Factors associated with World Health Organisation (WHO) clinical staging and related characteristics in HIV positive patients: An application of multistate, missing data and modelling techniques.

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Student: Murapa Batsirai

Supervisor 1: Prof Eustasius Musenge

Supervisor 2: Dr Kennedy Otwombe

A RESEARCH REPORT SUBMITTED TO THE SCHOOL OF PUBLIC HEALTH, UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG, IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN EPIDEMIOLOGY (BIOSTATISTICS)

June 2019

Declaration

I, Batsirai Amen Murapa, student No: 1734164 declare that this research report is my own work. It is being submitted for the degree of Master of Science in Epidemiology(Biostatistics) at the University of Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other University.

Batsirai Amen Murapa 7 June 2019

Dedications

This research is dedicated to my mother Noma and all Bayesian statisticians who believe in the inclusion of prior information in data analysis.

Abstract

Background

Human Immunodeficiency Virus (HIV) remains a significant problem in sub-Saharan Africa which has the highest number (25.6 million) of people living with HIV (PLHIV). South Africa is amongst the top rank of sub-Saharan Africa countries with the highest HIV prevalence. Many studies have been done to have an in-depth understanding of HIV and AIDS disease resulting in various interventions being implemented to improve the lives of people infected by the disease. These studies are usually done using longitudinal designs which have the advantage of enabling researchers to observe patient changes (outcomes) over time; however, they are prone to missingness due to unobserved data as patients may miss scheduled visits.

This study aims to determine transition probabilities between WHO stages I, II, III and IV over time and compare Rubin's and Bayesian methods in determine factors associated with WHO stage ailments and symptomatic conditions amongst HIV infected patients on patient level data from the Adult Wellness study.

Methods

The researcher conducted a secondary data analysis of the Adult Wellness study which was conducted from 2002 to 2010, to be able to quantify changes in ailments and symptomatic conditions over time, the researcher fitted the general multi-state Markov model which assumes that patients may develop more severe ailments and symptomatic conditions. The states were defined based on WHO stages, that is, stage I, II ,III and IV. The researcher also fitted three random effects ordered logistic regression models to determine factors associated with these WHO stage outcomes. The researcher employed the maximum likelihood estimation (MLE) on the first model fitted on raw data and second model after multiple imputation (MI) to account for missing data. The last model adopted Bayesian estimation (BE) to the raw data. Finally, the researcher performed a sensitivity analysis using simulated data and fitted all the three models described earlier.

Results

A total of 2,609 patients accounted for 12,102 observations were analysed. Majority of the patients were females (77.4%) antiretroviral therapy ART naïve (61.5%) having attained Grade 0-12 (77.9%). The Markov multi-state model showed that patients in WHO stage II were 1.33 times more likely to move to WHO stage III than WHO stage I whilst patients in WHO stage III were 2.26 times (0.16118/0.07124) more likely to move to WHO stage II than progressing to WHO stage IV. The probability of remaining in WHO stage I was 59% after eight year follow up period. Relative to patients with no ailments and symptomatic conditions, patients with one HIV ailments or symptomatic condition has an INCREASED RISK of progressing to advanced WHO stages, (model with raw data OR 2.07, 95%CI: 1.23-3.30). There were some unexpected results were patients on ART and cotrimoxazole (CTX) drugs showed to have increased odds of becoming worse than their counterparts. This was evident in all the three models: raw data MLE model OR 1.5828 (95% CI 1.1948,2.0969) and OR 2.0670 (1.5619,2.7355); MI MLE model OR 1.2252 (95% CI 1.0884,1.3793) and OR 1.5438 (95% CI 1.3186,1.8074); and raw data BE model OR 1.6096 (95% CI 1.2055,2.0952) and OR 2.0758 (1.5507,2.7270) for ART and CTX respectively. Both MLE and BE for the raw data gave similar estimates; however, these estimates were different from the MI MLE model which were more precise (smaller standard errors).

Conclusions

The results showed that patients had an increased chance of remaining in the same state than either advancing in WHO stages of ailments and symptomatic conditions or recovering. If the level of missing data is reasonable, it is recommended to apply the MI techniques. Multiple imputations and Bayesian missing data methods should be used together and determine which one produce better results per each situation. Finally simulated results showed that multiple imputation and Bayesian models become different as the percentage of missing data increases.

Preface

This Project Report has been done in partial fulfilment of a Master of Science in Epidemiology (Biostatistics) in the Academic years 2017 to 2018. I received data from the Perinatal Health Research Unit (PHRU), hence I conducted a secondary data analysis of the Adult Wellness Study which was conducted between 2002 to 2010. For this reason, I would like to express my gratitude to PHRU for granting me the permission to use the dataset. I would also like to extend my gratitude to the sub-Saharan African Consortium for Advanced Biostatistics (SSACAB).

The research report starts with the background and literature review in Chapter 1, methods in Chapter 2, results in Chapter 3 and discussion and conclusions in Chapter 4.

Acknowledgement

I want to thank my supervisors Professor Eustasius Musenge for guiding me through my research and reviewing my thesis; and Dr Kennedy Otwombe for their advice and input in the writing of this research report. I also extend my gratitude to Mrs Zvifadzo Matsena-Zingoni for the final review of my research report and help me structure it for submission. I am also thankful for the help I received from my lecturers from Wits School of Public Health. I would also like to thank Sub-Saharan African Consortium for Advanced Biostatistics (SSACAB) training programme for funding my studies. Finally I would like to thank the Perinatal Health Research Unit for providing me with the data they collected during the Adult Wellness study.

List of Figures

List of Tables

Contents

Abbreviation

- AIDS - Acquired Immunodeficiency Syndrome
- ART - Antiretroviral Treatment
- BE - Bayesian Estimation
- BMI - Body Mass Index
- CD-MCAR - Covariate Dependent Missing Completely at Random
- CD4 - Cluster of Differentiation 4
- CI - Confidence Interval
- CTX - Cotrimoxazole
- GEE - Generalized Estimating Equations
- HIV - Human Immunodeficiency Virus
- LOCF - Last Observation Carried Forward
- LTFU - loss-to-follow-up
- MAR - Missing at Random
- MCAR - Missing Completely at Random
- MICE - Multiple Imputations Chained Equations
- MI - Multiple Imputation
- MLE - Maximum Likelihood Estimator
- MNAR - Missing not at Random
- MCMC - Markov chain Monte Carlo

- MSM - Markov Multi-State Model
- PHRU - Perinatal HIV Research Unit
- PCR - Polymerase Chain Reaction
- PLHIV - People leaving with HIV
- WHO - World Health Organization
- SSACAB - Sub-Saharan African Consortium for Advanced Biostatistics
- SAS - Statistical Analysis System
- STI - Sexually Transmitted Infections
- TB - Tuberculosis
- VCE - Variance-Covariance Estimate

Key Definitions

- Ailments and symptomatic conditions - These are clinical features an HIV patient is diagnosed with besides the HIV infection. These include ailments like diarrhoea but also include clinical features like weight loss.
- Multi-state Markov model - Is a fictional way of describing a transition process in which something (in our case ailments and symptomatic conditions) moves through a series of stages or states in continuous time.
- General Markov multi-state model - This is a Markov model where transitions happens in both directions.
- Type 1 censoring - Censoring is called type one censoring if the censoring times are known in advance.

Chapter 1

Introduction and Background

Introduction

The chapter begins with an overview of the burden of Human Immunodeficiency Virus (HIV) in sub-Saharan Africa, with more emphasis on South Africa. The effects of missing data in longitudinal studies in HIV care treatment programs are explored. The researcher will also look at the movement of HIV patient's ailments and symptomatic conditions from one World Health Organization (WHO) stage to another. The literature on missing data and transition of individuals from one WHO stage to another are reviewed, and the chapter ends with a description of the aims and objectives of the study.

1.1 Background

Globally, the number of people living with Human immuno-deficiency virus (PLHIV) remains high. In 2017 there were more than 36.7 million people who are HIV infected with the majority of them being adults (15-49 years) and 160,000 are children below the age of 15. HIV-related deaths are still high with an estimated 940 000 died in 2017. The majority of these HIV patients are found in Sub-saharan Africa hence there are more died in this region. In 2016 sub-Saharan had the highest number of people living with HIV (PLHIV) (25.6 million)[1]. By 2016, 68 percent of PLHIV in the world were from this region [2].

South Africa which is within sub-Sahara Africa, ranks amongst the highest in HIV prevalence (12.6 percent as of July 2017) and approximately 7 million PLHIV [1]. The most substantial absolute number of HIV patients on antiretroviral treatment (ART) globally is found in South Africa with 3.1 million people on ART [3]. As of 2016, South Africa was spending R23 billion every year on HIV and AIDS programs [4].

PLHIV are at high risk of developing comorbid conditions such as tuberculosis (TB) and symptomatic manifestations weight loss, due to an immuno-compromised system [5]. These ailments and symptomatic conditions do not remain the same over time. They can range from mild such as skin rash to severe ones such as pneumocystis pneumonia. World Health Organization (WHO) classify them into groups based on the severity of illness known as WHO stages [6]. These ailments and symptomatic conditions can cause death if medical care is not adhered to.

The functionality of patients with multiple ailments is decreased. Ailments and symptomatic conditions also affect the manner in which patients attend scheduled health care services such as the provision of ART, leading to missing data from unattended visits. Disadvantages of missing data in reseach include loss of information, a low statistical power to detect differences and wider confidence intervals (due to increased standard errors) [7]. Missing data affects decisions drawn from such inferences due to biased point and interval estimates. Previous studies also confirm that statistical inferences from missing data may lead to wrong conclusions [8].

This study will estimate the transition of people with ailments and symptomatic conditions based on WHO staging and determine factors associated with these ailments and also use missing data techniques.

1.2 Problem Statement

1.2.1 Epidemiological problem statement

Having HIV infection is not an illness but it gives rise to several ailments due to immunosuppression if a patient does not adhere to treatment [9]. Hence it is paramount to find factors associated with these ailments and symptomatic conditions and how they change overtime. This may help in the reduction the these ailments and symptomatic conditions so that HIV patients can live a health life.

1.2.2 Statistical problem statement

Several HIV related longitudinal epidemiological studies have been undertaken to understand the effects of HIV on disease progression [10]. Longitudinal studies are often affected by missing data, as a result, complete case analysis is normally used [11, 12]. In addition, most of these studies control for missing data using complete case analysis [12, 13]. This suggest that statistical inferences from such analysis may lead to incorrect conclusions and mislead decision making.

1.3 Justification

1.3.1 Epidemiological justification

The classification of HIV ailments and symptomatic conditions using WHO stages have been done [14]. However, to our knowledge, none has been done in South Africa where factors associated with these WHO stage ailments and systematic conditions including changes occurring over time using the general Markov chain models.

1.3.2 Statistical justification

In clinical research, modelling of longitudinal data using random effects ordered logistic regression models or general estimating equation (GEE) can be fitted to determine factors associated with ailments and symptomatic conditions of HIV patients. However, these models do not give information on transition probability from one state to the other for patients on treatment overtime, hence general Markov multi-state model is preferred instead.

The data being used is longitudinal and is generally affected by missing data. The most commonly used missing data technique is multiple imputations, but also implementing Bayesian missing data technique will give us more room for choosing prior distributions and perhaps obtain better results [15, 16].

1.4 Literature Review

1.4.1 Ailments and symptomatic conditions in HIV infected patients

HIV ailments and symptomatic conditions are known to be complicated since AIDS was defined in 1981 [17]. The initial clinical presentation may look like symptoms of any other endemic diseases [18]. However, the extent of the manifestation of ailments and symptomatic conditions depends on the infected individual's health status and organs affected [18]. The most common ailments and symptomatic conditions among HIV infected patients in Africa are diarrhoea, tuberculosis, fever, weight loss and a variety of opportunistic infections and dermatological symptoms [18].

Majority of these HIV-related ailments and symptomatic conditions are in sub-Saharan Africa where there is a high burden of HIV [19]. South Africa having the highest absolute number of HIV infections has an influence on the burden of HIV related ailments and symptomatic conditions in their population [1, 20].

1.4.2 Points to note when modelling data with general Markov multi-state

- The next event depend on the present not the past (this is the Markov property) [21].
- If a Markov multi state model (MSM) is fitted without taking into consideration data sampling times then the results maybe biased[21].
- To fit an MSM model to longitudinal data the researcher must consider why observations were made at those times. This is analogous to missing data problem where the researcher need to know why a particular observation missing may give you information about the observed [21].
- If an MSM is fitted ignoring the information about sampling times then the results may be biased. This is so because sampling times are usually random and they should be modelled together with observed process[21].

1.4.3 Review of studies modelling disease progression with general Markov multi-state models

Kendra Brown Mhoon et al. fitted a continuous time general Markov multi-state model to show the stages of the construct from a health intervention. This study looked at whether a person was willing to change their behaviour after an intervention (Healthy Outlook on the Mammography Experience), a population-based randomised group intervention [22]. M. J Sweeting et al. also used general Markov multi-state model to model the progression of hepatitis C from none or mild to severe cirrhosis [23].

1.4.4 Missing data in HIV health research

HIV is a chronic condition that affects human health resulting in weight loss and being vulnerable to opportunistic infections like Tuberculosis(TB). These chronic conditions are usually studied using longitudinal studies, of which most of the data will have missing information. Missing data in general, gives rise to biased estimates, hence compromising results [24]. An example of such study was conducted in Zambia and South Africa from 2006 to 2010 on TB and AIDS reduction whereby laboratory data from HIV testing was used to confirm HIV status of the participants. Those participants with missing laboratory HIV results were asked to verbally confirm their status. In the event that there were no results from both the laboratory and self reporting,the missing values were imputed using Rubin's techniques assuming the data were missing at random (MAR) [25]. The second example comes from a Medical Outcome study. This study looked at how missing health related quantities of life could impact conclusions in HIV clinical trials. Patients were followed up after every 8 weeks for 48 weeks, however some were lost to follow-up during the study [26]. Imputation was done using last observation carried forward(LOCF)(this is were a researcher uses the previously observed value to impute the present missing one) [26]. Kennedy Otwombe et al also used MI in their study looking at, "factors associated with mortality in HIV-infected people in rural and urban South Africa". They assumed MAR mechanism for patients who had dropped out or missed visits [12].

1.4.5 Categories of missing data

Let R_{it} be 1 if the i^{th} patient has an observation at time t and 0 if the observation is missing at time t. Therefore R is a binary variable of whether the patient has been observed or not. The researcher can now consider classifying variables that have missing data by their relationship to the probability that $R_{it} = 1$. If factors are uncorrelated to this probability then they can not be related to the missing data process.

In 1976 Donald B. Rubin classified missing data into three mechanisms according to how they became missing [27]. The researcher consider missing data to be missing completely at random (MCAR) if and only if probability that $P(R_{it} = 1)$ does not depend on any of the observed variables. This can occur when HIV status data point is missing because of a random error in data entry [28]. When data are MCAR, estimates that are estimated are unbiased. However, missing completely at random assumption is not usually a reasonable assumption. For example, in longitudinal studies if there are missing data, there almost certain that there will be more missing data in the study. So, the predictor time or visit would usually be associated with the chance that an observation is missing.

There is minor, but essential, variation of the definition when missingness depends on covariates it is termed "covariate dependent missing completely at random" (CD-MCAR), which is primarily applicable to missing outcome data. In this scenario, the probability of the outcome's missingness can depend on the covariates which are included in the statistical model only.

Let X^{obs} representing all the observed data for covariates which will be used in our model, can formally write this scenario mathematically as

$$P(R_{it} = 1|Y, X) = P(R_{it} = 1|X^{obs}) \tag{1.1}$$

If by chance that an observation is missing is dependent on information which was not observed, like the the observation itself, missing data are said to be missing not at random (MNAR) [29]. For example, if the HIV status from a prevention trial is more likely to be missing in patients that are HIV [28].

Mostly, missing data are not MNAR or MCAR. Instead, the chance that observation is mainly missing depends on information of the data that is not missing [30]. This is a more flexible assumption because this assumption handles different plausible scenarios. Under MAR mechanism the chance of missing data may depend on covariates as well as observed outcome(s) [30]. Mathematically, with Y^{obs} as all the observation this can be represented formally as

$$P(R_{it} = 1|Y, X) = P(Rit = 1|Y^{obs}, X^{obs}) \tag{1.2}$$

1.4.6 Points to note when analysing data with missingness

- When non observed data are MCAR, any method you use to analyse data will give unbiased estimates [31].
- When the outcome depends on covariates, and those covariates are used to fit a regression model. Any method you use for analysis will give credible estimates [32].
- When the outcome variable data are missing at random, maximum likelihood analysis like linear mixed models will give credible estimates but other methods like generalised estimating equations may give biased estimates.
- When the observed data are MNAR, any standard methods of analysis may give biased estimates [31].
- Going from MCAR to CD-MCAR allows the usual situation in were missing data

depends on observed covariates. Moving from CD-MCAR to MAR accommodates even more dependence of the missing data on observed covariates and outcomes and will, therefore, include missing data mechanisms that have a higher probability of being utilised in practice. This makes likelihood methods attractive because they continue to give reliable estimates even if the data are MAR. To reflect this fact, data which are MAR, MCAR or CD-MCAR are sometimes said to have ignorable missingness [33, 31].

1.4.7 Modelling missing data with random effects ordered logistic regression

Random effects ordered logistic regression models are widely used in modelling data with missingness. The main reason for this is that they use the data in its raw form instead of summaries unlike generalized estimating equations(GEE). Also if an observation is missing it has no effect on other observations that are not from that same patient [34, 35]. It also an advantage of random effects ordered logistic regression models is that the researcher does not have to be consistent with time, which means unobserved time values does not affect observed ones in analysis [36, 35].

Another advantage of these models is that the researcher do not have to assume compound symmetry or sphericity in our model. The researcher can assume compound symmetry or sphericity , but he/she can allow the model to choose a set of explanatory variables or use explanatory patterns that are supplied [36, 37, 38].
They also use missing at random assumption that is more applicable compared to missing completely at random used by GEE which is difficult to attain in Adult Wellness data [39].

1.4.8 Review of studies Bayesian and Rubin's missing data methods

Some studies have been done comparing Bayesian and Rubin's method. A study done by G. Frank Li et al. (2016) was analysing data from Schizophrenia clinical trial showed that the two methods gave similar estimates [11]. A simulated study done by Kaifeng Lu et al. (2010) demonstrated the Bayesian approach to be conservative on parameters when analysing dichotomised data while Rubin's usually overestimate the true variance [40]. Another simulated studies done by Ibrahim J.G et al and another by J.G Qingxia Chen and Joseph G on linear regressions showed that the results from the Bayesian and Rubin's methods were similar [41, 42].

1.5 Research Question

1.5.1 Epidemiological Research Question

What are the factors associated with ailments and symptomatic conditions amongst adult HIV infected patients and how these ailments change over time?

