamma$.

Integrating $\lambda(a)$ using integration by parts results in a non-linear model

$$\pi(a) = 1 - e^{-\int_0^a \lambda(s)ds}$$

$$= 1 - \exp\left\{\frac{\alpha}{\beta} a e^{-\beta a} + \frac{1}{\beta}\left(\frac{\alpha}{\beta} - \gamma\right)\left(e^{-\beta a} - 1\right) - \gamma a\right\} \quad \text{(eqn 14)}$$

The full derivation of equation can be found in appendix A.2. This was solved in WinBUGS under hierarchical models.

2.6.6.3 Bayesian modelling of the force of infection

In the Bayesian approach, Nonlinear Hierarchical models were used to estimate the force of infection. In order to achieve the desired posteriori distribution, a priori distribution was assumed for the parameter space. The prior distribution was assumed to follow a truncated normal distribution. The number of HIV infected individuals in given age groups was assumed to follow independent binomial distributions.

There were three non-linear Farrington models and Log-logistic model which were fitted in Winbugs. The first non-linear Farrington model had two parameters, where the third parameter was assumed to be zero. The second Farrington model was a three-parameter with alpha3 not equal to zero and the third model was a three-parameter fitted on cumulative data. In order to estimate the FOI, Monte Carlo simulations approximated the FOI using the posterior means. The models were compiled with two Markov chains and initial values were loaded for each model. After

specifying and loading the model, a 1000 update burn-in was followed by 100,000 updates where 100th iteration was stored in the sample. This gave the parameter estimates in the model.

2.6.7 Estimating Force of Infection using Susceptible-Infected (SI) Models

Epimodel package was used to simulate and analyse a basic Susceptible-Infected (SI) model without demographics. Susceptible people were those who did not have the infection but were at risk whereas the infected were those who had the infection. Figure 2.3 shows the SI compartments with the FOI rate.

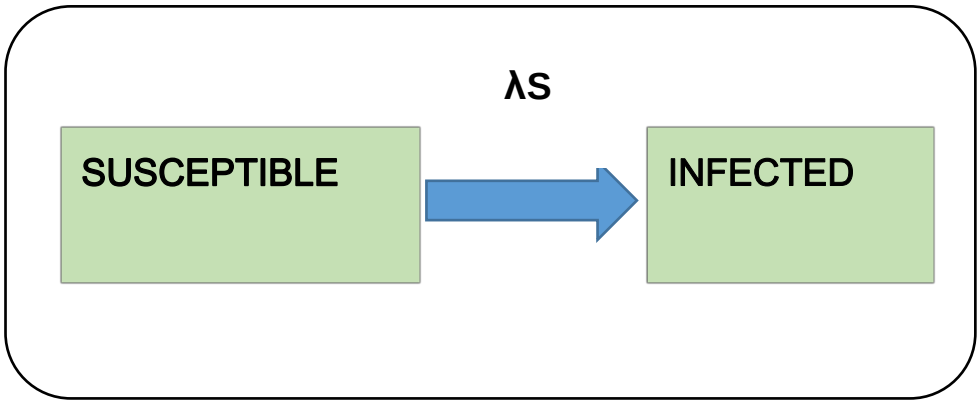


Figure 2.3: Susceptible-Infected HIV Model

The equations representing the size of the compartments over time were as follows;

$$\begin{cases} \frac{dS}{dt} = -\lambda S \\ \frac{dI}{dt} = \lambda S \end{cases} \qquad \text{eqn(15)}$$

where λ which is the force of infection is given by

$$\lambda = \frac{\beta c I}{N}$$

$$\therefore \begin{cases} \frac{dS}{dt} = -\frac{\beta c I}{N} S \\ \frac{dI}{dt} = \frac{\beta c I}{N} S \end{cases} \qquad \text{eqn(16)}$$

Where $\boldsymbol{\lambda}$ is the force of infection and is given by $\frac{\beta c I}{N}$. β is the probability of transmission per contact, c is the rate of contact per person per unit time, I is the number of infected and N is the size of the population (5). The number age group was used as a time factor and the force of infection was obtained by averaging the hazard function of the age groups. The probability of transmission was found to be around 0.30 to 0.38 (28), the average number of people infected and the average population size for the age groups were used. The contact rate was calculated using the estimates found above and with this formula $c = \frac{N\lambda}{\beta I}$. Furthermore, the contact rate was assumed to be constant as this was a basic SI model. Sensitivity analysis was also performed and SI plots were also reported.

2.6.8 Model goodness of fit

Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) were used for comparison and assessing the goodness of fit in R software. The Deviance Information Criterion (DIC) was used for comparison and assessing the goodness of fit in Winbugs. The models with lower AIC or BIC or DIC were considered to be better (5).

2.7 Ethics approval

The study protocol was presented to and approved by the Faculty of Health Sciences at the University of Witwatersrand. Ethical clearance was sought from the Human Research Ethics Committee of the University of Witwatersrand. It was granted before data analysis and the clearance certificate number is M170147. Permission to use Malawi Demographic Health Survey (2015) data from the ICF International was sought using an online application. All personal identifiers were removed from the data to protect the confidentiality of the participants. The data were only accessible to the study researchers.

Chapter 3 Results

This chapter presents the results for the study in tables and graphic format. Firstly, the first objective results are presented on section 3.1 as descriptive and bivariate statistics. The second objective results are presented in section 3.2 in form of linear and non-linear models, and also models using maximum likelihood estimation and Bayesian estimations. The third objective results are presented in section 3.3 and concluding remarks are presented in section 3.4

3.1 Descriptive and Bivariate Analysis

The survey data contained 14,010 individuals who were tested for HIV and had their demographic information. The results presented in table 3.1 showed that 12,736 (90.9%) were HIV negative and 1,274 (9.1%) were HIV positive individuals. The northern region had a prevalence of 7%, the Central had 5.9% and the Southern region had a prevalence of 13 % as shown in figure 3.1. The mean age at first sex was 8.6 years for those who were HIV positive and 11.3 years for those who were HIV negative. The mean total number of sexual partners was 18 partners for those who were HIV positive and 14 for those who were HIV negative. Female individuals had a prevalence of 11% whereas male individuals had a prevalence of 7 % as shown in figure 3.2. Furthermore, hotspots of HIV prevalence were observed in the southern districts as shown in figure 3.3.