1.5.2 Statistical Research Question

Can (Multiple Imputation) MI technique give better precise estimation compared to (Bayesian Estimation) BE when evaluating factors associated with ailments and symptomatic conditions?

1.6 Aim and Objectives

1.6.1 Epidemiological Aim

This study aims to determine factors associated with WHO stage ailments and symptomatic conditions of HIV infected patient overtime on ART treatment.

1.6.2 Statistical Aim

This study also aims to determine whether there is an improvement offered by Bayesian missing data method compared to the most commonly used of Rubin's method.

1.6.3 Specific objectives

1. To determine transition intensities between WHO stages of ailments and symptomatic conditions amongst adult HIV infected patients enrolled in the Adult Wellness study, 2002 to 2010

2. To ascertain factors associated with WHO stage ailments and symptomatic conditions amongst adult HIV infected patients using raw data with BE and MI data with MLE (a complete case analysis).

3. To compare parameters estimates produced by Rubin's and Bayesian missing data methods using Adult Wellness data.

4. To compare parameter estimates produced by Rubin's and Bayesian missing data methods using simulated data

Chapter 2

Methodology

Introduction

In this chapter, the Adult Wellness study, study design, data management and processing, data quality, inclusion and exclusion criteria, data source and variables and exploratory analysis are described. The exploratory analysis described the baseline characteristics, and also continuous variables are summarised with between, within and overall standard deviations of the panel data. Categorical variables are summarised as counts and percentages of overall, between and within components of the panel data. On inferential statistics, the researcher fitted general Markov multi-state model which constitutes transition rates, probabilities of moving from one state to the other. The researcher also fitted random effects ordered logistic regression models with raw data, multiple imputations and Bayesian missing data methods. Finally, MI method and Bayesian missing data were compared methods using simulated data.

2.1 The Overview of the Adult Wellness Study

Study design and study site

The Adult Wellness study was a follow-up study conducted from 2002 to 2010. It was a nurse-based wellness programme set up to provide pre and post-treatment care for HIV positive adults. The participant enrolment times were different, hence this study was an open cohort. The patients were follow-up at every 6 months from time of individual enrolment for 8 years. The study was carried out at two study sites which were Perinatal HIV Research Unit (PHRU) clinic in Johannesburg and Tintswalo Hospital also known as Acornhoek Hospital in Mpumalanga.

Study population

The study enrolled adult HIV patients aged 18 years and above who were receiving voluntary counselling and testing from centres surrounding the PHRU and Tintswalo hospital.

Study sampling and inclusion criteria

The participants were sampled using convenience sampling. Patients who tested HIV positive and had a CD4 cell count less than 200 cells/mm3 were initiated on ART treatment [12]. The study included only those patients aged 18 years or older who tested HIV positive by either Elisa or Polymerase Chain Reaction(PCR), receiving basic clinical care at PHRU clinic or Tintswalo Hospital and lastly those who had provided a written informed consent to participant in the study [12].

Data collection and management

Questionnaires were completed at baseline and on every subsequent visit that occurred after every six months for eight years. The patients' information collected include CD4 cell counts (cell/uL), weight (kg), creatinine and haemoglobin levels measured after every 6 months. Patients were treated for any other ailments and symptomatic conditions they would present with during the course of the study period [12]. During the study, loss-to-follow-up (LTFU) was minimised actively by ways of letters, telephone calls and home visits [12]. The project data were double entered by independent data clerks into an MS ACCESS database, and any discrepancies were corrected by the data manager and queries were resolved by the principal investigator.

2.2 Current study design

This current study is a secondary data analysis of a retrospective cohort study Adult Wellness Study conducted between 2002 to 2010.

2.3 Current data management and processing

Data was received in Statistical Analysis System (SAS) database format and converted to Stata 15 format which is the software used for data management and some of the analysis in this study. Additionally analyses were done using R program [36]. The data management process includes merging the received dataset files into one master file, labelling of the variables and generating new categorical variables from existing variables using Stata statistical software (Release 15, College Station, TX: StataCorp LLC).

2.4 Current study inclusion/exclusion criteria

The merged dataset had a total of 44 461 observations. Many of these had repeated visit dates, and some had missing WHO states (the researcher did not want to impute the outcome variable hence the researcher decided to remove missing values of the outcome). The researcher also removed observations of dead patients because there were only interested in WHO states without the absorbing state. Finally, the researcher removed observations that had one observation since they had no transitions that occurred on them. Figure 2.1 quantifies the number of observations that were dropped and why they were dropped.

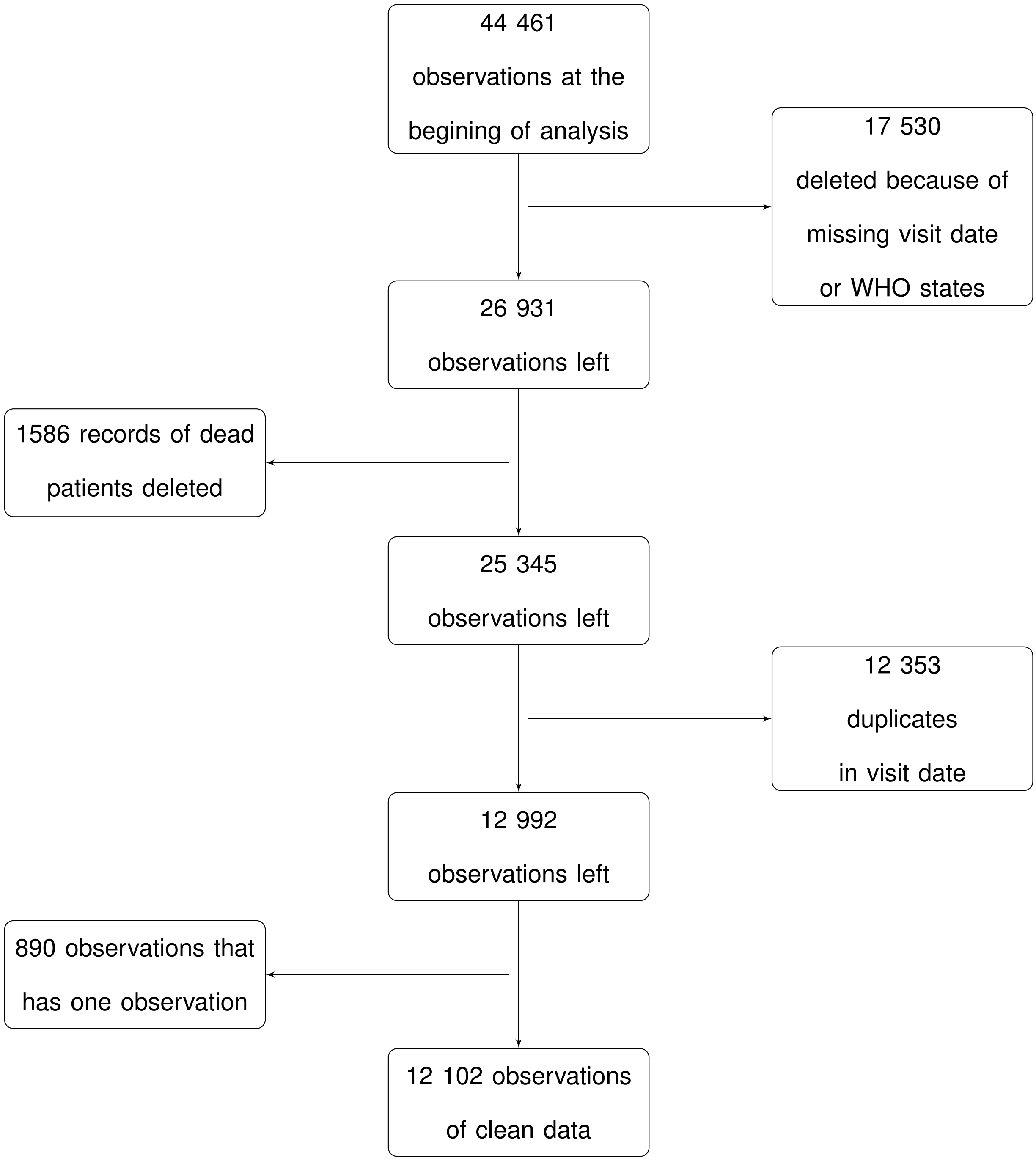


Figure 2.1: The flow diagram of selection of the HIV infected patients for analysis from the Adult Wellness study, 2002-2010

2.5 Missing data frequencies and percentages in Adult Wellness data

Table 2.1: The missingness patterns of each covariate used for analysis from the raw data.

Variable	**Observations**	**Missings**	**Percentage Missings**	**Non Missing**	**Non Missing percentage**
Age in years	12102	12097	99.96	5	0.0413
Patient on ART	12102	1920	15.87	10182	84.13
Patient on cotrimoxazole	12102	1436	11.87	10666	88.13
Education level	12102	6431	53.14	5671	46.86
Smoking status	12102	1453	12.01	10649	87.99
Patient receiving salary	12102	6031	49.83	6071	50.17
Marital status	12102	4853	40.1	7249	59.9
self employed	12102	6192	51.17	5910	48.83
Sex Age	12102	11047	91.28	1055	8.718
Alcohol	12102	1481	12.24	10621	87.76
Height	12102	5658	46.75	6444	53.25
Weight	12102	5701	47.11	6401	52.89
condom	12102	0	0	12102	100
sex partners	12102	8722	72.07	3380	27.93
BPSys	12102	4430	36.61	7672	63.39
BPDia	12102	4430	36.61	7672	63.39
STI	12102	2475	20.45	9627	79.55
CD4 count	12102	11603	95.88	499	4.123

2.6 Outcome and exposure variables

2.6.1 Exposure Variable

The exposure variables used were measured either at baseline or during follow-up visits. Variables like site were patient were treated were measured only at baseline while variables like body mass index(BMI) were measured at each visit. See Appendix B.1 of the exposure variables.

2.6.2 Outcome variables

The outcome variable was WHO staged ailments and symptomatic conditions. It's an ordinal variable. There were grouped in four categories which are WHO stage I (showing no symptoms), mild disease (mild disease) WHO stage II, (WHO stage III (moderate disease) and WHO stage IV (severe diseases). See Appendix B.2 shows ailments and symptomatic conditions staging.

2.7 Data analysis procedures

The data analysis conceptual framework is shown in Figure 2.6 from exploratory analysis to fitting the models.

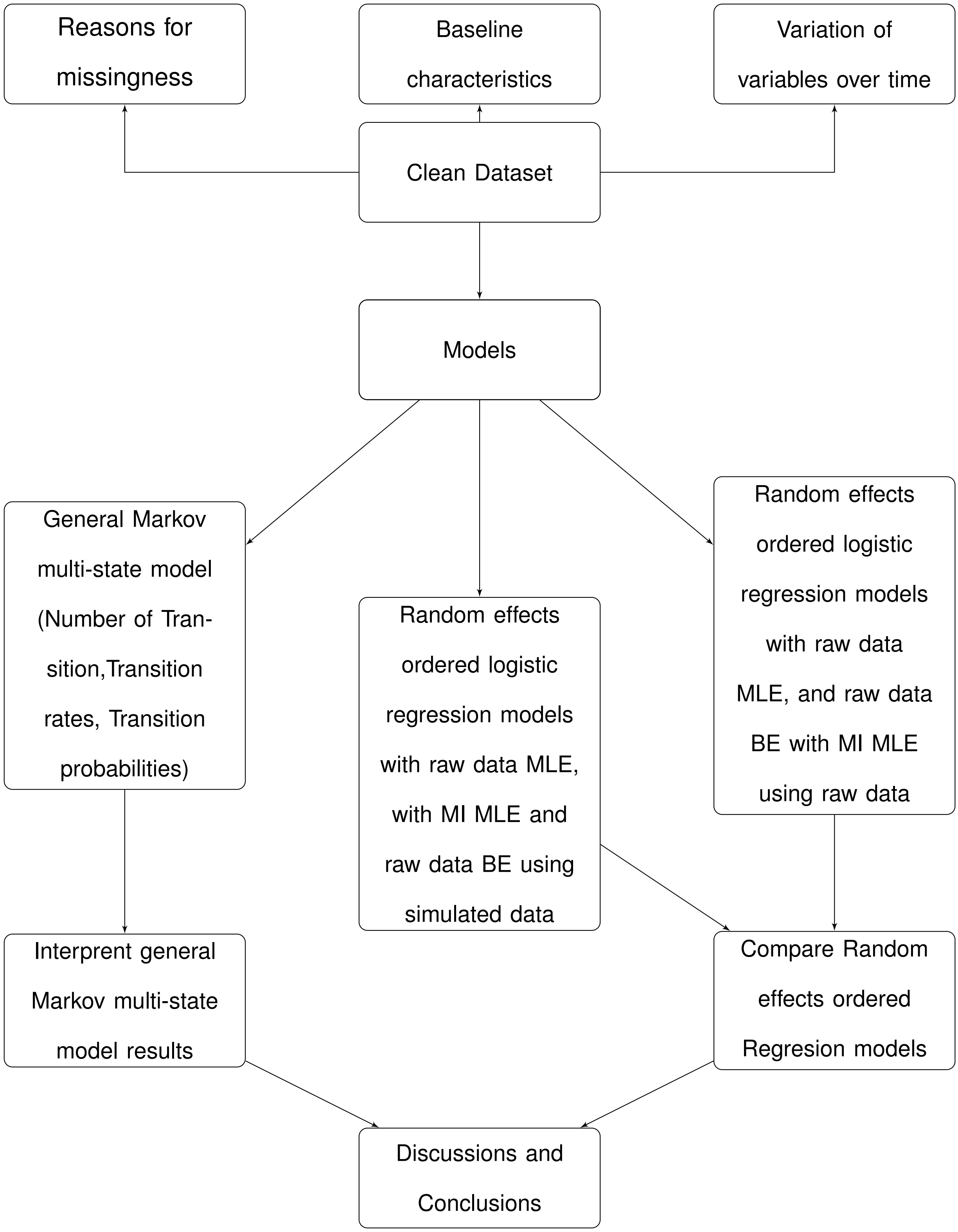


Figure 2.2: Schematic presentation diagram of the analysis stages employed in this study.

2.7.1 Exploratory analysis of Adult Wellness data

The researcher first found the reasons for missingness and then performed an exploratory data analysis whereby the data was summarised appropriately either as categorical variable or continuous variable. Categorical variables were summarised as counts (frequencies) and percentages of overall, between and within components of the panel data. The overall part summarises results in terms of person-years observed individuals had or did not have a condition of interest, between summarises the number of individuals that had or did not have a condition during follow up and within gives us a fraction of time an individual spent with a particular condition. Continuous variables were as means and standard deviation. The researcher estimated the between, within and overall standard deviations of the panel data and the global mean. The within and overall are calculated over person-years of data.

2.7.2 Multi-state Markov models

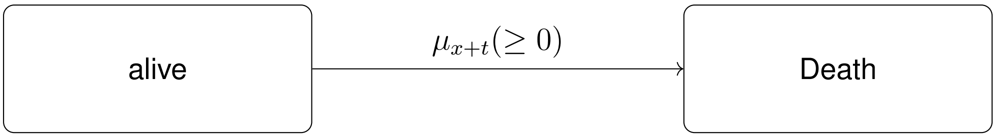
For objective one the researcher fitted the multi state model assuming a Markov process in a continuous time homogeneous approach. The researcher estimated the number of transitions patients made between the defined WHO states, i to j, which can be illustrated as:

$$\begin{bmatrix} n_{11} & n_{12} & n_{13} & n_{14} \\ n_{21} & n_{22} & n_{23} & n_{24} \\ n_{31} & n_{32} & n_{33} & n_{34} \\ n_{41} & n_{42} & n_{43} & n_{44} \end{bmatrix}$$

where n_{ij} for $i = 1, 2, 3, 4$ and $j = 1, 2, 3, 4$ is the ij^{th} patient that have moved from state i to state j.

Multiple state models

Multiple state models were outlined by Sverdrup and Waters in 1965 and 1984 respectively, and others. In this section the researcher show a basic two state model, which include birth and death known as the survival model [43]. This can be represented in a diagram



The chance that a living individual dies at any given age dependent on the transition intensity which is given by $\mu_{x+t}(t \geq 0)$

Assumptions of Markov model

- The probabilities that an individual will die at any age is dependent on the state it is currently in. This is called the Markov assumption.
- $dtq_{x+t} = \mu_{x+o(dt)}$
- μ_{x+t} for $0 \leq t < 1$ is a constant μ

Modelling time to death

This section describes the simple time to death model because it is the basic model of multistate models. The assumption of the model is that time to death can be modelled as with a random variable T taking values in a range (0 , w) where w can be taken as limiting age:
Let

$$F(t) = P[T \leq t] \tag{2.1}$$

be the distribution of T, and define

$$S(t) = 1 - F(t) = P[T > t] \tag{2.2}$$

to be the survivor function of T and define the force of mortality to be

$$\mu_t = \lim_{dt=0} \frac{1}{dt} P[t < T \leq t + dt/T > t] \tag{2.3}$$

assuming that the limit exist.
Now define a set of random variables $\left\{T_x\right\}_{x=0}^{x=w}$ where T_x is defined as the life time after age x. $T_0 = T$. The distribution and survival function of T_x denoted $F_x(t)$ and $S_x(t)$ respectively and satisfy $F_x(t) = P[T \leq x + t/T > x]$ and $S_x(t) = P[T > x + t|T > x]$. It can be shown that the force of mortality based on T_x, given by

$$\lim_{dt\to 0} \frac{1}{dt} P[t < T_x \leq t + dt/T_x > t] = \mu_{x+t} \tag{2.4}$$

$F_x(t)$ can be represented as ${}_t q_x$ and $S_x(t)$ ${}_t p_x$. The force of mortality can be described through the approximate relationship valid for small change in t (dt):

$${}_{dt}q_x \approx \mu_x dt \tag{2.5}$$

a function g(t) is defined to be O(t) (little-oh-of-t) if: $\lim_{t\to 0} \frac{g(t)}{t} = 0$. This is true since g(t) tends to zero faster than t. The researcher can show from the definition μ_{x+t} that:

$${}_{dt}q_x = \mu_x dt + g(dt) \tag{2.6}$$

where g(t) is a certain function o(t) the researcher can write it as follows

$${}_{dt}q_x = \mu_x dt + o(dt) \tag{2.7}$$

since the precision of g(t) is ignorable. The integrated hazard often arise in survival analysis it is defined by:

$$\Delta_t = \int_0^t \mu_s ds \tag{2.8}$$

The probability density function (pdf) of T_x, denoted by $f_x(t)$ is give in terms of the force of mortality

$$\begin{aligned} f_x(t) &= \frac{\partial}{\partial t} F_x \partial(t) \\ &= \lim_{dt \to 0} \frac{F_x(t+dt) - F_x(t)}{dt} \\ &= \frac{S(x+t)}{S(x)} \times \lim_{dt \to 0} \frac{F_x(t+dt) - F_x(t)}{dt} \\ &= {}_{\mathrm{t}}\, p_{\mathrm{x}} \mu_{x+t} \end{aligned} \tag{2.9}$$

Since $F_x(t) = 1 -_{\mathrm{t}} p_{\mathrm{x}}$ the researcher can write this as

$$\frac{\partial}{\partial t}{}_{\mathrm{t}} p_{\mathrm{x}} = \mu_{x+t} \tag{2.10}$$

If the researcher integrate equation 2.10 with boundary condition ${}_0p_x$ = 1 gives the formula

$${}_{\mathrm{t}}p_{\mathrm{x}} = exp\left(-\int_0^t \mu_{x+s} ds \right) \tag{2.11}$$

[43].

The two state model - probabilities

The researcher have specified the two state model in Figure 2.3 in terms of a transitions, the researcher can show how to calculate it's probabilities.

Let ${}_{\mathrm{dt}}q_{\mathrm{x+t}} + (dt)$, be the survival probability, conditioned on the state occupied at age x + t [43]. By the Markov assumption, nothing can affect the chance of survival or death after age $x + t$:

$$
\begin{aligned}
{}_{\text{dt+t}}q_{\text{x+t}} + (dt) &= {}_{\text{t}}\, p_{\text{x}} \times P[\text{ Alive at } x+t+dt/\text{Alive at} x+t] \\
&+ {}_{\text{t}}q_{\text{x}} \times P[\text{ Alive at } x+t+dt/\text{Dead at} x+t] \\
&= {}_{\text{dt}}\, P_{\text{x}} \times {}_{\text{dt}}\, P_{\text{x+t}} \times 0 \\
&= {}_{\text{t}}\, P_{\text{x}}(1-\mu_{x+t}dt)
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\frac{\partial}{\partial t}{}_{\text{t}}p_{\text{x}} &= \lim_{dt\to 0^+} \frac{{}_{\text{x+t}}p_{\text{x}} - {}_{\text{t}}\, p_{\text{x}}}{dt} \\
&= -{}_{\text{t}}p_{\text{x}}\mu_{x+t} + \lim_{dt\to 0^+} \frac{{}_{\text{t}}p_{\text{x}}}{dt} \\
&= -{}_{\text{t}}p_{\text{x}}\mu_{x+t}
\end{aligned}
\tag{2.12}
$$

In the Markov models, equation (2.12) is part of a family of Kolmogorov forward equations. Now consider observing N individuals during a given time, for example between x and $x+1$ and that the data are analysed retrospectively hence N is not a random variable [43]. The researcher also does not assume that individuals are observed until death and that all lives are independent [43].

Now assume type 1 censoring (defined in definitions section) for i = 1,...,N let $x + a_i$ to be the time an i^{th} individual begins to be observed, and let $x + b_i$ be the age at which observation of the i^{th} individual stop given that he/she has survived to this age *x + 1*, is the age of an individual or the age at which the observations end of the i^{th} individual in a study[43]. For this model to work the researcher need to know a_i and b_i in advance. Now define

$$
D_i = \begin{cases} 1 = \text{if an } i^{th} \text{ individual dies} \\ 0 = \text{if an } i^{th} \text{ individual's death in not observed.} \end{cases}
$$

This indicates the occurrence of death and define a random variable T_i as follows:

$x + T_i$ = the age at which observation of the i^{th} life ends. Notice that D, and T_t are not independent, since:

$$D_i = 0 \Longleftrightarrow T_i = b_i, D_i = 1 \Longleftrightarrow a_i < T_i < b_i \tag{2.13}$$

It is usual when analysing this type of data to use time spent under observation, so define $V_i = T_i - a_i$ V_i is called the waiting time. It contain a mixed probability distribution, with a mass function at $b_i - a_i$.
The pair (D_i, V_i) consists of a statistic, which means that the observed outcome of observation is a sample (d_i, v_i) sampled from the probability distribution $f_i(d_i, v_i)$, the joint distribution of (D_i, V_i). It is usually given in terms of $D_i = 0$ and $D_i = 1$

$$f_i(d_i, v_i) = \begin{cases} b_i - a_\mathsf{i} P_\mathsf{x} + a_i & (d_i = 0) \\ v_\mathsf{i} P_\mathsf{x} + a_\mathsf{i} \mu_\mathsf{x} + a_i + v_i & (d_i = 1) \end{cases}$$

$$= \begin{cases} exp(-\int_{i=0}^{b_i + a_i} \mu_x + a_i + tdt) & (d_i = 0) \\ exp(-\int_0^{v_i} \mu_i + a_i + tdt)\mu_x + a_i + v_i & (d_i = 1) \end{cases}$$

$$= exp(-\int_0^{v_i} \mu_x + a_i + tdt)\mu_x^{d_i} + a_i + v_i \tag{2.14}$$

Now lets assume that constant μ_{x+t} does not vary μ for 0≤t<1 and $f_i(d_i, v_i)$ takes on the simple form

$$f_i(d_i, v_i) = e^{\mu v} \mu^{d_i} \tag{2.15}$$

The joint probability distribution function of (D_i, V_i) is given by (consider the independent assumption): [43]

$$\prod_{i=1}^{i=N} e^{\mu v_i}\mu^{d_i} = e^{\mu(v_1+\ldots+v_N)}\mu^{d_1+\ldots+d_N} e^{\mu v}\mu^{d}$$
$$= e^{\mu v}\mu^{d}$$

Define a random variable D to be all deaths and V to be the waiting in time and the joint probability density function of D_i and V_i can be written in terms of **D** and **V**

Maximum Likelihood Estimator

The maximum likelihood estimator(MLE) for μ is give by

$$L(\mu; d, v) = e^{\mu v}\mu^{d} \tag{2.16}$$

which yield the MLE for μ:

$$\hat{\mu} = \frac{d}{v} \tag{2.17}$$

Since estimator $\hat{\mu}$ is an estimator of d and v and the moments of 2.17 are given by

$$E[D_i - \mu V_i] = 0 \tag{2.18}$$

$$Var[D_i - \mu V_i] = E[D_i] \tag{2.19}$$

Note equation 2.19 can be expressed as $E[D_i] = \mu[V_i]$. Since a_i and b_i are then if integrate $(D_i$,$V_i)$ over all possible events to obtain

$$\int_0^{b_i-a_i} e^{\mu v_i}\mu dv_i + e^{\mu(b_i+a_i)} = 1 \tag{2.20}$$

Note if the researcher differentiate equation 2.19 with respect to μ equation 2.20 is obtained differentiate again equation 2.21 is obtained. Now under type 1 censoring equation 2.22 can be replaced with equation 2.23

$$\int_0^1 \left(\int_{e_i}^1 e^{\mu c_i - e_i} \mu f_{C/E}(c_i/e_i) d_{c_i} \right) f_E(e_i) de_i = 0 \tag{2.21}$$

To find the distribution of $\hat{\mu}$

$$\frac{1}{N}(D - \mu V) = \frac{1}{N} \sum_0^N \left(D_i - \mu V_i \right) \tag{2.22}$$

It is sensible to assume that when N $\rightarrow \infty$ then the frequencies of $\{a_i\}$ and $\{b_i\}$ converges to a certain distribution.
Then note that

$$\lim_{N \to \infty} (\hat{\mu} - \mu) = \lim_{N \to \infty} \frac{N}{V} \left(\frac{D}{V} - \frac{\mu V}{N} \right) \tag{2.23}$$

Now considering the law of large numbers $\frac{V}{N} \rightarrow E[V_i]$ by the central limit theorem and in probability:

$$\frac{1}{N}\left(D - \mu V \right) \sim Normal\left(0; \frac{E(D)}{N^2} \right) \tag{2.24}$$

so asymptotically

$$\hat{\mu} \sim Normal\left(\mu; \frac{\mu}{E[V]} \right) \tag{2.25}$$

Simulated studies have shown that the asymptotic distribution provide reliable results if *E[D]*>10

General Markov multi-state transition model

The two state Markov model can be expressed in general to more arbitrary states. Let g and h represents any two states [43]. If state g and h are recognised differently, let μ^{gh}_{x+t} be transition of moving from g to h at age x+t:
Now let

$$P[\frac{\text{In state}_{\text{t}}p}{\text{x}}] = P[\frac{\text{In state h at age x + t}}{\text{In state g at age x}}] \tag{2.26}$$

where there is no need for g and h to be distinct. if g = h, there is no need for an individual to remain in state g. So for any state

$${}_{\text{t}}p^{\bar{g}g}_{\text{x}} = P[\frac{\text{In state g from age x to x+t}}{\text{in state g to age x}}] \tag{2.27}$$

Suppose returning to state g is impossible ${}_{\text{t}}p^{\bar{g}g}_{\text{x}} =_{\text{t}} p^{gg}_{\text{x}}$,now by the Markov property for any two states g and h ${}_{\text{dt}}p_{\text{x+t}}dt = \mu^{gh}_{x+t}dt + 0(dt)(t \geq 0)$ and the chance that a life makes more than one transition at time dt is o(dt) [43]. If the researcher still consider that the transition from g to h is impossible then the researcher can derive the forward Kolmogorov equation

$$\frac{\partial}{\partial t} {}_tP^{gh}_x = \sum_{j \neq h} \mu^{jh}_{x+t} - {}_tP^{gh}_x \mu^{hj}_{x+t} \tag{2.28}$$

$$\frac{\partial}{\partial t} {}_tP^{\hat{g}g}_x = - {}_tP^{\hat{g}g}_x \sum_{j \neq h} \mu^{gj}_{x+t} \tag{2.29}$$

solving equation 2.31 the following is obtained

$$\frac{\partial}{\partial t} {}_tP^{\hat{g}g}_x = exp\left(\int_0^t \sum_{j \neq g} \mu^{gj}_{x+s} ds \right) \tag{2.30}$$

Now consider the three state model that include alive, ill and death.

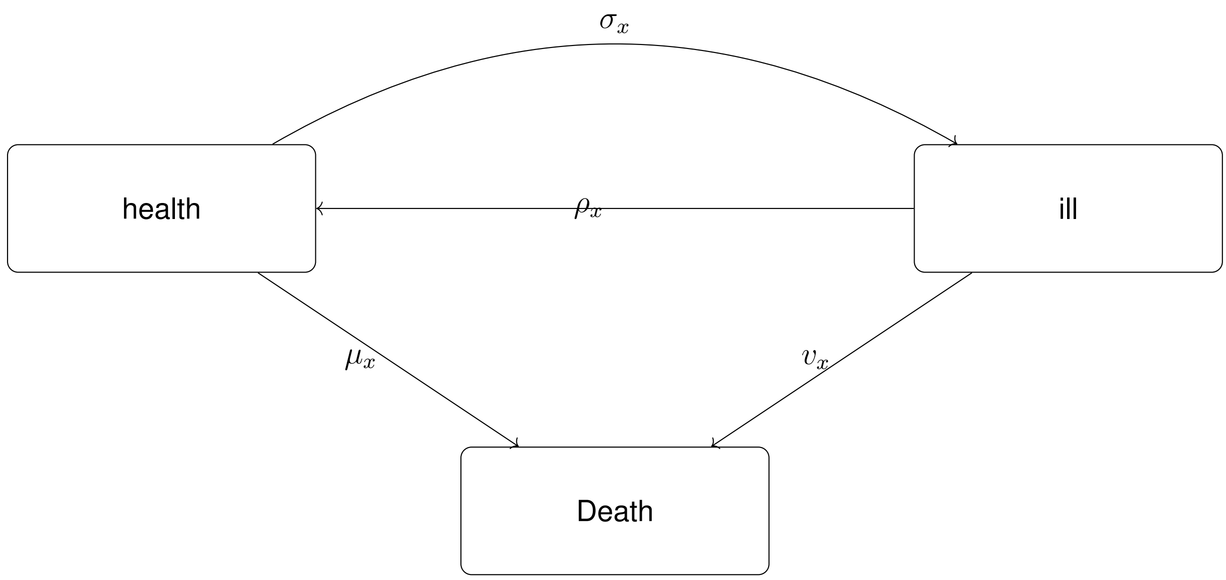


Figure 2.3: The diagrammatic presentation of the three state Illness-death Markov model.

If the transition intensities are not varying, equation (2.31) shows that each length t in the health state contributes a factor of $e^{(\mu+\sigma)t}$ or $e^{-(\mu+\sigma)t}$ respectively, to the maximum likelihood, so it is enough to record time spent waiting in each state.

V_i = Is time spent in the i^{th} health state

W_i = Is time spent in the i^th ill state

S_i = Total number of individuals moving from health $\rightarrow$ ill by the i^{th} life

R_i = Total number of transitions from ill $\rightarrow$ health by the i^{th} life

D_i = Total number of transitions from health $\rightarrow$ death

U^{th} = Total number of transitions from ill $\rightarrow$ dead

and define totals $V = \sum_1^N V_i$ (and so on), it can be shown that the likelihood estimators of our four parameters μ, σ,v, ρ when data is provided, is:

$$L(\mu, v, \sigma, \rho) = e^{(\mu+\sigma)v} e^{-(\mu+\sigma)w} \mu^d v^\mu \sigma s \rho^r \tag{2.31}$$

[43]. This can be factorised into the form $e^{\mu v}\mu^{d}$, so the MLEs are: $\hat{\mu} = \frac{D}{V}$; $\hat{v} = \frac{U}{W}$; $\hat{\sigma} = \frac{S}{V}$; $\hat{\rho} = \frac{R}{W}$

The asymptotic properties of these four estimators follows from equations (2.20) and (2.21), and the fact that $(D_i - \mu V_i)$ $(U_i - vW_i)$, $(S_i - aV_i,)$, $(R - \mu V_i)$ are uncorrelated therefore $E[(D_i - \mu V_i)(U_i - vW_i)] = 0$ and also the estimators are not independent U_i and D_i are both 0 or 1 but $D_iU_i \neq 1$ but there are asymptotically independent hence they have an asymptotically multivariate Normal probability distribution.

This three state Markov model can be extended to more states without death as done in this research report in section.

General Markov multi-state transition model (Application to our study)

The general Markov multi-state model is commonly used to describe disease progression over time in continuous time [44]. A widely used model is in matrix form named the transition matrix. This represents a series of more advanced stages of the disease and an 'absorbing' state if it is available and this is usually death [45]. The patient condition may get worse or recover from different stages of the disease or die at any stage of the disease [45]. Observations of staging are made for patients that will be followed over time [45].

More generally, Markov models of this nature can be fitted and allow the inclusion of many states these may include periods of treatment, competing causes of death or hospital stay [45]. The primary assumption for general Markov multi-state model is that the future of the transition process depends only on the current state, and not on the previous states [45]. This is the model used to track changes in ailments and symptomatic conditions of HIV patients in Adult Wellness study. The diagram in Figure 2.4 shows the transition rates of patients from one state to the other.

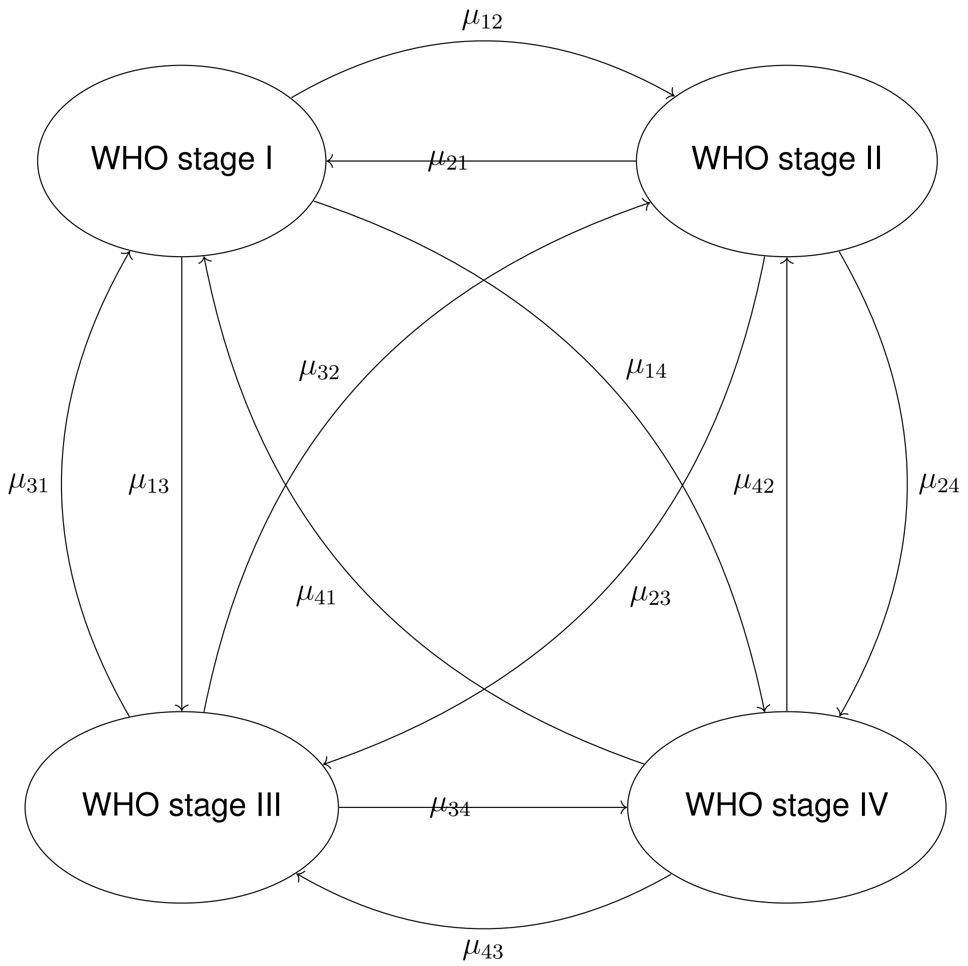


Figure 2.4: The schematic of the proposed four state multi-state Markov model with the corresponding transition intensities

where μ_{ij} for $i = 1, 2, 3, 4$ and $j = 1, 2, 3, 4$ transition rates of moving from state i to state j, $i \neq j$. Using Figure 2.4, the likelihood of the transition process and find the transition rates and transition probabilities. The likelihood of the process is given by the following formulae:

$$\begin{aligned} L(\mu_{ij}) = & \left(\mu_{12}^{n_{12}} e^{-\mu_{12}t_1} \mu_{13}^{n_{13}} e^{-\mu_{13}t_1} \mu_{14}^{n_{14}} e^{-\mu_{14}t_1}\right) \times \left(\mu_{21}^{n_{21}} e^{-\mu_{21}t_2} \mu_{23}^{n_{23}} e^{-\mu_{23}t_2} \mu_{24}^{n_{24}} e^{-\mu_{24}t_2}\right) \\ & \left(\mu_{31}^{n_{31}} e^{-\mu_{31}t_3} \mu_{32}^{n_{32}} e^{-\mu_{32}t_3} \mu_{34}^{n_{34}} e^{-\mu_{34}t_3}\right) \times \left(\mu_{41}^{n_{41}} e^{-\mu_{41}t_4} \mu_{42}^{n_{42}} e^{-\mu_{42}t_4} \mu_{43}^{n_{43}} e^{-\mu_{43}t_4}\right) \end{aligned} \tag{2.32}$$

where μ_{ij} for $i = 1, 2, 3, 4$ and $j = 1, 2, 3, 4$ is the ij^{th} rate of a patient i to state j and t_1 and t_2, t_3, t_4 are the time spent in state I, state II, state III and state IV respectively. applying natural logarithms the following is obtained

$$\begin{aligned} &= ln(\mu_{12}^{n_{12}}e^{-\mu_{12}t_1}) + ln(\mu_{13}^{n_{13}}e^{-\mu_{13}t_1}) + ln(\mu_{14}^{n_{14}}e^{-\mu_{14}t_1}) + ln(\mu_{21}^{n_{21}}e^{-\mu_{21}t_2}) + \\ &ln(\mu_{23}^{n_{23}}e^{-\mu_{23}t_2}) + ln(\mu_{24}^{n_{24}}e^{-\mu_{24}t_2}) + ln(\mu_{31}^{n_{31}}e^{-\mu_{31}t_3}) + ln(\mu_{32}^{n_{32}}e^{-\mu_{32}t_3}) + \\ &ln(\mu_{34}^{n_{34}}e^{-\mu_{34}t_3}) + ln(\mu_{41}^{n_{41}}e^{-\mu_{41}t_4}) + ln(\mu_{42}^{n_{42}}e^{-\mu_{42}t_4}) + ln(\mu_{43}^{n_{43}}e^{-\mu_{43}t_4}) + \end{aligned} \tag{2.33}$$