Table 3.1: Descriptive statistics adjusting for survey design features

Factor	Category	Accounting weights and design features			Design-based F Statistics (p-value)
		HIV status			
		Negative (%) (12,736, 90.9%)	Positive (%) (1,274, 9.1%)	Total (14,010)	
Region	Northern	94.7	5.8	11.8	47.4, (<0.000)
	Central	94.2	5.3	44.2	
	Southern	86.7	13.3	44	
Place of residence	Rural	92.3	7.7	82.0	38.9, (<0.000)
	Urban	84.9	15.1	18	
Education level	No education	87.3	12.7	3.7	4.3, (0.006)
	Primary	91.4	8.6	9.0	
	Secondary	91.6	8.4	60.3	
	Higher	89.5	10.5	27.1	
Wealth index	Poorest	92.0	8.0	17.6	7.5, (<0.000)
	Poorer	93.2	6.8	19.5	
	Middle	91.5	8.5	19.4	
	Richer	91.2	8.8	19.7	
	Richest	87.9	12.1	23.8	
Gender	Male	93.4	6.6	44.7	65.6, (<0.000)
	Female	89.0	11.0	55.3	
Marital status	Never married	96.9	3.1	29.8	89.5, (<0.000)
	Married	90.1	9.9	57.5	
	Living together	88.6	11.4	4.1	
	Widow/divorced/other	77.7	22.3	8.6	
Religion	Catholics	92.2	7.8	18.9	1.68, (0.192)
	Protestants	90.7	9.3	67.6	
	Muslim/other	90.6	9.4	13.5	
Age in years	Mean(SE)*	27.3(0.1)	34.0(0.35)		362.05, (<0.000)
Age at first sex	Mean(SE)*	11.4(0.1)	8.4(0.19)		148.51, (<0.000)
Total sexual partners	Mean(SE)*	13.7(0.2)	17.9(0.45)		75.04, (<0.000)