Now simplify 2.33

$$\begin{aligned} &= n_{12}ln(\mu_{12}) - \mu_{12}t_1 + n_{13}ln(\mu_{13}) - \mu_{13}t_1 + n_{14}ln(\mu_{14}) - \mu_{14}t_1 + n_{21}ln(\mu_{21}) - \mu_{21}t_2 + \\ &n_{23}ln(\mu_{23}) - \mu_{23}t_2 + n_{24}ln(\mu_{24}) - \mu_{24}t_2 + n_{31}ln(\mu_{31}) - \mu_{31}t_3 + n_{32}ln(\mu_{32}) - \mu_{32}t_3 + \\ &n_{34}ln(\mu_{34}) - \mu_{34}t_3 + n_{41}ln(\mu_{41}) - \mu_{41}t_4 + n_{42}ln(\mu_{42}) - \mu_{42}t_4 + n_{43}ln(\mu_{43}) - \mu_{43}t_4 \end{aligned} \tag{2.34}$$

Differentiating equation 2.34 with respect to μ_{ij} and equating to zero for maximum for each rate the following is obtained

$$\frac{\partial L(\mu_{ij})}{\partial \mu_{ij}} = \begin{bmatrix} - & \frac{\partial L(\mu_{ij})}{\partial \mu_{12}} & \frac{\partial L(\mu_{ij})}{\partial \mu_{13}} & \frac{\partial L(\mu_{ij})}{\partial \mu_{14}} \\ \frac{\partial L(\mu_{ij})}{\partial \mu_{21}} & - & \frac{\partial L(\mu_{ij})}{\partial \mu_{23}} & \frac{\partial L(\mu_{ij})}{\partial \mu_{24}} \\ \frac{\partial L(\mu_{ij})}{\partial \mu_{31}} & \frac{\partial L(\mu_{ij})}{\partial \mu_{32}} & - & \frac{\partial L(\mu_{ij})}{\partial \mu_{34}} \\ \frac{\partial L(\mu_{ij})}{\partial \mu_{41}} & \frac{\partial L(\mu_{ij})}{\partial \mu_{42}} & \frac{\partial L(\mu_{ij})}{\partial \mu_{43}} & - \end{bmatrix}$$

$$= \begin{bmatrix} - & \frac{n_{12}}{\mu_{12}} - t_1 & \frac{n_{13}}{\mu_{13}} - t_1 & \frac{n_{14}}{\mu_{14}} - t_1 \\ \frac{n_{21}}{\mu_{21}} - t_2 & - & \frac{n_{23}}{\mu_{23}} - t_2 & \frac{n_{24}}{\mu_{24}} - t_2 \\ \frac{n_{31}}{\mu_{31}} - t_3 & \frac{n_{32}}{\mu_{32}} - t_3 & - & \frac{n_{34}}{\mu_{34}} - t_3 \\ \frac{n_{41}}{\mu_{41}} - t_4 & \frac{n_{42}}{\mu_{42}} - t_4 & \frac{n_{43}}{\mu_{43}} - t_4 & - \end{bmatrix}$$

now equating to zero for for maximum and making μ_{ij} subject of formula i.e

$$\frac{\partial L(\mu_{ij})}{\partial \mu_{ij}} = \mathbf{0} \tag{2.35}$$

the following is obtained

$$\mu_{ij} = \begin{bmatrix} - & \frac{n_{12}}{t_1} & \frac{n_{13}}{t_1} & \frac{n_{14}}{t_1} \\ \frac{n_{21}}{t_2} & - & \frac{n_{23}}{t_2} & \frac{n_{24}}{t_2} \\ \frac{n_{31}}{t_3} & \frac{n_{32}}{t_3} & - & \frac{n_{31}}{t_3} \\ \frac{n_{41}}{t_4} & \frac{n_{42}}{t_4} & \frac{n_{43}}{t_4} & - \end{bmatrix}$$

where μ_{ij} $i = 1, 2, 3, 4$ and $j = 1, 2, 3, 4$ is the ij^{th} is the rate of moving from state i to state j for $i \neq j$.

Note the diagonal is empty because the patients that are in the same state will not be considered to have moved.

The following matrix represents how transition probabilities will be represented from R program.

$$\begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & P_{34} \\ P_{41} & P_{42} & P_{43} & P_{44} \end{bmatrix}$$

where P_{ij} for $i = 1, 2, 3, 4$ and $j = 1, 2, 3, 4$ is the $(ij)^{th}$ probability of moving from state i to state j and t_i are durations of staying in the same state. These probabilities can be derived from Figure (2.4) using occupancy probabilities. They give the next probability matrix of moving from one state to other (where $i \neq j$) or staying in the same state (where $i = j$).

$$\begin{bmatrix} e^{-(\mu_{12}+\mu_{13}+\mu_{14})t_1} & \frac{\mu_{12}}{A}\left(1-e^{-(\mu_{12}+\mu_{13}+\mu_{14})t_1}\right) & \frac{\mu_{13}}{A}\left(1-e^{-(\mu_{12}+\mu_{13}+\mu_{14})t_1}\right) & \frac{\mu_{14}}{A}\left(1-e^{-(\mu_{12}+\mu_{13}+\mu_{14})t_1}\right) \\ \frac{\mu_{21}}{B}\left(1-e^{-(\mu_{21}+\mu_{23}+\mu_{24})t_2}\right) & e^{-(\mu_{21}+\mu_{23}+\mu_{24})t_2} & \frac{\mu_{23}}{B}\left(1-e^{-(\mu_{21}+\mu_{23}+\mu_{24})t_2}\right) & \frac{\mu_{23}}{B}\left(1-e^{-(\mu_{21}+\mu_{23}+\mu_{24})t_2}\right) \\ \frac{\mu_{31}}{C}\left(1-e^{-(\mu_{31}+\mu_{32}+\mu_{34})t_3}\right) & \frac{\mu_{32}}{C}\left(1-e^{-(\mu_{31}+\mu_{32}+\mu_{34})t_3}\right) & 1-e^{-(\mu_{31}+\mu_{32}+\mu_{34})t_3} & \frac{\mu_{34}}{C}\left(1-e^{-(\mu_{31}+\mu_{32}+\mu_{34})t_3}\right) \\ \frac{\mu_{41}}{D}\left(1-e^{-(\mu_{41}+\mu_{42}+\mu_{43})t_4}\right) & \frac{\mu_{42}}{D}\left(1-e^{-(\mu_{41}+\mu_{42}+\mu_{43})t_4}\right) & \frac{\mu_{43}}{D}\left(1-e^{-(\mu_{41}+\mu_{42}+\mu_{43})t_4}\right) & 1-e^{-(\mu_{41}+\mu_{42}+\mu_{43})t_4} \end{bmatrix}$$

where A = $\mu_{12}+\mu_{13}+\mu_{14}$; B = $\mu_{21}+\mu_{23}+\mu_{24}$; C = $\mu_{31}+\mu_{32}+\mu_{34}$; D = $\mu_{41}+\mu_{42}+\mu_{43}$

2.7.3 Maximum likelihood random effects ordered logistic regression

In pertaining to objective two and three fitted random effects ordered logistic regression model to find factors associated with WHO stage ailments and symptomatic conditions. A stepwise forward selection method was done for choosing covariates to include in the final models fitted setting a liberal p-value at 0.1 [46]. Random effects ordered regression models were compared using standard errors and confidence intervals. In general random effects ordered logistic regression model is defined as;

$$Pr(y_{it} > k|k, x_{it}, v_i) = H(X_{ij} + v_i - k_k) \tag{2.36}$$

for i=1,...,n panels, t = 1,...,n_i and v_i are independent and identically distributed(iid) $N(0, \sigma_v^2)$, and k is a set of cutpoints $k_1, k_2, ..., k_{k-1}$. K is the count of observed outcomes; and $H(.)$ is the logistic cumulative distribution function.

From equation 2.37 the researcher can find the probability of getting outcome (k) for response y_{it} as

$$P_{itk} = Pr(y_{ij} \quad = k, |k, x_{it}, v_i) = Pr(k_{k-1} < x_{ij}\beta + v_i + \epsilon_{ij} + \leq k_k) \tag{2.37}$$

rearranging equation 2.39 the following is obtained

$$= Pr(k_{k-1} - x_{ij}\beta - v_i < \epsilon_{ij} \leq k_k - x_{ij}\beta - v_i) \tag{2.38}$$

Now splitting the equation 2.39 after recognising that the contents of 2.41 is an inequality the following is obtained.

$$= H(k_k - x_{ij}\beta - v_i) - H(k_{k-1} - x_{ij}\beta - v_i) \tag{2.39}$$

Substituting into the probability distributions the solution is

$$= \frac{1}{1 + exp(-k_k + x_{ij} + v_i)} - \frac{1}{1 + exp(-k_{k-1} + x_{ij}\beta + v_i)} \tag{2.40}$$

where k_0 is considered to be $-\infty$ and K_k is considered to be $+\infty$. from 2.41.Now writing the model in terms of a latent linear outcome, where our ordinal outcome of WHO stage ailments and symptomatic conditions y_{ij} are derived from the latent responses, such that

$$y^*_{it} = x_{it}\beta + v_i + \epsilon_{ij} \tag{2.41}$$

and

$$y_{ij} = \begin{cases} 1 & \text{if} \quad y^*_{ij} \leq k_1 \\ 2 & \text{if} \quad k_1 < y^*_{ij} < k_2 \\ . & \\ . & \\ k & \text{if} \quad k_{k-1} < y^*_{ij} \end{cases} \tag{2.42}$$

The error terms ϵ_{it} are distributed as a binary (logistic) distribution with mean 0 and standard deviation $\frac{\pi}{\sqrt{3}}$ and are independent of v_i. Given a set of random effects v_i, the conditional distribution is given by

$$f(y_{it}, k, x_{it}\beta + v_i) = \prod_{k=1}^{k} P_{itk}^{I_k(y_{it})} \tag{2.43}$$

for response y_{it}

Evaluating the product on the right side of equation 2.44 the following equation is obtained

$$= exp \sum_{k=1}^{K}(I_k(y_{it}log(p_{itk}))) \tag{2.44}$$

where

$$I_k(y_{ij}) = \begin{cases} 1 & \text{if} \quad y_{ij} = k \\ 0 & \text{otherwise} \end{cases}$$

For panel i where i=1,...,M the conditional distribution of $y_i = (y_{ij}, ..., y_{in_i})'$ is

$$\prod_{t=1}^{n_i} f(y_{it}, k, x_{it}\beta + v_i)$$

and the panel-level likelihood is given by

$$L_i(\beta, k, \sigma_v^2) = \int_{-\infty}^{+\infty} \frac{e^{v_i^2/2\sigma_v^2}}{\sqrt{2\pi}\sigma_v} \{\prod_{t=1}^{n_i} f(y_{ij}, k, x_{ij}\beta + v_i)\} dv_i \tag{2.45}$$

$$= \int_{-\infty}^{+\infty} g(y_{it}, k, x_{it}, v_i) dv_i \tag{2.46}$$

This integral equation can be estimated with M points Gauss Hermite quadrature

$$\int_{-\infty}^{+\infty} e^{-x^2} h(x) dx \approx \sum_{m=1}^{M} w_m^* h(a_m^*) \tag{2.47}$$

This is equivalent to

$$\int f(x)dx \approx \sum_{m=1}^{M} exp\{(a_m^*)^2\}f(a_m^*) \tag{2.48}$$

where the w_m^* represents the weights of the quadrature and a_m^* represents the abscissas of the quadrature. The log-likelihood, L, is the summation of all the logs of the longitudinal level likelihood L_i The estimation for the log-likelihood is by mean-variance adaptive Gauss-Hermite quadrature, which estimates the panel-level likelihood using

$$L_i \approx \sqrt{2}\hat{\sigma}_i \sum_{m=1}^{M} w_m^* exp\{(a_m^*)^2\}g(y_{it}, k, x_{it}, \sqrt{2}\hat{\sigma}_i a_m^* + \hat{\mu}_i) \tag{2.49}$$

where $\hat{\mu}$ (mean) and $\hat{\sigma_i}$ (standard deviation) are estimates for panel i [47, 48]. Iterations are used to estimate these parameters. The researcher use the following equation to do so

$$L_{i,j} \approx \sum_{m=1}^{M} \sqrt{2}\hat{\sigma}_{i,j} w_m^* exp\{(a_m^*)^2\}g(y_{it}, k, x_{it}, \sqrt{2}\hat{\sigma}_{i,j-1} a_m^* + \mu_{i,\hat{j}-1}) \tag{2.50}$$

Letting

$$\tau_{i,m,j-1} = \sqrt{2}\sigma_{i,\hat{j}-1} a_m^* + \mu_{i,\hat{j}-1} \tag{2.51}$$

$$\hat{\mu}_{i,j} = \sum_{m=1}^{M} (\tau_{i,m,j-1}) \frac{\sqrt{2}\sigma_{i,\hat{j}-1} w_m^* exp\{(a_m^*)^2\}g(y_{it}, k, x_{it}, \tau_{i,m,j-1})}{l_{i,j}} \tag{2.52}$$

and

$$\hat{\sigma}_{ij} = \sum_{m=1}^{M} (\tau_i, m, j-1)^2 \frac{\sqrt{2}\sigma_{ij-1} w_m^* exp\{(a_m^*)^2\}g(y_{it}, k, x_{it}, \tau_i)}{l_{i,j}} - (\hat{\mu}_{i,j})^2 \tag{2.53}$$

This is repeated until $\hat{\mu}_{i,j}$ and $\hat{\sigma}_{i,j}$ converges and applied on every iteration. The log likelihood can be calculated by a non adaptive Gauss-Hermite quadrature with the option, where $\rho = \frac{\sigma_v^2}{\sigma_v^2+1}$:

$$L = \sum_{i=1}^{n} w_i log\{P(y_{i1}, ..., y_{in_i}|k, x_{i1,...,x_{in_i}})\} \tag{2.54}$$

$$\approx \sum_{i=1}^{n} w_i log\left[\left\{\frac{1}{\sqrt{\pi}}\sum_{m=1}^{n} w_m^* \prod_{m}^{n_i} f(y_{it}, k, x_{it}\beta + a_m^*\left(\frac{2\rho}{1-\rho}\right)^{1/2}\right\}\right] \tag{2.55}$$

These two quadrature formulas need the integrated function to be well estimated by a polynomial of equal degree to quadrature points. Panel size has an effect on

$$\prod_{t=1}^{n_i} f(y_{it}, k, x_{it}\beta + v_j) \tag{2.56}$$

how it is approximated by a polynomial. As ρ and panel size increase, the quadrature approximation can be less accurate. As ρ increases, the random effects ordered logistic regression models can also become to small to be identified. Adaptive quadrature give more accurate results for correlated data and large panels compared to non adaptive quadrature [47, 48].

2.7.4 Rubin's method of missing longitudinal data

Theoretical background of Rubin's method

The first method which will be used in increasing completeness is Rubin's Method developed by Donald B Rubin in 1987 [27]. This method is a maximum likelihood technique which means it estimates the parameters of a statistical model using given data [29, 27]. The first stage of Rubin's method is to impute missing data say m times. These data will be drawn from a particular probability distribution, for example, normal distribution or binomial. Then the researcher estimate the parameters from these datasets to get m sets of estimated parameters [49]. The final stage is to produce a single set of estimated parameters by pooling (combining estimated parameters using a certain scale) [15]. The strength of this method is that it is capable of modelling both the outcome and covariates of data that is missing [40].

Usually three to five imputations yield the same results [50, 51]. For example, if five imputations are done to the dataset it can be shown in the following diagram:

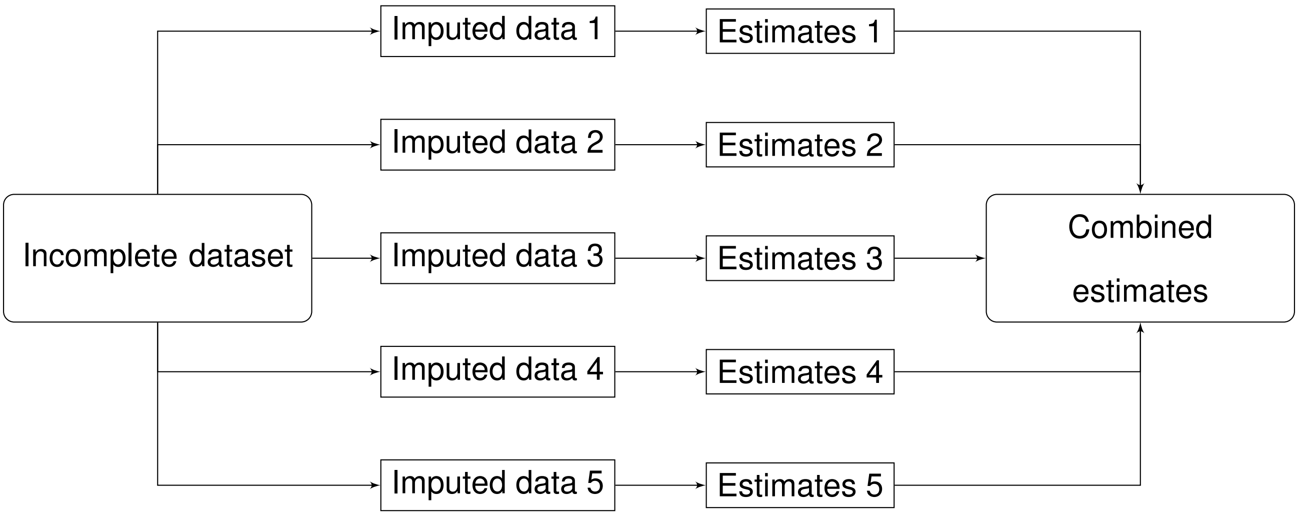


Figure 2.5: Schematic presentation of the multiple imputation chained equations (MICE) adopting Rubin's method

Application of Rubin's method in our study

The first imputation method used is called Rubin's method or multiple imputations. It has the following algorithm. With m = 5 imputations, computed 5 different sets of the point and variance estimates for a parameter q. Then our point estimate for q obtained from multiple imputations is the mean of the 5 complete data estimates [52, 53]. The MI of q (where $\{q_i, u_i : i = 1, 2, \ldots, 5\}$) estimates is

$$q_m = \frac{1}{m}\sum_{i=1}^{m} q_i \tag{2.57}$$

the variance-covariance estimate (VCE) of q_m(total variance) is $T = \bar{U} + (1 + \frac{1}{m})\beta$ Where $\bar{U} = \frac{\sum_{i=1}^{m}(q_i - q_m)(q_i - q_m)'}{m-1}$ is the between-imputation variance-covariance matrix [54, 55, 56].

The formula 2.58 for multiple imputations is not used on a single variable but rather

as a regression model. This is where the variable which is being attributed is considered as an outcome, and others variables that are awaiting imputation are regarded as independent variables. Then the researcher imputed the dependent variable using formulae 2.58 [32, 57]. This process is called multiple imputations chained equations (MICE) [58, 59, 60]. At each stage of the algorithm, an imputation was produced for the missing variable, and then this imputed value was used to generate imputation of the following variable [61]. This process was repeated using Gibbs sampling procedure until the process reaches convergence [61]. Different chains are used to generate multiple imputations from each variable [61].
For continuous variables, the researcher used linear regression model and other variables as predictors [61]. For categorical variables, logistic regression, ordinal or multinomial was fitted using polytomous models [61].

2.7.5 Bayesian method for missing longitudinal data

The second method used in this study is called the Bayesian imputation method. This method involves specifying probability distributions for all parameters as well as assigning statistical distributions for the missing data based on the observed [62]. These distributions are called prior distributions. The missing data are then imputed from these distributions [62]. The Bayesian method is perhaps more flexible (since it allows us to specify priors) and robust for modelling data which are not complete [62].

Full Bayesian missing data model

The Bayesian missing data model used to fill in missing data was defined as follows: Posterior[p(parameters/data)] = Likelihood $\times$ Priors. The prior distributions of parameters are normal i.e

$$g(\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(\beta-\mu)^2/2\sigma^2} \tag{2.58}$$

Where μ and σ^2 are the mean and variance of the prior distribution respectively. β is a vector of coefficients and the likelihood is the same as the one in section 2.74

$$L(Y|X,\beta,\sigma^2) = \sum_{i=1}^{n} w_i log\left[\left\{\frac{1}{\sqrt{\pi}}\sum_{m=1}^{n} w_m^* \prod_{m}^{n_i} f(y_{it},k,x_{it}\beta + a_m^*\left(\frac{2\rho}{1-\rho}\right)^{1/2}\right\}\right] \quad (2.59)$$

, hence our full Bayesian model is defined as follows:

$$\text{Posterior[p(parameters/data)]} =$$

$$g(\beta) \times L(Y|X,\beta,\sigma^2) = \sum_{i=1}^{n} w_i log\left[\left\{\frac{1}{\sqrt{\pi}}\sum_{m=1}^{n} w_m^* \prod_{m}^{n_i} f(y_{it},k,x_{it}\beta + a_m^*\left(\frac{2\rho}{1-\rho}\right)\right\}\right] \quad (2.60)$$
$$\times \frac{1}{\sqrt{2\pi\sigma^2}} e^{(\mathbf{B}-\mu)^2/2\sigma^2}$$

[63, 64] This equation can not be solved analytically hence Metropolis hustling algorithm was used to find the value of posterior parameters.

Bayesian random effects ordered logistic regression model

In the Bayesian model instead of imputing the dataset, assigned normal prior distributions to the random effects ordered logistic regression model covariates parameters in section (2.8.11). Metropolis hustling algorithm was used to estimate the covariates. A total of 7500 Markov multi-state Monte Carlo (MCMC) sample were done with a thining of 7 to reduce autocorrelation between adjacent samples. To increase efficiency and convergence used blocking with a default size of 50.

The normal distribution used for prior distribution is given by the following formulae [65].

$$\beta_i \sim N(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{(x-\mu)^2/2\sigma^2} \quad (2.61)$$

were μ is the variance of the normal distribution and σ_2 is the variance of the normal distribution, where μ = 0 and σ_2 = 10000.

2.7.6 Simulation data models

For objective four the researcher performed a sensitivity analysis comparing Bayesian and Rubin's missing data methods on a simulated dataset. The simulated data constitute 15000 observations and 999 unique individuals. The variables are as follows d-is a continuous variable, W-is a binary variable and m is a multinomial variable and the outcome is ordinal. The random effects ordered logistic regression model similar to those fitted in the Adult Wellness data where fitted again using the simulated data. The models were adjusted for continuous, nominal and binary variables as well. Two different datasets were simulated with different level of missingness, one with 25% of missing values and the other with 50% of missing values. Comparison of there simulated data models estimates were done.

2.7.7 Goodness of fit techniques

For random effects ordered regression models (Adult Wellness data models) in using Rubin's method used the Wald test to check the significance of each variable and on Bayesian approach used Bayesian information criterion to choose the best model among nested ones [15, 66].

2.8 Ethical considerations

To protect the identities of HIV infected participants, the datasets used unique study identifications instead of patient names. The dataset were stored in student's personal laptop which has a password to restrict access. Ethical approval was sought and was granted by the University of Witwatersrand ethics committee. The clearance certificate M180335, approved on the 6th of April 2018, See attached copy of certificate in Appendix F.1.

Chapter 3

Results

Introduction

This chapter show results which are baseline characteristics and summaries of data over time were presented. Results for the general Markov multi-state model as well as those from the random effects ordinal logistic regression models adjusting for individual random effects for raw data MLE, raw data BE and MI data MLE are presented.

3.1 Exploratory Analysis

3.1.1 Reasons for missingness

The loss to follow-up patterns in the study were relatively pronounced. The study recruited a total of 2609 participants who were entering at different time points over the 8 year period. The maximum time patients were followed was 5 years. Of these recruits, 80.61% (n=2103) completed one year, 57% (n=1507) completed two years, 39.13% (n=1021) completed three years, 22.27% (n=581) completed four years and 6.32% (n=165) completed five years with reference to the total number recuited in the study. A total of 662 participants were lost to follow up (LTFU), these accounted for 25.37% of the total study participants recruited. Majority of the patients who became LTFU could not be traced. Of the recruited 2609, 1414 (54.2%) were missing. Missing data was observed as a result of the LTFU participants and those who were missing the scheduled visits which were set in the study. The other reasons for missing documented were dropped-out, opted out and withdrawals.

3.1.2 Baseline characteristics

The descriptive baseline characteristics of the participants are shown in table 3.1. The overall mean age of HIV patients in the study was 33.5 (C.I 29.1-39.5) whilst the mean age by gender was 36.81 (C.I 35.5-43.6) and 32.4 (C.I 28.1-38.3) for male and female respectively. Majority of the patients were drinking 8 and above bottles of alcohol per week (73.7%) of which most of them were females. Majority of the participants in the study had not used condom with any of their last four sexual partners (60.1%, n=1372), had between 4-6 sexual partners (94.8%, n=2166), did not know HIV status of their partners (58.6%, n=255) and had attained school grade 0-12 (77.9%) . The majority of HIV patients in the study were non-smokers (78.5%, n=1140), unemployed (95.6%) aged above 25 years (93.4%) and single (69.1%).

Table 3.1: Descriptive baseline demographic characteristics of HIV infected patients in the Adult Wellness study, 2002-2010

Variable	**Male(n = 517, 22.6)**	**Female(n = 1767, 77.4)**	**overall(2284)**
mean age (C.I) in years	36.81(32.5−43.6)	32.4(28.1−38.3)	33.5(29.1−39.5)
Alcohol Drinks/week			
None	1(0.2)	2(0.1)	3(0.1)
1 to 7	250(48.4)	348(19.7%)	598(26.2)
>= 8	266(51.5)	1,417(83%)	1683(73.7)
Condom			
No	300(58)	1072(60.7)	1372(60.1)
Yes	217(42)	695(39.3)	912(39.9)
Sexual Partners			
1 t o 3	5(1.0)	47(2.7)	52(2.3)
4 to 6	505(97.7)	1661(94)	2166(94.8)
>= 7	7(1.3)	59(3.3)	66(2.9)
HIV partner status			
No	31(50.8)	96(25.7)	127(29.2)
Yes	6(9.8)	47(12.6)	53(12.2)
Unknown	24(39.3)	231(61.7)	255(58.6)
Education level			
None	39(13.3)	87(9.0)	126(9.9)
Grade 0 − 12	227(77.2)	756(78.0)	983(77.9)
Tertiary	28(9.5)	126(13)	154(12.2)
Smoking status			
No	189(37.7)	1512(87.2)	1701(76.1)
Yes	312(62.3)	222(12.8)	534(23.9)
Patient receiving salary			
No	267(71.8)	873(80.8)	1140(78.5)
Yes	105(28.2)	208(19.2)	313(21.5)
Self employed			
No	341(93.9)	1020(96.2)	1361(95.6)
Yes	22(6.1)	40(3.8)	62(4.4)
Sex age Category			
<16	4(0.8)	11(0.6)	15(0.7)
16−20	14(2.7)	102(5.8)	116(5.1)
20−25	3(0.6)	16(0.9)	19(0.8)
>25	496(95.9)	1638(92.7)	2134(93.4)
Marital Status			
Single	279(53.9)	1298(73.5)	1577(69.1)
Livingtogether/Maried	189(36.6)	293(16.5)	482(21.1)
Divorced	25(4.8)	72(4.1)	97(4.3)
Widowed	24(4.6)	104(5.9)	128(5.6)

Table 3.2 shows the baseline clinical characteristics of the patients. Generally, one third of the patients had at least one ailment and symptomatic condition. Few were on ART (n=744, 38.55%) at baseline and taking CXT treatment (46.6%). Majority of the patients were in good clinical state with CD4 cell count of above 500 cell/uL of blood. The proportion with sexually transmitted infections was 21.1% (n=390) and more than half of the patients were obese.

Table 3.2: Descriptive baseline clinical characteristics of HIV infected patients in the Adult Wellness study, 2002-2010

Variable	**Male(n = 517, 22.6)**	**Female(n = 1767, 77.4)**	**overall(n = 2284)**
Number of Ailments and symptomatic conditions			
0	341(65.9)	1142(64.6)	1483(64.9)
1	79(15.3)	306(17.3)	385(16.9
2	50(9.7)	170(9.62)	220(9.6)
>=3	47(9.1)	149(8.4)	196(8.6)
Patient on ART			
No	241(55.1)	945	1186(61.5)
Yes	197(44.9)	547	744(38.55)
Patient on cotrimoxazole			
No	179(40.6)	861(57.2)	1040(53.4)
Yes	262(59.4)	645(42.8)	907(46.6)
CD4 category			
<=200	3(0.6)	8(0.5)	11(0.5)
201 to 350	2(0.4)	17(1)	19(0.8)
351 to 500	3(0.6)	14(0.8)	17(0.7)
>500	509(98.4)	1728(97.8)	2237(98)
Blood pressure systolic			
Normal	8(1.6)	17(1)	25(1.1)
High	509(98.4)	1750(99)	2259(98.1)
Blood pressure diastolic			
Low	300(58)	1056(59.8)	1356(59.4)
Normal	16(3.1)	44(2.4)	60(2.6)
High	201(38.9)	667(37.8)	868(38)
Sexual transmitted infections			
No	338(80.1)	1119(78.5)	1457(78.9)
Yes	84(19.9)	306(21.5)	390(21.1)
Body Mass Index			
Underweight	187(36.2)	450(25.5)	637(27.9)
Healthy weight	69(13.4)	305(17.3)	374(16.4)
Overweight	13(2.5)	51(2.9)	64(2.8)
Obese	248(47.9)	961(54.4)	1209(52.9)

3.1.3 Time varying patient characteristics

The person-years, estimated using the age showed that the between (which is the average age a person was observed in years during follow-up) ranges from 18.79 to 61.63 years with a standard deviation of 7.8 years. The within number ranges from 12.20 to 61.3 years. The overall part in the table summarises results in terms of person-years. Between repeats the results but this time in terms of patients and within gives us a fraction a patient that had a specific condition of the variable of interest. During follow-up, 2044 (80.7%) patients spent more time (7207 person-years) without using condoms and 2571 patients spent 11288 person-years with 4 to 6 sexual partners. A total of 389 patients (96.9%) never new the status of their sexual partners, 233 patients ((59.9%) who had attained high school level accounted for 4573 person-years of follow-up. The majority of patients were non-smokers (n= 2044,(80.7%) and accounted for 8221 person years of follow-up. Only 557(26.6%) had salaries accounted for 285 person-years of follow-up. The majority of patients who had sex debut of after 25 years accounted for 11063 person-years of follow-up while 2225(85.3%) patients who were not married accounted for 8967 follow-up years.

Table 3.3: The number of times a certain demographic characteristic occurred during the follow up period amongst HIV infected patients in the Adult Wellness study

	N(Total)	Overall	Between	Within
		Person-years(Percentage)	**Person-years(Percentage)**	**Fraction of the time**
Estimated Age				
mean(SD)	12102	36.5(8.2)	36.5(7.8)	36.5(2.8)
Min-Max		16.0- 62.0	18.8- 61.6	12.2 - 61.3
Alcohol Drinks/week				
None	12102	17(0.1)	14(0.5)	20.6
1 to 7		2911(24.1)	838(32.1)	79.2
>= 8		9174(78.8)	2109(80.8)	92.1
Condom				
No	12102	7207(59.6)	2044(78.3)	75
Yes		4895(4)	1660(63.6)	64.9
Sexual Partners				
1 − 3	12102	405(3.4)	209(8)	43.1
4 − 6		11288(93.3)	2571(98.5)	94.5
>= 7		409(3.38)	222(8.51)	39.7
Partner status				
No	3377	959(28.4)	202(31.1)	96.3
Yes		449(13.3)	85(13.1)	92.2
Unknown		1969(58.3)	389(59.9)	96.9
Education level				
None	5671	590(10.4)	226(12.2)	78.1
Grade 0 - 12		4573(80.6)	1505(81.5)	96.8
Tertiary		508(9)	233(12.6)	91.6
Smoking				
No	10649	8221(77.2)	2044(80.7)	94.6
Yes		2428(22.8)	705(27.8)	85.1
Patient receiving salary				
No	6071	4715(77.7)	1724(82.3)	94.7
Yes		1356(22.3)	557(26.6)	83.6
Self employed				
No	5910	5625(95.2)	2004(97.1	98.8
Yes		285(4.8)	104(5.0)	82.1
Sex age Category				
<16	12102	157(1.3)	82(3.1)	43.6
16-20		707(5.8)	362(13.9)	42.7
20-25		175(1.5)	96(3.7)	36.8
>25		11063(91.4)	2564(98.3)	92.9
Marital Status				
Single	12102	8967(74.1)	2225(85.3)	85.1
Livingtogether/Maried		1990(16.4)	783(30)	60.5
Divorced		502(4.2)	171(6.6)	62.7
Widowed		643(5.3)	235(9)	57.5

Table 3.4 displays the number of times a certain ailments and symptomatic conditions occurred during the follow-up period amongst HIV infected patients in the Adult Wellness study. Non-development of ailments or symptomatic conditions was observed 2559 times (98.1%) of the total visits made. These patients accounted for 10526 person-years of follow-up. Patients not receiving ART treatment were observed 1953 times accounting for 6203 person-years of follow-up in the study while patients not taking cotrimoxazole contributed 1795 person-years and were observed in 76.4% of total

times. Patients with CD4 count greater than 500 were observed for 2568(98.4%) times and accounted for 11717 person-years. Patients who had never had sexually transmitted infections were observed 2317 times and accounted for 7595 person-years whilst patients who were obese were observed 2413 (79.1%) times and accounted for 8784 person-years of follow-up.

Table 3.4: The number of times a certain ailments and symptomatic conditions occurred during the follow up period amongst HIV infected patients in the Adult Wellness study

Variable	N(Total)	Overall	Between	Within
		Freq(Percentage)	Freq(Percentage)	Percentage
Number of ailments and symptomatic conditions				
0	12102	10526(88.9)	2559(98.1)	87.93
1		806(6.7)	651(24.9)	27.96
2		449(3.7)	414(15.9)	24.82
>=3		321(2.6)	307(11.8)	24.13
Patient on ART				
No	10182	6203(60.9)	1953(79.1)	80.7
Yes		3979(39.1)	1250(50.6)	71.5
Patient on cotrimoxazole				
No	10666	5853(54.9)	1795(71.9)	76.4
Yes		4813(45.1)	1557(62.4)	72.2
CD4 category				
<=200	12102	74(0.6)	37(1.4)	36.6
201 to 350		181(1.5)	66(2.5)	45.7
351 to 500		130(1.1)	46(1.8)	50
>500		11717(96.8)	2568(98.4)	99
Blood pressure systolic				
Low	12102	16(0.2)	5(0.2)	36.4
Normal		116(0.9)	70(2.7)	33.9
High		11970(98.9)	2602(99.8)	99.3
Blood pressure diastolic				
Low	12102	7407(61.2)	2146(82.3)	74.69
Normal		261(2.2)	132(5.1)	42.61
High		4434(36.6)	1583(60.7)	60.01
STI				
No	9627	7595(78.9)	2317(91.0)	84.7
Yes		2032(21.1)	1111(43.6)	52.5
Body Mass Index				
Underweight	12102	1951(16.1)	934(35.8)	44.7
Healthy weight		1165(9.6)	663(25.4)	35.9
Overweight		202(1.7)	137(5.3)	33.6
Obese		8784(72.6)	2413(92.5)	79.1

3.2 General Markov multi-state model of Adult Wellness data

Table 3.5 shows the number of the observed transitions between the defined WHO states from state i (row) to state j (column), where i = j shows the transitions of staying in the same state. Generally patients were becoming worse in this cohorts as evident by the bigger number of transitions made in the forward direction, that is, (454 + 356+ 116=926) compared to backward (recovery) transitions made (401+ 254+70=716) during the follow-up period. There were more transitions observed for staying in the same state over time, that is staying in WHO stage 1 a total of 230 transitions were observed. Similarly for the other states, WHO II (n=1021), WHO III (n=3124) and WHO IV (n=281).