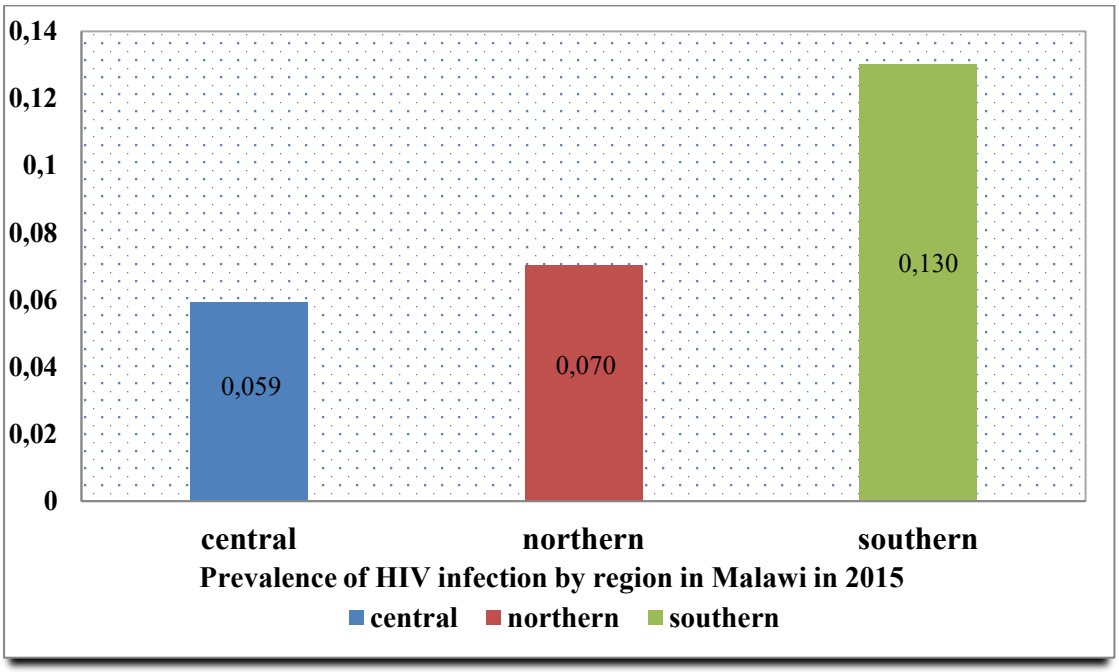


Figure 3.1: Graph showing HIV prevalence by region in Malawi in 2015

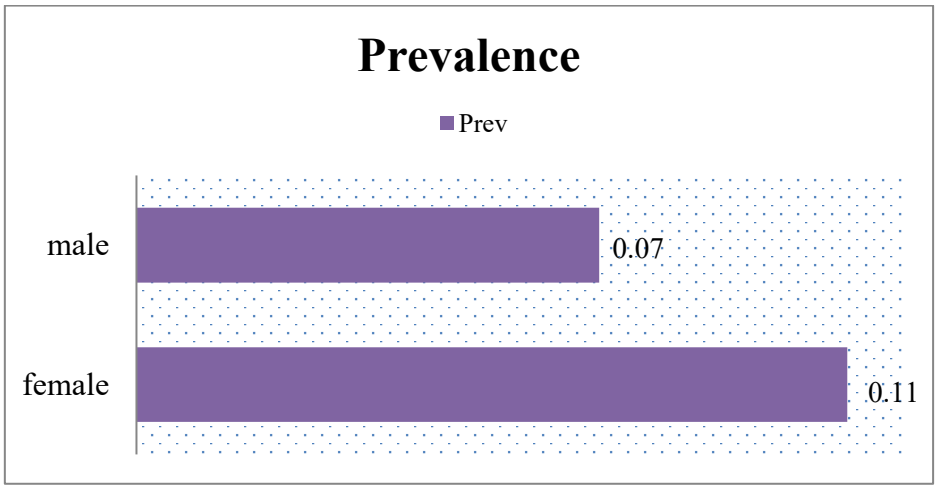


Figure 3.2: Bar Chart showing prevalence of HIV infection by gender

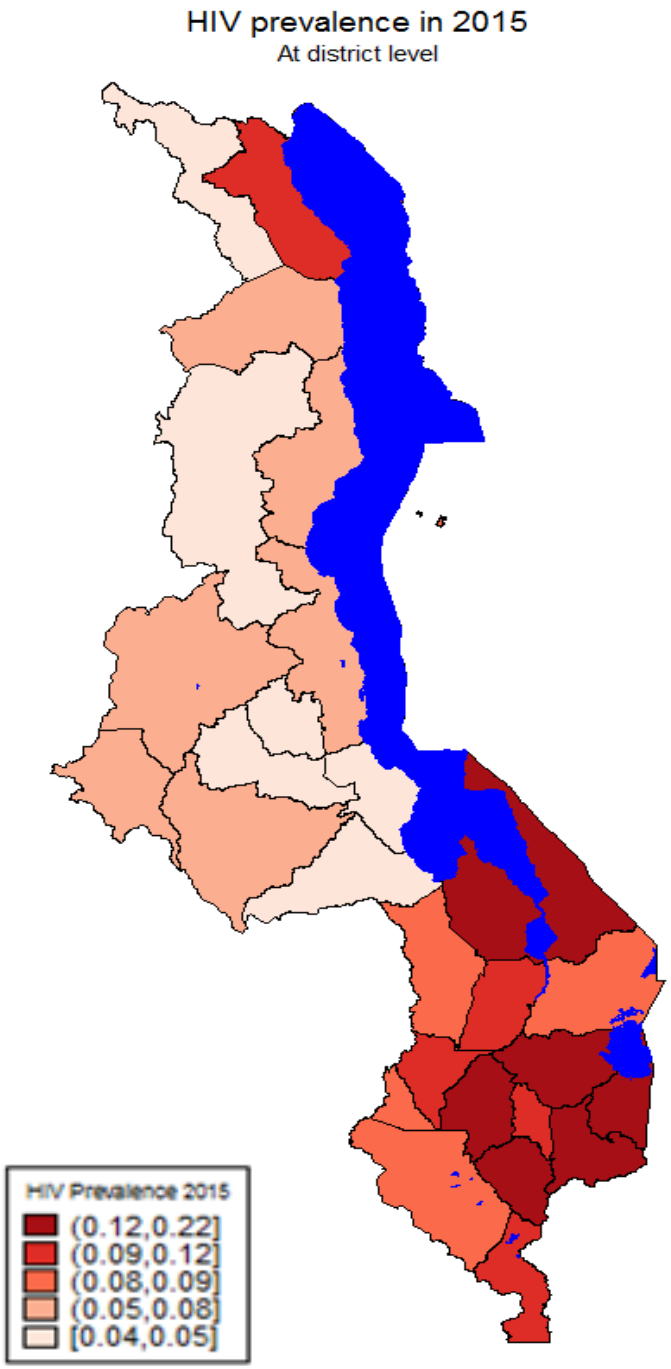


Figure 3.3: Map showing HIV prevalence by district in Malawi in 2015

3.2 Result of multivariable analysis and non-linear models

The second objective was to estimate the age-specific Forces of Infection using non-linear models. Firstly, multivariable analyses were done to assess the explanatory variables which were associated to HIV Infection. Survey logistic and multilevel logistic models adjusting for random effects were fitted. Survey survival models were also fitted, where the hazards estimates were obtained for each age which were later used in non-linear and Susceptible Infected (SI) models.

3.2.1 Result of multivariable analysis

Table 3.2 show the results of survey logistic analysis, survey survival analysis and multilevel analysis adjusting for the random effects.

3.2.1.1 Results of Survey logistic models

The variables which were statistically significant in the model at alpha level of 5% were place of residence, age in years, gender, marital status and total sexual partners. The odds of being HIV positive increased by 1.06 times as the Age increased yearly. The odds of having HIV among those who lived in urban area were 2.51 times higher than the odds among those who lived in rural area. Among males, the odds of having HIV was 54% lower than those of the females. Furthermore, the odds of having HIV increased by 1.03 times as the number of total sexual partners increased by one person.

3.2.1.2 Results of Multilevel models

The variables which were statistically significant after adjusting for the random effects (clusters) were age in years, place of residence, gender, marital status, total sexual partners and region. Individuals who lived in urban areas were 3 times more likely to be HIV positive than those in rural areas. The odds of having the HIV infection increased by 1.07 times as the age increased yearly. Furthermore, Male individuals were 52% less likely to be HIV positive as compared to the female individuals and the odds of having the HIV infection increased as the total number of sexual partners also increased. After adjusting for the other covariates, the odds of being HIV positive were 4% less likely among individuals from central region, 2.7 times higher among those from southern region than those from Northern region.

3.2.1.3 Results of Survival models

In addition to the variables which were statistically significant in the models above, education level was also significant in the survival models. The risk of having HIV was 35% lower among those who had no education than those who had higher education. The risk of being HIV positive was 7% higher among those with primary education, 39 % higher among those with secondary education as compared to those who had higher education. Furthermore, the risk of having HIV was 62 % higher among those from urban areas than those from rural areas. Males were 55% less likely to have HIV as seen in the other two models.

Table 3.2: Results of Logistic analysis, Survival analysis and Multilevel analysis adjusted for random effects

Variable	Category	Logistic estimates Odds ratio (95% CI), p-value	Multilevel logistic estimates Odds ratio (95% CI), p-value	Survival analysis Hazard ratio (95% CI), p-value
Age in years		1.06 (1.05, 1.07), 0.00	1.07 (1.06, 1.08), 0.00	
Place of residence	Rural	1	1	1
	Urban	2.51 (1.93, 3.26), 0.00	3.01 (2.23, 4.07), 0.00	1.62 (1.38, 1.90), 0.00
Gender	Female	1	1	1
	Male	0.46 (0.38, 0.56), 0.00	0.48 (0.40, 0.59), 0.00	0.45 (0.40, 0.52), 0.00
Marital status	Never married	1	1	1
	Living together	1.46 (0.95, 2.23), 0.08	1.36 (0.86, 2.14), 0.19	0.30 (0.21, 0.43), 0.00
	Married	1.26 (0.88, 1.82), 0.21	1.29 (0.89, 1.87), 0.18	0.27 (0.21, 0.35), 0.00
	Widowed/divorced/other	2.42 (1.64, 3.58), 0.00	2.65 (1.78, 3.93), 0.00	0.39 (0.29, 0.51), 0.00
Total sexual Partner		1.03 (1.02, 1.03), 0.00	1.02 (1.02, 1.03), 0.00	1.01 (1.01, 1.02), 0.00
Region	Northern		1	1
	Central		0.96 (0.67, 1.38), 0.84	1.11 (0.88, 1.40), 0.37
	Southern		2.71 (1.97, 3.72), 0.00	2.14 (1.75, 2.60), 0.00
Education Level	Higher			1
	No education			0.65 (0.47, 0.90), 0.01
	Primary			1.07 (0.81, 1.41), 0.64
	Secondary			1.39 (1.06, 1.81), 0.02
Random effects parameters Multilevel analysis				
Variance (cons)			0.60 (0.38, 0.92)	

3.2.2 Estimating the Force of Infection using non-linear models

This section presented results of estimating the age-group specific forces of infection. Firstly, Farrington and Hierarchical models were fitted in Winbugs software to estimate the age specific forces of HIV infection. Secondly, non-linear models were fitted in Stata 15 using age specific Hazard rates from survival models.

3.2.2.1 Results of Farrington and Bayesian models

This section presented results from Farrington models and hierarchical models used to estimate the force of infection as shown in table 3.3. The first model to be fitted was the exponential model assuming alpha3 is equal to zero. The deviance was 471 and the number of parameters was 1 which was lower than the true number of parameters. For the exponential model with alpha3 greater than zero, the deviance was 473.39 and number of parameters was 0.86. The exponential model (cumulative data) deviance was 11803.7 and log-logistic model's deviance was 6952.07. The effective number of parameters for exponential model (alpaha3=0) was 1 which was a bit lower than the true number of parameters (2) and for the other models the effective number of parameters were very low than the true numbers of parameters. Although, the effective number for the model with alpha3 equal to zero was a bit lower it was still closer to the true number of parameters, hence it was considered to be the best fit. It was also the best fit since the deviance was slightly lower.

Furthermore, the posterior mean of the average age at infection for the exponential model with alpha3 equal to zero is 39.25 years and the posterior mean for the number of uninfected individuals is 0.85. This was also seen in the exponential model when alpha3 is included in the model and when the cumulative data is also

used. The posterior mean of the average age at infection decreased to 22.22 years and the mean of uninfected individuals decreased to 0.50 in the Log-logistic model.

Table 3.3: Posterior means for the parameters

Parameter	Exponential (alpha3=0) mean (95% CI)	Exponential (alpha3>0) mean (95% CI)	Exponential (alpha3>0) HIV cumulative mean (95% CI)	Log-logistic mean (95% CI)
Alpha1	0.001 (0.001, 0.001)	0.001 (0.001, 0.001)	0.001 (0.001, 0.001)	0.001 (0.001, 0.001)
Alpha2	0.075 (0.072, 0.079)	0.094 (0.087, 0.100)	0.1 (0.01, 0.1)	0.001 (0.001, 0.002)
Alpha3		0.001 (0.001, 0.002)	0.001 (0.001, 0.001)	0.05 (0.003, 0.098)
Uninfected individuals	0.85 (0.84, 0.86)	0.85 (0.84, 0.86)	0.87 (0.87, 0.87)	0.50 (0.50, 0.50)
Average age at infection	39.25 (38.95, 39.54)	39.27 (38.98, 39.54)	39.95 (39.95, 39.95)	22.20 (22.19, 22.20)
Goodness of fit				
Dbar	469.91	472.53	6952.07	11803.7
Dhat	468.91	471.67	6952.07	11803.7
pD	1	0.86	0.001	0
DIC	471	473.39	6952.07	11803.7

Where**;**

Dbar- sample mean of Binomial Deviance

Dhat-deviance evaluated at posterior mean

pD- effective number of parameters

DIC-Deviance Information Criterion

3.2.2.2 Frequentist non-linear estimation of the Force of Infection

Table 3.4 shows results obtained from the multilevel non-linear models where the Hazard was the response variable. The parameters estimated were alpha1 = -0.03, alpha2 =0.05 and alpha3 = -0.14. The model included random effects at the ID level and they all had the same variance. The variance of the random intercept (Var(U1)) was 0.0001 and the variance of the error term (Var(residual)) was 6.3E-250. Figure 3.4 shows the Hazard rates across the ages 15 to 49 years. The hazard rates increase slowly from ages 15 to 30 years, then further increases from age 30 to 40 years with a further sharp increase from age 40 to 49 years.

Table 3.4: Results of Multilevel non-linear models

Parameter	Coefficient
Alpha1	-0.03
Alpha2	0.05
Alpha3	-0.14
Random-effects parameters	
id: identity	
Var (U1)	0.0001
Var (residual)	6.30E-250

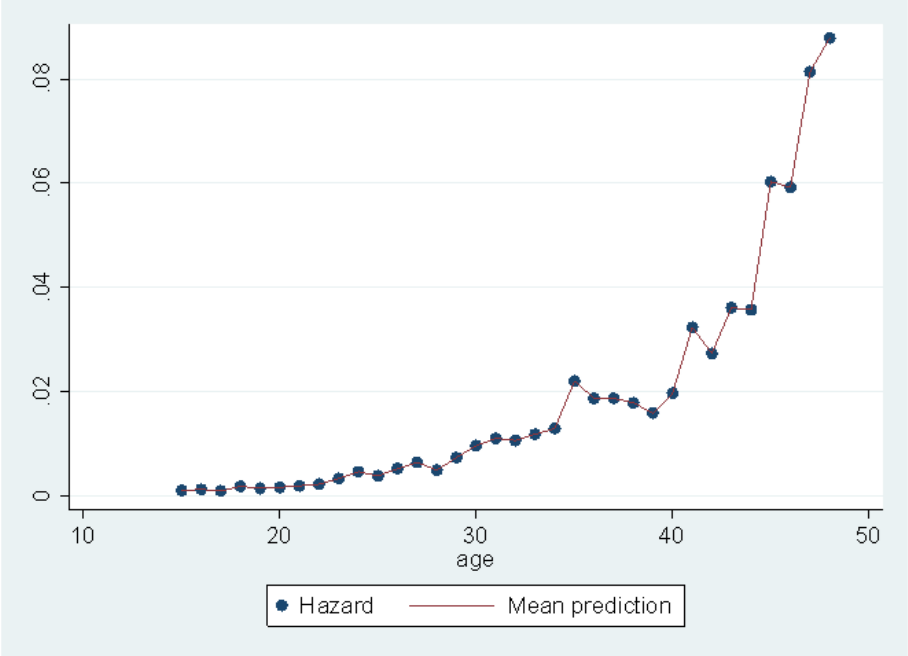


Figure 3.4: Hazard age-specific rates

3.2.3 Results of SI (Susceptible, Infected) Model

This section presents results of modelling the force of infection for HIV infection using Susceptible Infected (SI) models. In figure 3.5, we have the graph showing the Compartment sizes on the left and the disease incidence on the right-hand side. It can be shown that from time 8 (which is age 22) to time 12 (age 26) the size of the infected class increased rapidly and from time 13 (age 26) to age 49 the disease levelled off. The number of people moving from susceptible class to the infected class increased from time 5 (age 19) and reached its peak at time age 22. A sensitivity analysis was also done for the disease prevalence and incidence as

shown in figure 3.6. It can be shown that using different parameters for the probability transmission per act and the acts per person per time, the disease prevalence reached its peak at around age 26 and disease incidence reached its peak at around age 22. At time 12 age 26, the disease incidence was 112 infections, 116 susceptible individuals, 13839 infected individuals as shown in figure 3.7.