Table 3.5: Frequency of the observed transitions between the four defined WHO staged ailments and symptomatic conditions among the HIV patients in Adult Wellness study.

	To			
From	**WHO I**	**WHO II**	**WHO III**	**WHO IV**
WHO I	2390	454	511	54
WHO II	401	1021	356	28
WHO III	382	245	3124	116
WHO IV	29	24	70	281

3.2.1 Transition rates

The results for the transition intensities are presented in table 3.6. Patients in WHO stage II were 1.33 times (0.66329/0.49619) more likely to recover than getting worse whilst patients in WHO stage III were 2.26 times (0.16118/0.07124) more likely to transition to state II than getting worse. Patients in WHO stage IV were 3.51 times (0.42440/0.12095) more likely to recover to WHO stage III than getting to WHO stage II directly.

Table 3.6: Estimates of the transition intensities and their 95% confidence intervals for the General multi-state Markov Model

Transitions	**Rates(95%C.I)**
WHO I → WHO I	-0.66488(-0.71195,-0.62092)
WHO I → WHO II	0.35696 (0.32067, 0.39735)
WHO I → WHO III	0.28352 (0.25454, 0.31580)
WHO I → WHO IV	0.02440 (0.01653, 0.03604)
WHO II → WHO I	0.66329 (0.59209, 0.74304)
WHO II → WHO II	-1.19150(-1.28825,-1.10201)
WHO II → WHO III	0.49619 (0.43888, 0.56099)
WHO II → WHO IV	0.03202 (0.01892, 0.05419)
WHO III → WHO I	0.21416 (0.19018, 0.24117)
WHO III → WHO II	0.16118 (0.13898, 0.18692)
WHO III → WHO III	-0.44658(-0.48180,-0.41393)
WHO III → WHO IV	0.07124 (0.05860, 0.08661)
WHO IV → WHO I	0.15675 (0.10040, 0.24472)
WHO IV → WHO II	0.12095 (0.07046, 0.20761)
WHO IV → WHO III	0.42440 (0.32961, 0.54646)
WHO IV → WHO IV	-0.70210(-0.83827,-0.58805)

3.2.2 Transition probabilities

The estimated transition probabilities between the four WHO states at the end of eight year period are presented in table 3.7. The researcher found that the probability of staying in the same state were higher than either making a forward or backward transition for any of the states, that is WHO stage I (0.59), WHO stage II (0.37), WHO stage III (0.70) and WHO stage IV (0.51). At the end of the 8 year period, patients in WHO stage I had a 22% chance of moving to WHO stage III whist patients in WHO stage II had a 37% chance of moving to WHO stage II. The probability of moving from WHO stage III to WHO stage I was 16% at the end of 8 years while patients in WHO stage IV had a 27% chance of making a reversal transition to WHO stage III.

Table 3.7: The estimated transition probabilities between the four WHO states at the end of eight years of HIV infected patients

	To			
From	**WHO I**	**WHO II**	**WHO III**	**WHO IV**
WHO I	0.5924047	0.16513873	0.2200398	0.02241678
WHO II	0.3120109	0.37141245	0.2898056	0.02677106
WHO III	0.1607609	0.099038	0.6955664	0.04463473
WHO IV	0.1319091	0.08263893	0.2783528	0.50709915

3.3 Random effects ordered logistic regression models

The comparative results between the different models of the random effects adjusting for the individual heterogeneity using the MLE and BE are presented in table 3.8 for both raw data and imputed data from the Adult Wellness study. The three fitted models gave varying patterns of the variable estimates.

The complete case analysis model (raw data MLE) showed that the variable number of ailments and symptomatic conditions, health facility site, ART and CXT statuses, level of education, smoking status and having a salary were significantly associated with developing ailments and symptomatic conditions. Patients who experience only one ailment and symptomatic condition were 2.14 times (OR 2.14; 95%CI: 1.35-3.37) more likely to advance in WHO staging compared to those who never had an ailment or symptomatic condition. Patients enrolled at Tintswalo Hospitals (reference: PHRU Hospital), receiving ART (reference: not on ART) ,CTX (reference: not on CTX) drugs and smokers (reference: non-smokers) had an INCREASED RISK of advancing in WHO stages, (OR 4.22; (95%CI: 2.70-6.62), (OR 1.58; (95%CI: 1.19-2.10), (OR 2.07; (95%CI: 1.56-2.74) and (OR 1.78; (95%CI: 1.05-3.05) respectively. HIV patients who had attained tertiary education (reference: no education) and having a salaried job (reference: Patient receiving salary) had a protective effect of advancing in WHO staging, (OR 0.43; (95%CI: 0.19-0.99) and (OR 0.52; (95%CI: 0.33-0.82) respectively.

The multiple imputed model (MI MLE) detected that the significant risk factors were number of ailments and symptomatic conditions, health facility site, ART and CXT statuses. Patients who experience only one, two and three or more ailments and symptomatic conditions were more likely to advance in WHO staging compared to those who never had an ailment or symptomatic condition, (OR 1.78; (95%CI: 1.48-2.13), (OR 1.98; (95%CI: 1.58-2.48) and (OR 2.36; (95%CI: 1.78-3.11) respectively. Patients enrolled at Tintswalo Hospitals (reference: PHRU Hospital), receiving ART (reference: not on ART) ,CTX (reference: not on CTX) drugs had an increased risk of advancing in WHO stages, (OR 1.84; (95%CI: 1.59-2.15), (OR 1.23; (95%CI: 1.09-1.38) and (OR 1.54; (95%CI: 1.32-1.81) respectively.

The Bayesian estimation complete case analysis model (raw data BE) indicated that the variable number of ailments and symptomatic conditions, gender, health facility site, ART and CXT statuses, level of education, smoking status and having a salary were significantly associated with developing ailments and symptomatic conditions. Patients who experience only one ailment and symptomatic condition were 2.21 times (OR 2.21; 95%CI: 1.36-3.43) more likely to advance in WHO staging compared to those who never had an ailment or symptomatic condition. Patients enrolled at Tintswalo Hospitals (reference: PHRU Hospital), receiving ART (reference: not on ART) ,CTX (reference: not on CTX) drugs and smokers (reference: non-smokers) had an increased risk of advancing in WHO stages, (OR 4.46; (95%CI: 2.78-6.60), (OR 1.61; (95%CI: 1.21-2.10), (OR 2.08; (95%CI: 1.55-2.73) and (OR 1.88; (95%CI: 1.07-3.04) respectively. HIV patients who had a attained tertiary education (reference: no education), females (reference: males) and having a salaried job (reference: no salary) had a protective effect of advancing in WHO staging, (OR 0.48; (95%CI: 0.20-0.97), (OR 0.61; (95%CI: 0.37-0.95) and (OR 0.53; (95%CI: 0.32-0.82)respectively.

Table 3.8: The three random effects ordered logistic regression models for raw data with maximum likelihood, maximum likelihood on multiple imputed data, and raw data with Bayesian estimation

	Complete case		**MI MLE model**		**Raw data BE Model**	
	Odds Ratio	**(95% Conf.I)**	**Odds Ratio**	**(95% Conf.I)**	**Odds Ratio**	**(95% Conf.I)**
Number of ailments and symptomatic conditions						
0	(base)		(base)		(base)	
1	**2.1359**	**(1.3531,3.3717)**	**1.777**	**(1.4832,2.1290)**	**2.2064**	**(1.3573,3.430)**
2	1.5545	(0.9486,2.5472)	**1.9801**	**(1.5795,2.4825)**	1.6113	(0.9406,2.5444)
>=3	1.6592	(0.8641,3.1859)	**2.3554**	**(1.7814,3.1143)**	1.7446	(0.8535,3.201)
Gender						
Male	(base)		(base)		(base)	
Female	0.6004	(0.3566,1.0108)	0.9749	(0.8114,1.171)	**0.6100**	**(0.3679,0.9545)**
Study site						
PHRU	(base)		(base)		(base)	
Tintswalo	**4.2285**	**(2.7024,6.6162)**	**1.844**	**(1.5854,2.1455)**	**4.4576**	**(2.7821,6.601)**
Patient on ART						
No	(base)		(base)		(base)	
Yes	**1.5828**	**(1.1948,2.0969)**	**1.2252**	**(1.0884,1.3793)**	**1.6096**	**(1.2055,2.0952)**
Patient on Cotrimoxazole						
No	(base)		(base)		(base)	
Yes	**2.067**	**(1.5619,2.7355)**	**1.543753**	**(1.3186,1.8074)**	**2.0758**	**(1.5507,2.7270)**
Education level						
None	(base)		(base)		(base)	
Grade 0 - 12	1.296	(0.7152,2.3481)	0.9847	(0.7938,1.2215)	1.416577	(0.7340,2.2592)
Tertiary	**0.4317**	**(0.1875,0.9938)**	0.8371	(0.5965,1.1747)	**0.483**	**(0.2006,0.9697)**
Smoking status						
No	(base)		(base)		(base)	
Yes	**1.7847**	**(1.0458,3.0456)**	1.1374	(0.9326,1.3881)	**1.8802**	**(1.0663,3.0378)**
Patient receiving a salary						
No	(base)		(base)		(base)	
Yes	**0.5165**	**(0.3261,0.8180)**	0.8908	(0.6937,1.1439)	**0.5303**	**(0.3194,0.8207)**

Bold entries represent significant values

3.4 Simulated data results

The sensitivity analysis using simulated data are shown in table 3.9. Comparing model estimates in the 25% missing data models, the parameter estimates from MI data and BE are very similar and slightly smaller than those from the complete case analysis. However, the MI model had smaller standard errors hence narrower confidence intervals compared to the BE. When the level of missing data was increased to 50%, the MI MLE model and BE followed a similar patterns but with increased standard errors hence larger confidence intervals.

Table 3.9: The three random effects ordered logistic regression models for simulated data with maximum likelihood, maximum likelihood on multiple imputed simulated data, and simulated data with Bayesian estimation.

Variables	**True Parameters**	**MI MLE models**		**Raw data BE Models**	
y	**Odds Ratio**	**Odds Ratio**	**Std. Err.[Con.Inter]**	**Odds Ratio**	**Std. Dev[95%Cred.Inter]**
		Models with about 25 percent missingness			
d	1.0014	1.0018	0.0046(0.9927,1.0109)	0.9958	0.0058(0.9848,1.0070)
w					
0	1.0000	1.0000	(base)	(base)	
1	0.9531	0.9553	0.0335(0.8901,1.0254)	0.9507	0.0428(0.8692,1.0373)
m					
1	1.0000	1.0000	(base)	(base)	
2	1.0057	0.9668	0.0434(0.8848,1.0565)	0.9308	0.0564(0.8264,1.0468)
3	0.9938	0.9577	0.0444(0.8735,1.0499)	0.9876	0.0627(0.8687,1.1155)
4	1.0082	0.9742	0.0504(0.8766,1.0826)	0.9884	0.0622(0.8717,1.1149)
		Models with about 50 percent missingness			
d	1.0014	1.0072	0.0060(0.9946,1.0201)	1.0312	0.0127(1.0065, 1.0570)
w					
0	1.0000	1.0000	(base)	(base)	
1	0.9531	0.9807	0.0418(0.8951,1.0745)	0.9830	0.0831(0.8310, 1.1566)
m					
1	1.0000	1.0000	(base)	(base)	
2	1.0057	0.9967	0.0701(0.8506,1.1678)	1.1010	0.1320(0.8644, 1.3839)
3	0.9938	1.0699	0.0618(0.9471,1.2086)	1.1349	0.1375(0.8897,1.4372)
4	1.0082	1.0130	0.0556(0.9039,1.1353)	1.0972	0.1357(0.8552, 1.3882)

The table shows missingness in the simulated dataset

3.5 Model diagnostics

After running the displayed model results in table 3.8, post estimation test were conducted for each model to assess model validity and fitness so as to get a parsimonious model. For the raw data MLE model, maximum likelihood ratio test was used to compare MLE raw data model with random effects and without random effects. A p-value of <0.001, which showed that the model with random effects was better. The random effects variance was 9.52 (95%CI: 7.841- 11.56) with standard error of 0.94. From this reported likelihood, it suggested that there was enough variability between patients to favour a random effect ordered logistic model versus a standard ordered logistic model. In addition, the quadrature approximation was assessed for after running the random effects model since a good random effects model fit depends on a combination of the quality of the quadrature approximation and the goodness of the data. The fitted quadrature approximation coefficients (Appendix E) which did not change by a relative difference of 0.01%, hence the reported findings can be confidently reported.

The Bayesian model gave an acceptance rate of 36.1% which is greater than 10% which shows no problem in convergence of the model. The average efficiency of the model was 9.9% which is greater than 1% hence the mixing of the model was good. Trace plots, autocorrelation plots, histogram plots and density plots are shown in Appendix D.1 for all the three Bayesian models fitted. The trace plots rapidly traverse which was a good sign of mixing of the model while the autocorrelation dies off rapidly with few lags. The histogram plots were approximately normally distributed and the first and second half curves were roughly superimposed on the main curve.

Overally the the multiple imputation model and the Bayesian model provided a good fit to the data. This is evident by the assumptions of our models that were not violated. The main assumption of the quadrature on the random effects ordinal logistic regression model was not violated as well as the convergence of all the models were good. The multiple imputation model was used as our final model for interpretation because of it’s superior smaller confidence intervals.

Chapter 4

Discussion and Conclusion

Introduction

This chapter discusses the results from our analysis by comparing them with other studies, for some of the key results. The researcher also address the strength and limitations of the study.

4.1 Discussions

The aim of this study was to determine the possible risk factors associated with developing severe ailments and symptomatic condition and determine the transitions of of these ailments and symptomatic conditions among HIV patients enrolled in the Adult Wellness study between 2002 to 2010. From the MI MLE model used the researcher found significant association of advancing in WHO staging with number of ailments and symptomatic conditions experienced during the follow up period, the health facility a patient was enrolled at, whether a patient is taking ART treatment and CTX drugs.

The risk factors identified in this study are not different to other studies conducted earlier [67]. The studies observed that the more the person experienced the HIV related opportunistic infections, the more likely they were to progress to advanced WHO stages. This is supported by the fact that, once a person is HIV infected, they are vulnerable to catch any infections since they would be immunocompromised.

This study identified that, patients enrolled at Tinswalo Hospital had an increased risk of developing more severe ailments and symptomatic conditions. This could be explained by the fact that the catchment area for Tinswalo is mainly rural, hence patients may present late at the hospital , challenges with hospital accessibility linked to transport costs and poor adherence to treatment [68]. Patients enrolled at PHRU through Baragwanath Hospital being in urban and at a big health facility were more likely to received better health care and resources were readily available with much more modern patient care equipment [69].

There was an increased risk of developing ailments and symptomatic conditions amongst patients who were receiving ART and CTX drugs. This was surprising since one would expect a person on ART or CTX to become better and not show any symptoms of deterioration. The study finding contrast what was reported in a study by Daria Podlekareva et al which detected a protective effect of ART combination therapy in relation to hospital admissions. The current study findings could be explained by adherence. Patients who have poor adherence to ART treatment are likelier to become worse than recover. Poor adherence is linked to ART drug toxicity which has medical side effect like stavudine (D4T) which was eventually banned [70].

This study could not detect any significant association between developing severe ailments and symptomatic conditions and gender, however being a female was protective. This finding can be explained by the fact that women usually seek treatment for every ailment compared to men who typically seek health care when their health condition has much deteriorated [71, 72].

The backward transitions (recovery) were much more likely than the forward transitions (getting worse). This is also evident by a relatively high probability of staying in the same state compared to those of moving from one state to another and also by the transitions rates that are small. This could be explained by medical attentions the patients could be getting after admissions into the study at Tinswalo and PHRU hence their health state remained under monitoring or improvement in adherence once they have suffered an ailment.

This study two three different types of modelling approaches. First, a complete case model (raw data MLE) was fitted to estimate the possible risk factors of developing ailments and symptoms. This model utilised only those observations which have complete information and exclude all missing observations. As a result, the precision of the estimates was compromised due to a reduced sample size of observation. The estimates produced had wide confidence interval which is an issue in inferential statistics and can overestimate the association as evident in this study. If such an approach is conducted, invalid conclusions and incorrect decisions are made which have implications to patient survival [31].

In order to correct this problem of missing data, a MI MLE model was fitted. This model had increased power due to an increase in the observation used in the modelling. The estimates from this model were found to be precise since the confidence interval were much narrower than those from the case limited model. The model with MI produced fewer significant covariates that is (number of ailments and symtomatic conditions, study site, patient on ART, Patients on cotrimoxazole) and produced relatively lower odds ratios. This means if one had used a complete case analysis model, some risk factors would have been detected of which they would not be correct for this setting. Also, there would have been either overestimation or underestimation of the association existing.

According to this research in modelling, one can either choose to use MLE and BE techniques. MLE technique is mainly data driven, that is, it estimates directly from the data available. As a result, this method is sensitive to sample size. If the sample size is small, statistical inference can be compromised due to large standard errors as evident in this study's findings (table 3.8).

A sensitivity analysis performed using simulated data showed that BE and MI models showed similar estimates of the predictor variables for the 25% and 50% level of missingness. However, Bayesian models gave results that had larger standard errors and wider confidence intervals compared to MI model. This can be explained by the fact that BE model used complete case analysis hence estimation was based on 50% of the data whilst MI used the complete data which is 100%. Basing on this, the addition of prior information does not necessarily undo the evidence from the complete dataset. This is similar to what is reported in the literature [7, 16].

4.2 Study strengths and limitations

This study's main strength was to model the possible risk factors of developing severe ailments and symptomatic condition adjusting for other confounding factors using multiple variable regression on baseline characteristics. However, the researcher did not use time varying covariates and this may have some implications on the findings. The modelling approach used was robust in a way that is accounted for individual heterogeneity. This is important in HIV studies since the HIV progression varies by individual genetics and behavior [73]. This kind of approach gives more precise estimates compared to models with no randon effect. This is similar to what was found in a study by

Musenge at al [74] when they found better estimates after using a model with spatial household random effect compared to the standard model. The multiple imputation method used produced better results compared to complete case models. This was done to account for missingness which was observed to be an issue in this study. Lastly, a sensitivity analysis was done on the three models fitted by use of simulated data. This showed similar pattern of results from the Adult Wellness data.

4.3 Conclusion

When analysing data from longitudinal observation studies, it is important to explore the data by performing various types of disposable modelling approaches to see which one estimates better. MI model gave better precise estimates with narrower confidence intervals hence it should be used compared to BE model. Patients with ailments and symptomatic conditions can move from one state to the other at different rates and probability.

Bibliography

[1] United Nations Programme on HIV and AIDS. South Africa 2016. Available from: http://www.unaids.org/en/regionscountries/countries/South Africa [cited 20 February 2018].