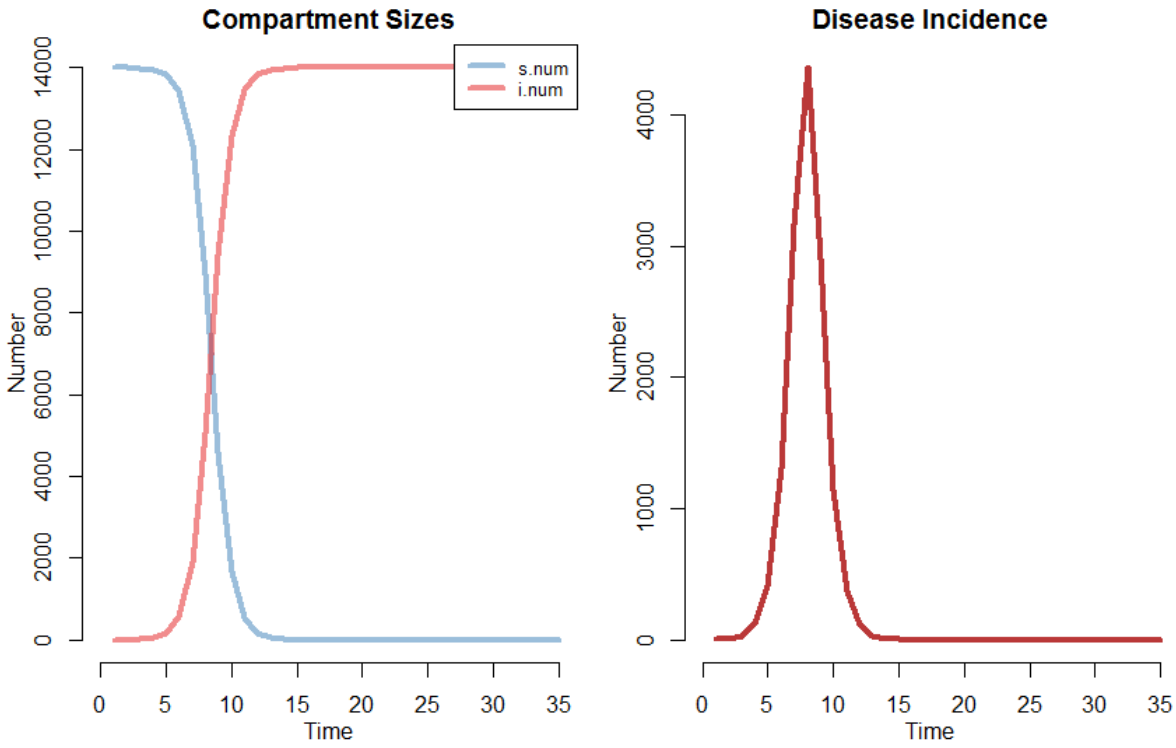


Figure 3.5: SI and Incidence plot

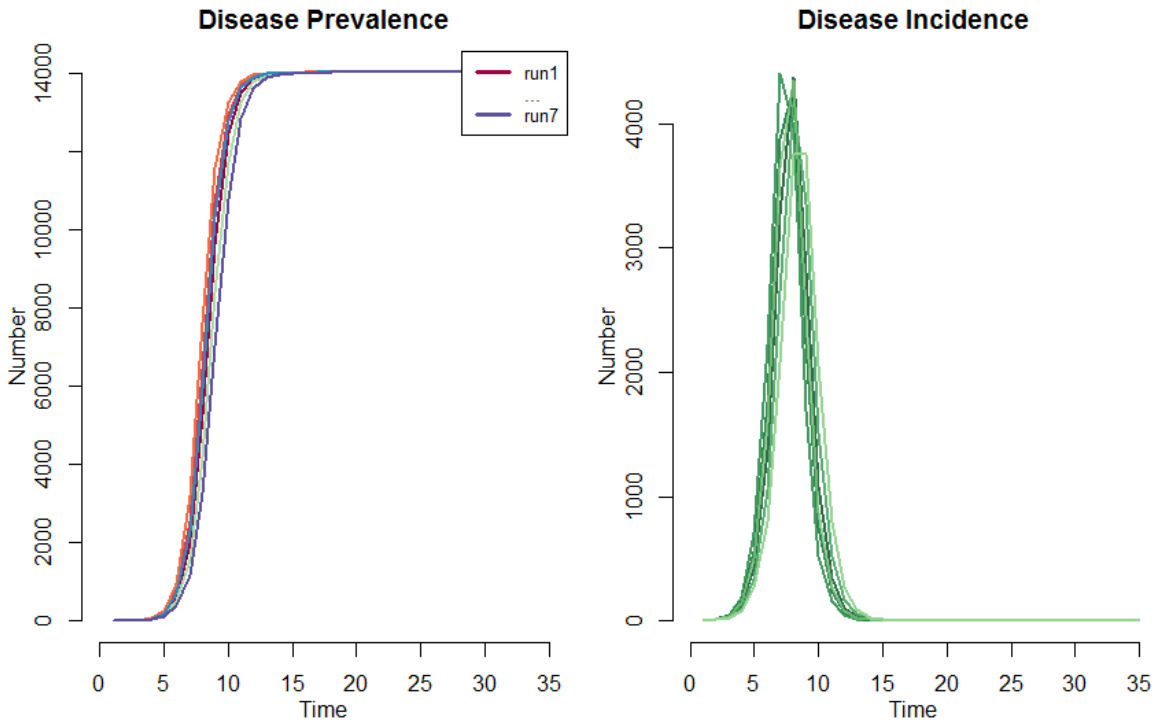


Figure 3.6: Sensitivity Analysis

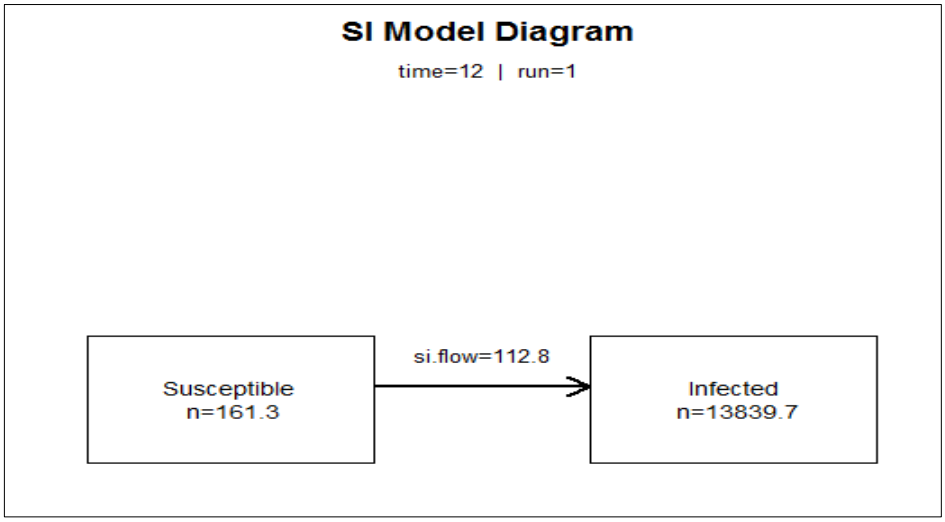


Figure 3.7: Compartment sizes at time 12 (age 26)

3.3 Concluding remarks on results

According to the results in the previous sections, it was shown that survey logistic, multilevel and survival models produced similar results. Almost all the factors significant in one analysis were also significant in the other analyses. Results from the Susceptible Infected (SI) model were from a basic SI model without other demographic parameters and mortality rates. Hence, this may not be the real situation in Malawi. Results are well explained in the discussion section.