[2] United Nations Programme on HIV and AIDS. Prevention Gap Report 2016. Available from: http://www.unaids.org/sites/default/files/media_asset/2016-prevention-gap-report_en.pdf [cited 12 September 2017].

[3] Statistics South Africa. Mid-year Population estimates 2015. *www.statssa.gov.za/publications/.../P03022014.pdf*, (July):1–18 [cited 12 september 2017].

[4] Bussiness Report. R23bn spent on HIV/Aids each year 2016. https://www.iol.co.za/business-report/economy/r23bn-spent-on-hivaids-each-year-2042364[cited 17 December 2017].

[5] National Institute of statistics, Rwanda Demographic and Health Survey, 2015, Https://dhsprogram.com/pubs/pdf/FR316/FR316.pdf [cited 23 january 2018].

[6] Joseph S Gonnella, Marvin Vincent E Gozum, Clara A Callahan, and Cathleen A Barnes. Disease Staging Clinical And Coded Criteria. *Ann Arbor, MI: Thomson Medstat*, 2003.

[7] James D Dziura, Lori A Post, Qing Zhao, and Zhixuan Fu. Strategies for dealing

with Missing data in clinical trials : From design to Analysis. *The Yale journal of biology and medicine*, 86:343–358, 2013.

[8] Jaeil Ahn, Suyu Liu, Wenyi Wang, and Ying Yuan. Bayesian latent-class mixed-effect hybrid models for dyadic longitudinal data with non-ignorable dropouts. *Biometrics*, 69(4):914–924, 2013.

[9] United States Department of Health and Human Services. When to Start Antiretroviral Therapy. Available from: Https://aidsinfo.nih.gov/understanding-hiv-aids/fact-sheets/21/52/when-to-start-antiretroviral-therapy [cited 5 september 2017]. .

[10] The National Institutes of Health (NIH):. Turning Discovery Into Health,https://www.nih.gov/,[cited 13 Oct 2017], 2015.

[11] G. Frank Liu, Baoguang Han, Xin Zhao, and Qun Lin. A Comparison of Frequentist and Bayesian Model Based Approaches for Missing Data Analysis: Case Study with a Schizophrenia Clinical Trial. *Statistics in Biopharmaceutical Research*, 8(1):116–127, 2016.

[12] Kennedy N Otwombe, Max Petzold, Tebogo Modisenyane, Neil A Martinson, and Tobias Chirwa. Factors associated with mortality in hiv-infected people in rural and urban south africa. *Global health action*, 7(1):25488, 2014.

[13] Surendra K Sharma and Manish Soneja. Hiv & immune reconstitution inflammatory syndrome (iris). *The Indian journal of medical research*, 134(6):866, 2011.

[14] World Health Organization. WHO case definitions of HIV for surveillance and revised clinical staging and immunological classification of HIV-related disease in adults and children. *World Health*, page 52, 2007.

[15] Min Tang, Eric V Slud, and Ruth M Pfeiffer. Goodness of fit tests for linear mixed models. *Journal of Multivariate Analysis*, 130:176–193, 2014.

[16] Yiran Dong and Chao-Ying Joanne Peng. Principled missing data methods for researchers. *SpringerPlus*, 2(1):222, 2013.

[17] Idindili Boniphace, Minzi Omari, Rumisha Susan Fred, Mugusi Ferdinand, Tanner Marcel, Box Dar, and Es Salaam. HIV / AIDS Clinical Manifestations and their Implication for Patient Clinical Staging in Resource Limited Settings in Tanzania. *Open Access*, pages 9–16, 2011.

[18] Sead Ahmetagic, Humera Porobić-jahic, Dilista Piljic, Amer Custovic, Damir Sabitovic, and Denis Zepic. Clinical and Epidemiological Characteristics of HIV Infection / AIDS in Hospitalized Patients. *Acadedemy of Medical Sciencies of Bosinia and Herzegovia*, 78(December 2014):27–30, 2015.

[19] Ayesha BM Kharsany and Quarraisha A Karim. Hiv infection and aids in sub-saharan africa: current status, challenges and opportunities. *The open AIDS journal*, 10:34, 2016.

[20] Karl Peltzer, Shandir Ramlagan, Salim Khan, and Bernhard Gaede. The social and clinical characteristics of patients on antiretroviral therapy who are ' lost to follow- up ' in KwaZulu-Natal , South Africa : a prospective study. *Journal of acquired immune deficiency sydrome*, 0376, 2012.

[21] Christopher H Jackson. Multi-State Models for Panel Data: The msm Package for R. *Journal of Statistical Software*, 38(8), 2011.

[22] Kendra Brown Mhoon, Wenyaw Chan, Deborah J. Del Junco, , and Sally W. Vernon. A Continuous-Time Markov Chain Approach Analyzing the Stages of Change Construct from a Health Promotion Intervention. *JP journal of biostatistics*, 4(3):213–226, 2013.

[23] M J Sweeting, V T Farewell, and D De Angelis. Multi-state Markov models for disease progression in the presence of informative examination times : An ap-

plication to hepatitis C. *Statistics in medicine*, 29(November 2008):1161—-1174, 2010.

[24] Ofer Harel, Jennifer Pellowski, and Seth Kalichman. Are we missing the importance of missing values in hiv prevention randomized clinical trials? review and recommendations. *AIDS and Behavior*, 16(6):1382–1393, 2012.

[25] Finbarr P. Leacy, Sian Floyd, Tom A. Yates, and Ian R. White. Analyses of sensitivity to the missing-at-random assumption using multiple imputation with delta adjustment: Application to a tuberculosis/HIV prevalence survey with incomplete HIV-status data. *American Journal of Epidemiology*, 185(4):304–315, 2017.

[26] Diane L. Fairclough, Herbert Thijs, I-Chan Huang, Henrik W. Finnern, and Albert W. Wu. Handling missing quality of life data in HIV clinical trials: what is practical? *Quality of Life Research*, 17(1):61–73, 2008.

[27] S. Yang and J. K. Kim. A note on multiple imputation for method of moments estimation. *Biometrika*, 103(1):244–251, 2015.

[28] Global Health Observatory (GHO) data. NCD mortality and morbidity 2015 http://www.who.int/gho/ncd/mortality_morbidity/en/. [cited 27 june 2018].

[29] A Rogier T Donders, Geert JMG Van Der Heijden, Theo Stijnen, and Karel GM Moons. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091, 2006.

[30] Jerome P. Reiter. Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika*, 95(4):933–946, 2008.

[31] Hyun Kang. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5):402–406, 2013.

[32] Jesper N Wulff and Linda Ejlskov. Multiple Imputation by Chained Equations in Praxis : Guidelines and Review. *Electronic Journal of Business Research Methods*, 15(1):41–56, 2017.

[33] Rita Faria, Manuel Gomes, David Epstein, and Ian R White. A Guide to Handling Missing Data in Cost-Effectiveness Analysis Conducted Within Randomised Controlled Trials. *Pharmacoeconomics*, 32(12):1157–1170, 2014.

[34] Michael J. Daniels and Chenguang Wang. Missing data methods in longitudinal studies: a review. *Test*, 18(1):1–41, 2009.

[35] David C. Howell. Mixed Models for Missing Data With Repeated Measures Part 1 2015.Https://www.uvm.edu/˜dhowell/StatPages/More_Stuff/Mixed-Models-Repeated/Mixed-Models-for-Repeated-Measures1.html [cited 20 May 2018].

[36] Michikazu Nakai and Weiming Ke. Review of the Methods for Handling Missing Data in Longitudinal Data Analysis. *Int. Journal of Math. Analysis*, 5(1):1–13, 2011.

[37] Rema Raman and Donald Hedeker. A mixed-effects regression model for three-level ordinal response data. *Statistics in Medicine*, 24(21):3331–3345, 2005.

[38] Nan M Laird and James H Ware. Random-Eff ects Models for Longitudinal Data. *Science*, 38(4):963–974, 2007.

[39] Kathryn M Aloisio, Sonja A Swanson, Nadia Micali, Alison Field, and Nicholas J Horton. Analysis of partially observed clustered data using generalized estimating equations and multiple imputation. *The Stata journal*, 14(4):863, 2014.

[40] Kaifeng Lu, Liqiu Jiang and Anastasios A. Tsiatis. Multiple Imputation Approaches for the Analysis of Dichotomized Responses in Longitudinal Studies with Missing Data. *Biometrics*, 27(3):320–331, 2015.

[41] Joseph G Chen, Qingxia and Ibrahim. A note on the relationships between multiple imputation, maximum likelihood and fully Bayesian methods for missing responses in linear regression models. *Statistics and its Interface*, 6(3):315–324, 2014.

[42] Joseph G Ibrahim, Ming-hui Chen, Stuart R Lipsitz, Amy H Herring, Joseph G Ibrahim, Ming-hui Chen, Stuart R Lipsitz, Amy H Herring, Joseph G I Brahim, Ming-hui C Hen, Stuart R L Ipsitz, and Amy H H Erring. Missing-Data Methods for Generalized Linear Models Missing-Data Methods for Generalized Linear Models : A Comparative Review. *Journal of the American Statistical Association*, 1459:332—-346, 2011.

[43] AS Macdonald. An actuarial survey of statistical models for decrement and transition data-i: Multiple state, poisson and binomial models. *British Actuarial Journal*, 2(1):129–155, 2018.

[44] Priyanka Srikanth. Using Markov chains to predict the natural progression of diabetic retinopathy. *International journal of ophthalmology*, 8(1):132–137, 2015.

[45] Christopher H Jackson, Linda D Sharples, Simon G Thompson, and Stephen W Duffy. Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52:193–209, 2003.

[46] Irwin DJ Bross. Critical levels, statistical language and scientific inference. *Foundations of statistical inference*, pages 500–513, 1971.

[47] Integre Tech, British Social, Attitudes Survey, Sophia Rabe-hesketh, Anders Skrondal, Sophia Rabe-hesketh, and Tolman Hall. Generalized multilevel structural equation modeling s. *Psychometrika*, 69(167-190), 2004.

[48] J. C. Naylor Smith and A. F. M. Applications of a Method for the Efficient Com-

putation of Posterior Distributions : http://www.jstor.org/stable/2347995. *Royal Statistical Society*, 31(3):214–225, 1982.

[49] Panteha Hayati Rezvan, Katherine J Lee, and Julie A Simpson. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC medical research methodology*, 15(1):30, 2015.

[50] John W Graham, Allison E Olchowski, and Tamika D Gilreath. How Many Imputations are Really Needed ? Some Practical Clarifications of Multiple Imputation Theory. *Prevention science*, 8:206–213, 2007.

[51] Workflow Mail Survey. Determining Sufficient Number of Imputations Using Variance of Imputation Variances: Data from 2012 NAMCS Physician Workflow Mail Survey. *Applied mathematics*, 5(2014):3421–3430, 2016.

[52] Craig D Newgard and Jason S Haukoos. Advanced Statistics: Missing Data in Clinical Research—Part 2: Multiple Imputation. *Academic Emergency Medicine*, 14:669—-678, 2007.

[53] Yang C Yuan. Multiple Imputation for Missing Data : Concepts and New. *SAS Institute Inc, Rockville, MD*, 49:1–13, 2010.

[54] John B Carlin, Ning Li, Philip Greenwood, and Carolyn Coffey. Tools for analyzing multiple imputed datasets. *The Stata Journal*, 3:226—-244, 2003.

[55] Joseph L Schafer, Maren K Olsen, Joseph L Schafer, and Maren K Olsen. Multiple Imputation for Multivariate Missing-Data Problems : A Data Analyst ' s Perspective Missing-Data Problems : A Data Analyst ' s Perspective. *Multivariate behavioral research*, 3171:545—-571, 2010.

[56] Xiao-Li Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4):538–558, 1994.

[57] Dick L Van Buuren, Stef and Boshuizen, Hendriek C and Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, 18:681—-694, 1999.

[58] Trivellore E Raghunathan, James M Lepkowski, John Van Hoewyk, and Peter Solenberger. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey methodology*, 27(1):85–95, 2001.

[59] S Van Buuren, J P L Brand, D B Rubin, J P L Brand, and D B Rubin. Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, 76:1049—-1064, 2007.

[60] Stef Van Buuren. Multiple imputation of discrete and continuous. *Statistical methods in medical research*, 16:219–242, 2007.

[61] Nicholas J Horton and Ken P Kleinman. and software to fit incomplete data regression models. *The American Statistician*, 61(1):79–90, 2007.

[62] Jennifer S.K. Chan. Bayesian informative dropout model for longitudinal binary data with random effects using conditional and joint modeling approaches. *Biometrical Journal*, 58(3):549–569, 2016.

[63] Hirotugu Akaike. A bayesian extension of the minimum aic procedure of autoregressive model fitting. *Biometrika*, 66(2):237–242, 1979.

[64] Brian Patenaude, Stephen M Smith, David N Kennedy, and Mark Jenkinson. A bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage*, 56(3):907–922, 2011.

[65] Alexander McFar1ane Mood. *Introduction to the theory of statistics: McGraw-hill*. 1974.

[66] Valen E Johnson et al. A bayesian $\chi 2$ test for goodness-of-fit. *The Annals of Statistics*, 32(6):2361–2384, 2004.

[67] F Hidata, A Worku, and F Urgessa. Contraception use and factors contributing to non-use of contraception among in-school adolescents in toke kutaye woreda, west shoa zone, oromia regional state, ethiopia. *Journal of Pregnancy and Child Health*, 2:186, 2015.

[68] Zewdie Mulissa, Degu Jerene, and Bernt Lindtjørn. Patients present earlier and survival has improved, but pre-art attrition is high in a six-year hiv cohort data from ethiopia. *PloS one*, 5(10):e13268, 2010.

[69] Fahui Wang and Wei Luo. Assessing spatial and nonspatial factors for healthcare access: towards an integrated approach to defining health professional shortage areas. *Health & place*, 11(2):131–146, 2005.

[70] Isabelle Andrieux-Meyer, Eric Goemaere, Sharonann Lynch, Tido Von Schoen-Angerer, Polly Clayden, Simon Collins, Nathan Geffen, Mark Harrington, and Tracy Swan. Why it's time to say goodbye to stavudine... everywhere: debate. *Southern African Journal of HIV Medicine*, 13(1):17–19, 2012.

[71] Paul M Galdas and Paul Marshall. Men and health help-seeking behaviour : literature review. *Journal of advanced nursing*, 49:616—-623, 2005.

[72] P M Pronyk, M B Makhubele, J R Hargreaves, S M Tollman, and H P Hausler. Assessing health seeking behaviour among tuberculosis patients in rural South Africa. *The International Journal of Tuberculosis and Lung Disease*, 5:619–627, 2001.

[73] Rachel Jewkes and Robert Morrell. Gender and sexuality: emerging perspectives from the heterosexual epidemic in south africa and implications for hiv risk and prevention. *Journal of the International AIDS society*, 13(1):6, 2010.

[74] Eustasius Musenge, Penelope Vounatsou, Mark Collinson, Stephen Tollman, and Kathleen Kahn. The contribution of spatial analysis to understanding hiv/tb mor-

tality in children: a structural equation modelling approach. *Global health action*, 6(1):19266, 2013.

Appendices

Appendix A

Plagiarism declaration report

Appendix B

Variables used in the study

B.1 Exposure and Outcome variables

Exposure variables

Table B.1: A list of the selected exposure variables from the Adult Wellness study

Number of ailments and symptomatic conditions → These are are combined counts of these conditions a patient is diagnosed with during follow-up
Patient on ART → This is an indicator variable which shows if a patient is on ART or not
Patient on cotrimoxazole → This is an indicator variable which shows if a patient is on cotrimoxazole or not
BPSys → Is the pressure in your blood vessels when your heart beats.
Blood pressure diastolic → Is the pressure in your blood vessels when your heart rests between beats.
Sexually transmitted infections → This is a binary variable which shows whether a patient have been treated for STI's in the year before your first visit or since last study visit
CD4 count → This is a measure of how many CD4 cells a patient had in their blood
Estimated age in years → This is a continuous variable which shows the age of a patient at each visit
Education level → This is the education a patient has attained by the time he/she make a visit
Smoking status → This is a binary variable which shows whether a patient have ever smoked, either in the past or now?
Patient receiving salary → This is a binary variable which shows whether a patient is on salary or not
Marital status → This is a nominal random variable which indicates whether a patient is single , married, divorced or widowed or not
self employed This is an indicator variable which shows whether a patient is self employed or not
Sex Age → Age at which a patient first had sex
Alcohol → This is a binary variable which shows whether a patient have ever taken alcohol or not.
Height → This is a variable which gives the height of a patient without shoes
Weight → Weight of a patient at each visit
condom → This is a a binary variable which indicates whether a patient has used condom with any of his/her last four sexual partners
sex partners → This is a count variable which gives the number of sexual partners an HIV patient had in his/her life.

Table B.2: Ailments and asymptotic conditions outcomes that were grouped and used for analysis in this study

Ailments and symptomatic conditions		
WHO Stage		
I	No Symptoms	Persistent Generalized Lymphadenopathy
II	Skin Rash, Patches Or Plaques	Prurigo (Chronic Itchy Skin)
	Herpes Zoster within last 5 years	
	Recurrent upper respiratory tract infections	Weight Loss <10% Body Weight
III	Weight Loss >10% Body Weight	Unexplained Fever >1 month
	Oral Candidiasis	Bedridden <50% Of Day Most Days In Previous Month
	Pulmonary TB Within The Past Year	
	Vulvo-Vaginal Candidiasis >1 month	Oral Hairy Leucoplakia
	Pneumonia Or Other Serious Bacterial Infection	Diarrhea >1 month
IV	Extra pulmonary TB	Esophageal Candida
	Pneumocystis Pneumonia	Kaposi's Sarcoma
	Herpes Simplex >1 month (Skin/mouth/nose/eye)	HIV Wasting Syndrome
	Bedridden >50 % Of Day In Previous Month	Cryptococcosis extra pulmonary
	HIV Encelopathy /Dementia	Cryptosporidiosis with diarrhea >1 month
	Atypical mycobacteriosis (non tuberculosis)	Non - typhoid Salmonella septicaemia
	Progressive multifocal leucoencephalopathy	Cytomegalovirus other than liver/lymphnode
	Toxoplasmosis of the brain	Lymphoma
	Any disseminated endemic mycosis	

Appendix C

Missingness in the simulated data

C.1 Missing data frequencies and percentages in simulated data(with 25 % missingness in each variable)

Table C.1: The missingness patterns of each covariate used for analysis from the simulated data(with 25 % missingness in each variable).

	with 25 percent missing in each variable				
Variable	Obs	**Missings**	**NonMiss**	**Freq.Missings**	**Freq.NonMiss**
w	15000	3712	24.75	11288	75.25
m	15000	3798	25.32	11202	74.68
d	15000	3740	24.93	11260	75.07

Variable	Obs	Missings	Feq.Missings	NonMiss	Feq.NonMiss
ARVTreatment	2720	453	16.65	2267	83.35
CPTTreat	2720	360	13.24	2360	86.76
edulev	2720	1457	53.57	1263	46.43
Smoked	2720	233	8.566	2487	91.43
salaried	2720	1249	45.92	1471	54.08
Mstatus	2720	990	36.4	1730	63.6
selfemployed	2720	1279	47.02	1441	52.98
SexAge	2720	2494	91.69	226	8.309
Alcohol	2720	237	8.713	2483	91.29
Height	2720	1181	43.42	1539	56.58
Weight	2720	687	25.26	2033	74.74
condom	2720	0	0	2720	100
sexpartyear	2720	2052	75.44	668	24.56
BPSys	2720	1000	36.76	1720	63.24
BPDia	2720	1000	36.76	1720	63.24
STIyn	2720	506	18.6	2214	81.4
CD4	2720	2627	96.58	93	3.419

C.2 Missing data frequencies and percentages in simulated data(with 50 percent missingness in each variable)

Table C.2: The missingness patterns of each covariate used for analysis from the simulated data(with 50 percent missingness in each variable).