Chapter 4 Discussion and conclusion

This chapter presents the discussion of the results of this research report, strengths and limitations of the study. It also presents conclusion and recommendations for future research studies.

4.1 Discussion

4.1.1 Discussion of Bivariate and Multivariable analysis

The study found out that HIV Infection in Malawi is associated with place of residence, gender, age in years, marital status and total number of sexual partners. The hazard risk of having HIV infection increases as the age increases. The prevalence of HIV infection was higher among female individuals than male individuals. Most countries in sub-Saharan Africa have reported that women have higher rate of HIV infection (29, 30).

Individuals who resided in urban areas had a higher risk of HIV infection unlike those in rural areas. Furthermore, hotspots of HIV infection were seen mostly in the southern region of Malawi. Several studies have also showed that individuals in urban areas have high HIV infection than rural counterparts and southern region of has higher infection rates (2, 31, 32). Individuals living in urban areas might involve themselves in risky behaviours or risky sexual encounters unlike those from rural areas.

Individuals who were never married had lower risk of HIV infection than those who were married, living together, widowed or divorced. It could be that married individuals might get the infection from their partners and those who are widowed might have gotten it from their deceased partners or after remarrying. Those who are divorced might have gotten it from their partners or they may have remarried after separation from their partners which could increase their chances of getting the infection. Other researchers in Malawi and Africa have also showed similar results (33-35).

The total number of sexual partners showed a significant association with the HIV infection. The risk of having HIV infection increased with increasing number of sexual partners. This might be due to the fact that having more sexual partners increasing the chances of contracting the HIV infection. Other researchers have also revealed a linear relationship between number of sexual partners and HIV infection risk (36).

4.1.2 Discussion of Bayesian and frequentist models of Force of Infection

The age-specific force of infection for HIV from the Bayesian analysis was best fitted with the Farrington two parameter model. The model assumed a decreasing rate of new infections as the HIV infection increases, with the highest infection at age 15 years. The decline in the force of infection could be because of increased anti-retroviral treatments which are being offered in Malawi and also behaviour changes due to awareness of the epidemic or AIDS related death. A study done by Misiri

using recurrence relation also found that HIV force of infection decreased as the ages increased (6). Other studies have modelled the FOI using splines although modelling splines only may lead to unstable projections (37). Our study used Bayesian modelling framework to estimate the force of infection for HIV. The Bayesian modelling is advantageous because it allows prior distributions to be incorporated into the models, and its flexibility in handling parameters (5, 38, 39).

Furthermore, the age specific force of infection using the Frequentist non-linear models was an increasing function, with older ages having higher HIV force of infection. However, the results from the Bayesian analysis were preferred because parameters of the frequentist non-linear model were yielding negative parameter results despite showing an increasing Force of infection.

The SI models showed that the HIV infection increased from the ages 22 to 26 years and then decreases rapidly as the age's increases. These results were from a basic SI model without other demographic parameters and mortality rates and may be different if these parameters were incorporated. Researchers have recommended compartmental models with demographic parameters, allowing for changes over time, mortality rates and age specific rates to be incorporated into models (40). This would also help in assessing the impact and cost effectiveness of interventions. However, this was beyond the scope of this research study.

Modelling the FOI is very important as its estimates helps to assess the impact of infectious disease interventions as most public health policies rely on these estimates (41).

4.2 Strengths and limitations of the study

The strengths and limitations of this study were divided into the following categories: type of data, classification of variables, confounding and effect modification.

The strength of this study was that it used a large dataset that represented the national, urban and rural populations and also various age groups. The other strength was that the data was cleaned and checked for any discrepancies. Only a few observations were removed from the analysis because they had undetermined HIV status results.

The confounding variables were dealt by the multivariable analyses and also the adjustment of random effects in the models.

The limitation of the study was that our study used cross-sectional data where we were not able to estimate time-varying FOI estimates since exposures and the HIV status were measured at one time. However, longitudinal studies are difficult to implement as there are more expensive.

4.3 Conclusions

The study showed the importance of modelling the force of infection estimates to help in eliminating the HIV pandemic. It was shown that the age specific force of infection decreased as the age's increases. The decline can be attributed by the prevention programs, increased availability of anti-retroviral treatments of HIV infection**.** However, HIV infection still remains a huge disease-burden in Malawi. This makes modelling the force of infection of HIV a very important aspect as its estimates help to develop interventions to deal with prevailing disease burdens and also checking the impact and cost effectiveness of interventions. The age-specific estimates of HIV force of infection are also very important because they easily help organizations or future interventions identify and target the most risky groups. Modelling the FOI using Bayesian methods is very useful because of their flexibility to handle parameters and also estimate probabilities of models using data provided. This research study will add on to what has already been done and also be a reference to future research studies. Future work on the modelling of the force of infection should also consider other approaches of estimating the HIV estimates.

4.4 Recommendations

The recommendations arising from this study are as follows:

- There should be more HIV research and prevention programs targeted towards the most risky groups in order to effectively eliminate the disease.
- More study researches involving modelling of the Force of Infection of HIV should be done or published in Malawi, especially comparing modelling using different techniques to provide references for other researchers
- Bayesian modelling is a very important technique and should be used more in estimating the HIV forces of infection

References

1. UNAIDS (2016) Global AIDS Update 2016. Geneva, Switzerland. Available: http://www.unaids.org/sites/default/files/media_asset/global-AIDS-update-2016_en.pdf [Accessed:28th September 2016]
2. Malawi, G. O. (2015) Malawi AIDS Response Progress Report:2015. Malawi Available: http://www.unaids.org/sites/default/files/country/documents/MWI_narrative_report_2015.pdf [Accessed: 26th October 2016].
3. MACRO & NSO. (2017) Malawi Demographic and Health Survey 2015-16. Zomba, Malawi, and Calverton, Maryland, USA: National Statistical Office and ICF Macro
4. Coates, T.J., Richter, L., Caceres, C. (2008) Behavioural strategies to reduce HIV transmission: how to make them work better. *The Lancet*, 372, 669-84.
5. Hens, N., Shkedy, Z., Aerts, M., Faes, C., Van Damme, P. & Beutels, P. (2012) *Modelling infectious disease parameters based on serological and social contact data: a modern statistical perspective,* New York: Springer Science & Business Media.
6. Misiri, H. E., Edriss, A., Aalen, O. O. & Dahl, F. A. (2012) Estimation of HIV incidence in Malawi from cross-sectional population-based seroprevalence data. *Journal of the International AIDS Society,* 15**,** 14-14.
7. Sutton, A., Hope, V., Mathei, C., Mravcik, V., Sebakova, H., Vallejo, F., Suligoi, B., Brugal, M., Ncube, F. & Wiessing, L. (2008) A comparison between the force of infection estimates for blood-borne viruses in injecting drug user populations across the European Union: a modelling study. *Journal of viral hepatitis,* 15**,** 809-816.
8. Akpa, O. M. & Oyejola, B. A. (2010) Modelling the Transmission Dynamics of HIV/AIDS epidemics: an introduction and a review. *The Journal of Infection in Developing Countries,* 4**,** 597-608.
9. Conn, P. B., Cooch, E. G. & Caley, P. (2012) Accounting for detection probability when estimating force-of-infection from animal encounter data. *Journal of Ornithology,* 152**,** 511-520.
10. Ross, R. (1916) An application of the theory of probabilities to the study of a priori pathometry. Part I. *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character,* 92**,** 204-230.
11. Rada, X., Martelli, C., Amaku, M., Sartori, A. & De Soarez, P. (2014) Modelling the Force of Infection for Hepatitis A in an Urban Population. *PLoS ONE,* 9).
12. Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, P. & Van Damme, P. (2006) Modelling age-dependent force of infection from prevalence data using fractional polynomials. *Statistics in Medicine,* 25**,** 1577-1591.

13. Pinkerton, S. D. (2012) HIV transmission rate modeling: a primer, review, and extension. *AIDS and Behavior,* 16**,** 791-796.

14. Johnson, L. (2004) An Introduction to the mathematics of HIV/AIDS modelling*:* Centre for Actuarial Research, UCT. (ed.). Cape Town.

15. Misiri, H. (2015) Estimating HIV incidence from grouped cross-sectional data in settings where anti-retroviral therapy is provided. *Research* [Online]. Available: https://www.labome.org/research/Estimating-HIV-incidence-from-grouped-cross-sectional-data-in-settings-where-anti-retroviral-therapy.html [Accessed 26th October 2016].

16. Malawi Ministry of Health. (2010) Human Resources for Health Country Profile Malawi.Available: http://ecsahc.org/wpcontent/uploads/2017/05/malawi_country_profile-Human-Resources.pdf. Accessed[17th September 2017].

17. Map of Malawi.Available: https://www.nationsonline.org/oneworld/map/malawi_map.htm [Accessed 17th September 2017].

18. Hemming, K. & Marsh, J. (2013) A menu-driven facility for sample-size calculations in cluster randomised controlled trials. *The STATA Journal,* 13**,** 114-35.

19. StataCorp. 2015. *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP.

20. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Available: http://www.R-project.org/. [Accessed 17th September 2017].

21. StataCorp. 2017. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC.

22. Lunn DJ, Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing, 10:325–337.

23. Pisati, M. (2007) SPMAP: Stata module to visualize spatial data. Statistical Software Components, S456812, Boston College Department of Economics, revised 18 Jan 2018.

24. StataCorp. 2013. Stata 13. Seemingly unrelated estimation (suest) procedures. Available: https://www.stata.com/manuals13/rsuest.pdf. [26 September 2017].

25. StataCorp. 2013. Stata 13 Logistic Regression Manual. College Station, TX: Stata Press. Available: https://www.stata.com/manuals13/rlogistic.pdf [Accessed 17th September 2017].

26. StataCorp. 2013. Stata 13 Multilevel mixed-effects reference manual. College Station, TX: Stata Press. Available: https://www.stata.com/manuals13/me.pdf [Accessed 17th September, 2017].

27. StataCorp. 2013. Stata 13 Survey data reference manual. College Station, TX: Stata Press. Available: https://www.stata.com/manuals/st.pdf [Accessed 17th September 2017.

28. Boily, M.C., Baggaley, R.F., Wang, L., Masse, B., White, R.G., Hayes, R.J., et al. (2009) Heterosexual risk of HIV-1 infection per sexual act: systematic review and meta-analysis of observational studies. *The Lancet infectious diseases*, 9, 118-29.

29. Cassels, S., Clark, S.J., Morris, M. (2008) Mathematical models for HIV transmission dynamics: tools for social and behavioral science research. *Journal of acquired immune deficiency syndromes*, 47, S34.

30. Stover, J., Andreev, K., Slaymaker, E., Gopalappa, C., Sabin, K., Velasquez C, et al. (2014) Updates to the spectrum model to estimate key HIV indicators for adults and children. *AIDS (London, England)*, 28, S427.

31. Chapotera G, Jayachandran V, Phuka J. Factors associated with HIV Infection among Educated Malawians: Analysis of the 2010 Demographic and Health Survey. 2016.

32. Zulu, L.C., Kalipeni, E., Johannes, E. (2014) Analyzing spatial clustering and the spatiotemporal nature and trends of HIV/AIDS prevalence using GIS: the case of Malawi, 1994-2010. *BMC infectious diseases*, 14, 285.

33. Anglewicz, P. (2012) Migration, marital change, and HIV infection in Malawi. *Demography*, 49, 239-65.

34. Geubbels, E., Bowie, C. (2006) Epidemiology of HIV/AIDS in adults in Malawi. *Malawi Medical Journal*, 18, 111-33.

35. Kembo, J., (2012) Risk factors associated with HIV infection among young persons aged 15–24 years: Evidence from an in-depth analysis of the 2005–06 Zimbabwe Demographic and Health Survey. SAHARA-J: *Journal of Social Aspects of HIV/AIDS*, 9, 54-63.

36. Chen, L., Jha, P., Stirling, B., Sgaier, S.K., Daid, T., Kaul, R., et al. (2007) Sexual risk factors for HIV infection in early and advanced HIV epidemics in sub-Saharan Africa: systematic overview of 68 epidemiological studies. *PloS one*, 2, e1001.

37. Hogan, D.R., Salomon, J.A. (2012) Spline-based modelling of trends in the force of HIV infection, with application to the UNAIDS Estimation and Projection Package. *Sexually transmitted infections*, 88, Suppl 2:i52-7.

38. Musenge, E., Vounatsou, P., Collinson, M., Tollman, S., Kahn, K. (2013) The contribution of spatial analysis to understanding HIV/TB mortality in children: a structural equation modelling approach. *Global health action*, 6, 19266.

39. Van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J.B., Neyer, F.J., Van Aken, M.A. (2014) A gentle introduction to Bayesian analysis: Applications to developmental research. *Child development*, 85, 842-860.