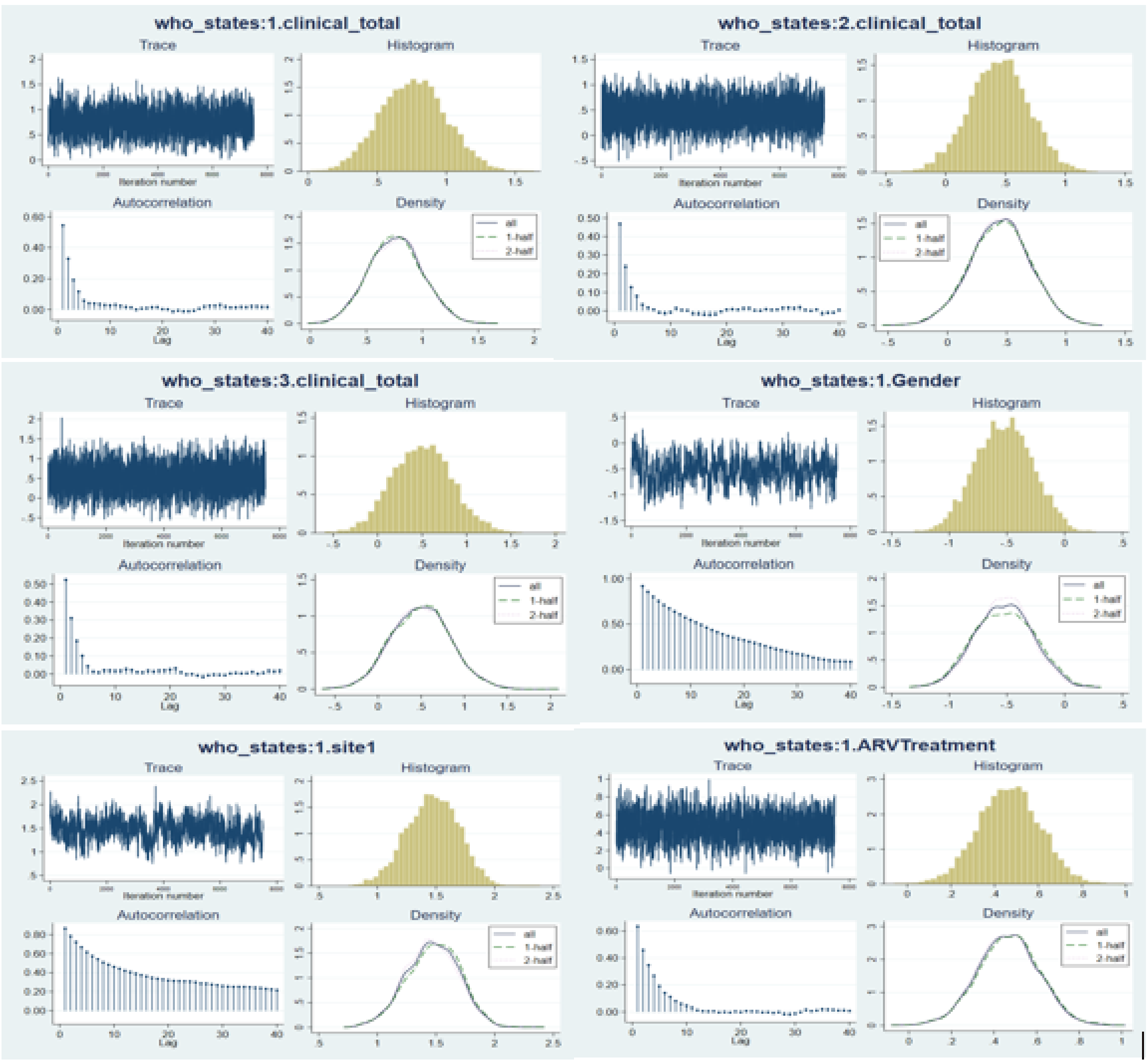
	50 percent missing in each variable				
Variable	**Obs**	**Missings**	**Freq.Missings**	**NonMiss**	**Freq.NonMiss**
w	15000	7513	50.09	7487	49.91
m	15000	7611	50.74	7389	49.26
d	15000	7481	49.87	7519	50.13

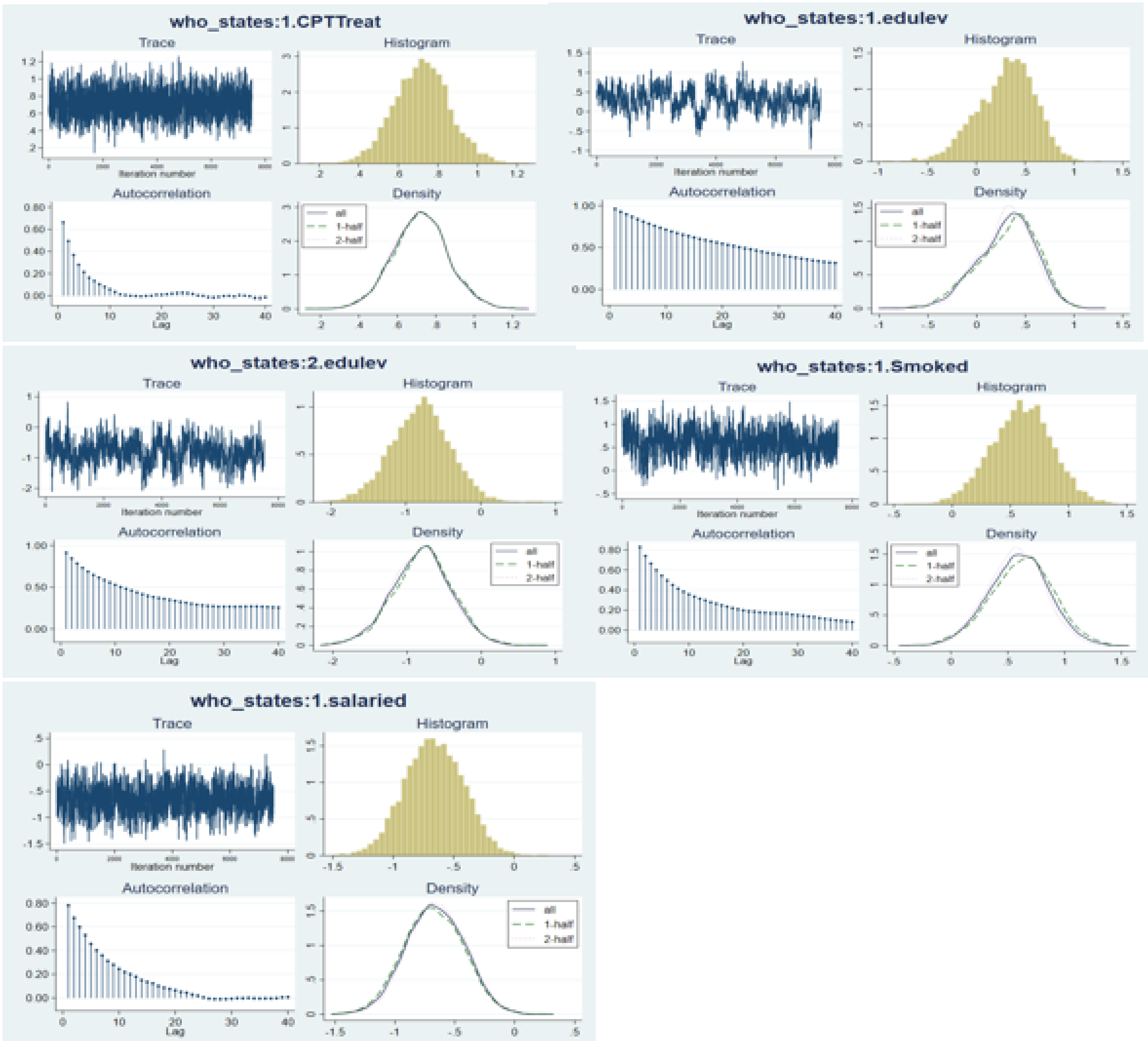
C.3 Missing data frequencies and percentages in baseline characteristics

Appendix D

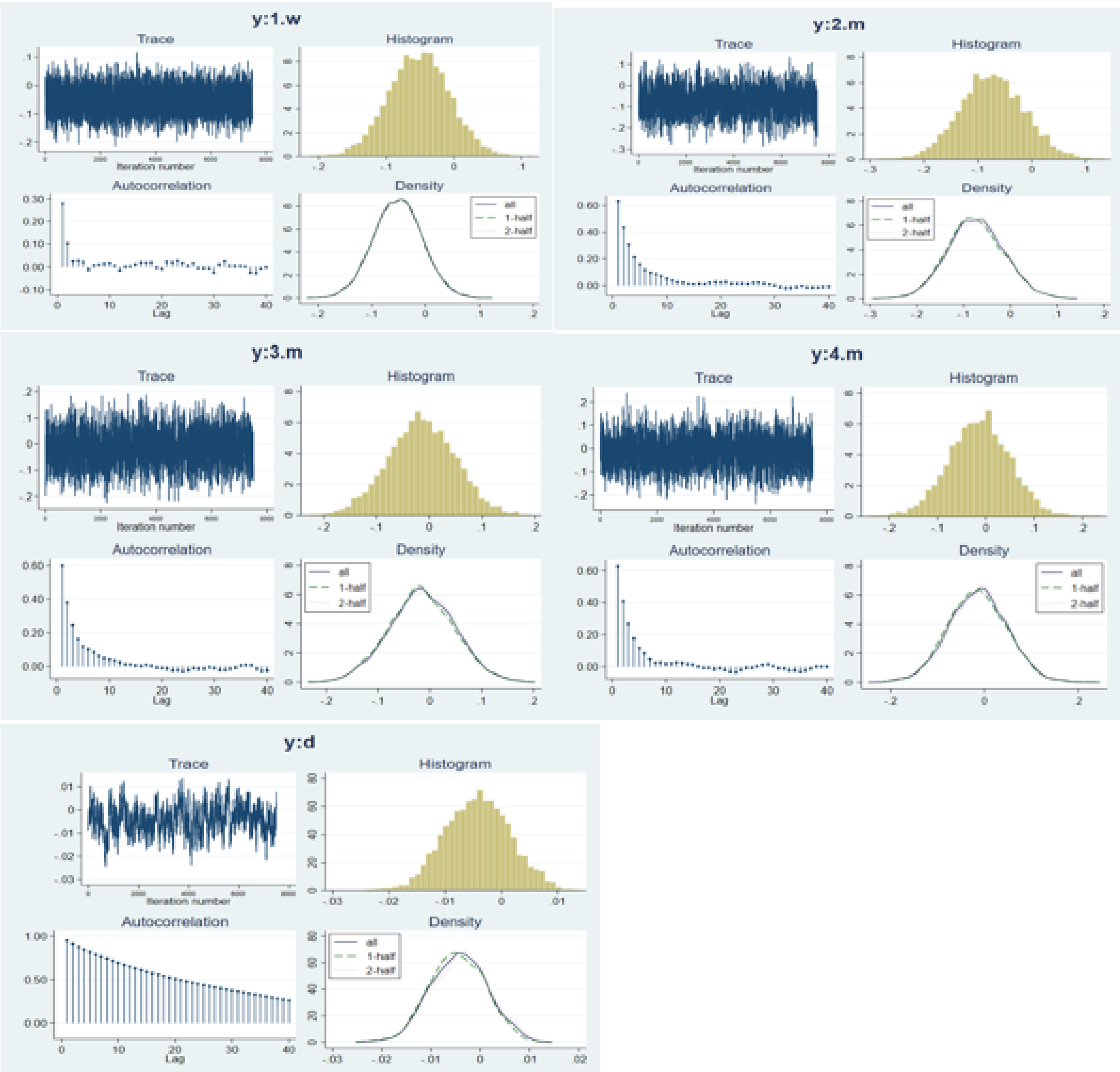
Diagnostic plots

D.1 Bayesian convergence plots for Adult Wellness data

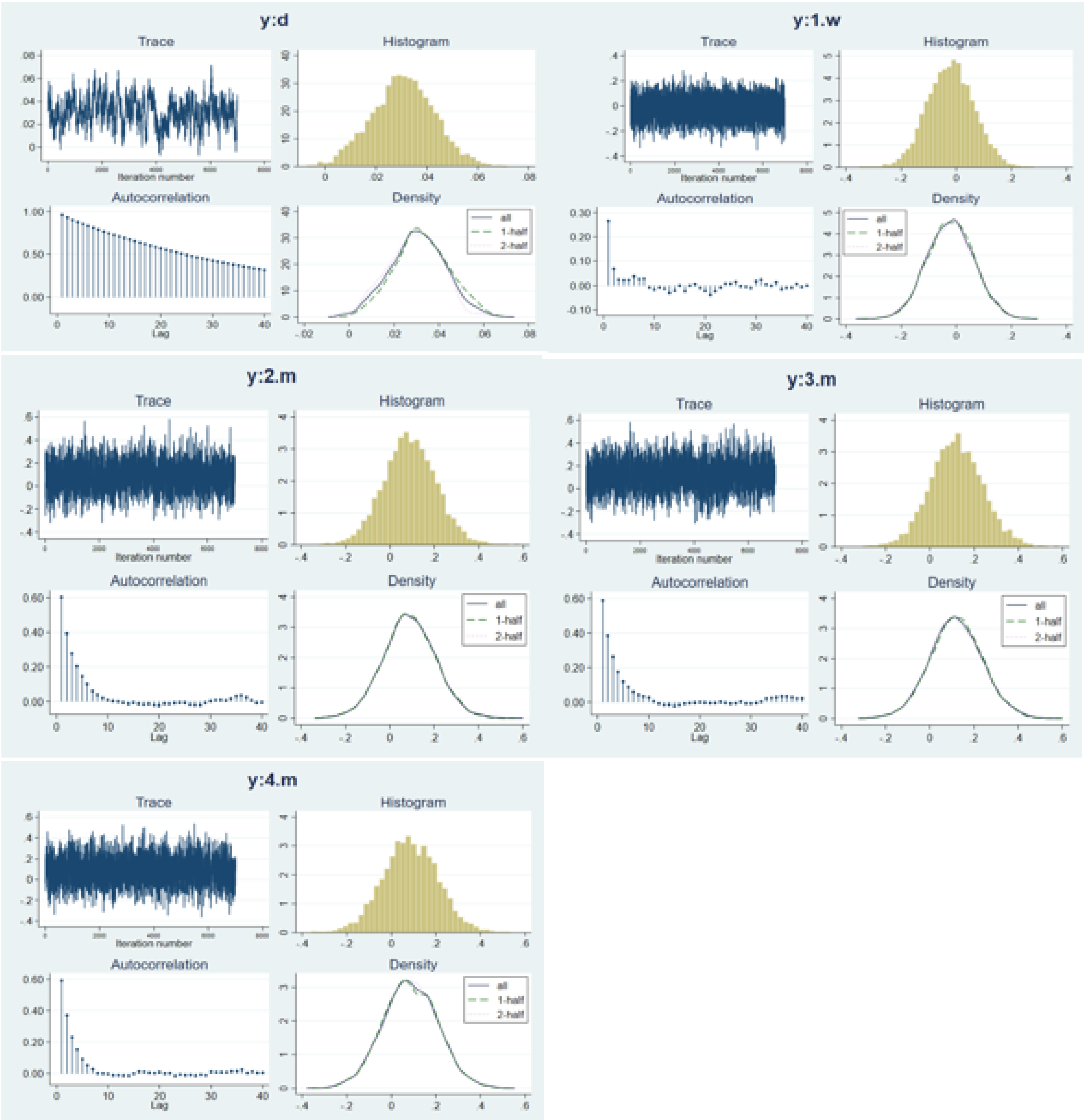



who_states:1.CPTTreat
Trace
Histogram
Iteration number
Autocorrelation
Density
Lag
all
1-half
2-half
who_states:1.edulev
Trace
Histogram
Iteration number
Autocorrelation
Density
Lag
all
1-half
2-half
who_states:2.edulev
Trace
Histogram
Iteration number
Autocorrelation
Density
Lag
all
1-half
2-half
who_states:1.Smoked
Trace
Histogram
Iteration number
Autocorrelation
Density
Lag
all
1-half
2-half
who_states:1.salaried
Trace
Histogram
Iteration number
Autocorrelation
Density
Lag
all
1-half
2-half

D.2 Bayesian convergence plots for simulated data (25 percent missingness in each variable)



D.3 Bayesian convergence plots for simulated data (50 percent missingness in each variable)



Appendix E

Quadrature for random effects models

	Model with raw dat		8 quadrature points		16 quadrature points	
	Coef	**(95% Conf.I)**	**Coef**	**(95% Conf.I)**	**Coef**	**(95% Conf.I)**
Number of Ailments and symptomatic conditions						
0	(base)		(base)		(base)	
1	0.7589	(0.3024,1.2154)	0.7574	(0.3011,1.2138)	0.7587	(0.3022,1.2153)
2	0.4411	(-0.0528,0.9350)	0.4399	(-0.0538,0.9337)	0.441	(-0.0530,0.9349)
>=3	0.5064	(-0.1460,1.5873)	0.5044	(-0.1481,1.1569)	0.5061	(-0.1463,1.185)
Gender						
Male	(base)		(base)		(base)	
Female	-0.5102	(-1.0312,0.0107)	-0.5089	(-1.0309,0.0130)	-0.5099	(-1.0301,0.0103)
Study site						
PHRU	(base)		(base)		(base)	
Tintswalo	1.4418	(0.9942,1.8895)	1.4319	(0.9849,1.8790)	1.4418	(0.9944,1.8893)
Patient on ART						
No	(base)		(base)		(base)	
Yes	0.4592	(0.1779,0.7404)	0.4573	(0.1759,0.7387)	0.4591	(0.1779,0.