40. Williams, B.G. (2014). Fitting and projecting HIV epidemics: Data, structure and parsimony. Available: https://arxiv.org/abs/1412.2788.

41. Hens, N., Aerts, M., Faes, C., Shkedy, Z., Lejeune, O., Van Damme, P., et al.(2010) Seventy-five years of estimating the force of infection from current status data. *Epidemiology & Infection*, 138, 802-812.

Appendix A

A.1 Power Computation

Table 1: Cluster adjusted power computation command and results
. clustersampsi, binomial power p1(0.0845) p2(0.102) m(19) k(847) rho(0.15) size_cv(0) alpha(0.05) base_correl(0)
Power calculation for a two sample comparison of proportions (using normal approximations) without continuity correction.
• For the user specified parameters:
• p1: 0.0845
• p2: 0.102
• significance level: 0.05
• baseline measures adjustment (correlation): 0.00
• average cluster size: 19
• number of clusters per arm: 847
• coefficient of variation (of cluster sizes): 0.00
• intra-cluster correlation (ICC): 0.1500
• clustersampsi estimated parameters:
• Firstly, assuming individual randomisation:
• power: 1.00
• Then, allowing for cluster randomisation:
• design effect:3.70
• power: 0.80

A.2 Farrington's model and Derivations

Under the Farrington's model, the FOI was defined as

$$\lambda(a) = (\alpha a - \gamma)e^{-\beta a} + \gamma \qquad \text{eqn (A.1)}$$

Where $\lambda(a)$ - age-specific FOI

α -constant

β - Coefficient for age

a -age group

γ -long-term residual value, as $a \to \infty$, $\lambda(a) \to \gamma$ .results in the three parameter Farrington model and then for the γ =0, gives the two parameter model.

The non-linear model to estimate the FOI was given by

$$\pi(a) = 1 - e^{-\int_0^a \lambda(s)ds}$$

$$= 1 - \exp\left\{\frac{\alpha}{\beta}ae^{-\beta a} + \frac{1}{\beta}\left(\frac{\alpha}{\beta} - \gamma\right)\left(e^{-\beta a} - 1\right) - \gamma a\right\} \qquad \text{eqn (A.2)}$$

This can be solved using the frequentist or Bayesian approaches. In the Bayesian approach, the parameters are assumed to follow distributions hence not fixed, unlike

the frequentist approach where the parameters are fixed and estimated from the data.

In the Bayesian approach, after reparameterizing to express these in terms of alphas to have the vector $\boldsymbol{\alpha} = (\alpha = \alpha_1, \beta = \alpha_2, \gamma = \alpha_3)$.

Therefore the new form for the FOI for the Farrington's model becomes:

$$\lambda(a) = (\alpha_1 a - \alpha_3) e^{-\alpha_2 a} + \alpha_3 \qquad \text{eqn (A.3)}$$

Where $\lambda(a)$ - age-specific FOI

The age-specific prevalence has the parametric form

$\pi(a) = \pi(a, \boldsymbol{\alpha})$ Where $\boldsymbol{\alpha}$ is the parameter vector eqn (A.4)

Similarly, integrating $\lambda(a)$ like in the Farrington's model results into a non-linear model;

$$\pi(a) = 1 - \exp\left\{ \frac{\alpha_1}{\alpha_2} a e^{-\alpha_2 a} + \frac{1}{\alpha_2}\left(\frac{\alpha_1}{\alpha_2} - \alpha_3 \right)\left(e^{-\alpha_2 a} - 1 \right) - \alpha_3 a \right\} \qquad \text{eqn (A.5)}$$

The number of the infected individuals y_a at age group a are assumed to follow independent Binomial distributions

$$y_a \square Bin(n_a, \pi_a) \text{ ; for } a = 1, 2, ..., m \qquad \text{eqn (A.6)}$$

Where n_a is the number of individuals in age group a and m is the number of age groups.

In order to achieve the desired posterior distribution of $P\left(\boldsymbol{\pi}\middle|\boldsymbol{y},\boldsymbol{n}\right)$, a prior distribution $p\left(\boldsymbol{\alpha}\right)$ is assumed for the parameter space $\boldsymbol{\alpha}$.

The α_j 's follow the truncated normal distribution in order to get non-negative FOI

$\alpha_j \sim truncatedN\left(\mu_j,\tau_j\right)$ where $j=1,2,3$

The posterior distribution is given by

Posterior distribution $\propto$ Likelihood * Prior

Where the likelihood is given as:

$$P\left(y_a\right)=\prod_{a=1}^{m}\binom{n_a}{y_a}\pi\left(a\right)^{y_a}\left(1-\pi\left(a\right)\right)^{n_a-y_a} \qquad \text{eqn (A.7)}$$

The joint prior is given by

$$P(\boldsymbol{\alpha}) \propto \prod_{j=1}^{3} \left\{ -\frac{1}{\tau_j} \exp\left(-\frac{1}{2\tau_j^2} \left(\alpha_j - \mu_j \right)^2 \right) \right\} \qquad \text{eqn (A.8)}$$

The joint posterior distribution is given by

$$P(\boldsymbol{\alpha} | \boldsymbol{y}) \propto \prod_{a=1}^{m} \binom{n_a}{y_a} \pi(a)^{y_a} \left(1 - \pi(a)\right)^{n_a - y_a} \times \prod_{j=1}^{3} \left\{ -\frac{1}{\tau_j} \exp\left(-\frac{1}{2\tau_j^2} \left(\alpha_j - \mu_j \right)^2 \right) \right\}$$

eqn(A.9)

As the joint posterior distribution cannot be simplified any further, the MCMC simulations were used to obtain the estimates (posterior means, medians and measure of variability for use in determining the credible intervals) and these were used to estimate the FOI.

Appendix B Plagiarism Form

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

I, **Justina Kaunda** (Student number: **1476346**) am a postgraduate student registered for the degree of **Master in Epidemiology (Biostatistics)** in the academic year **2016**.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature: J. Kaunda Date: 23rd August, 2019

Appendix C Ethics clearance certificate

R14/49 Miss Justina Kaunda

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M170147

NAME: (Principal Investigator)	Miss Justina Kaunda
DEPARTMENT:	School of Public Health Division of Epidemiology and Biostastics
PROJECT TITLE:	Modelling the Non-Linear Force of Infection for HIV in Malawi in 2010
DATE CONSIDERED:	27/01/2017
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Dr Eustasius Musenge
APPROVED BY:	Prof P Cleaton-Jones, Chairperson, HREC (Medical)
DATE OF APPROVAL:	03/02/2017

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary in Room 301, Third floor, Faculty of Health Sciences, Phillip Tobias Building, 29 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in January and will therefore be due in the month of January each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES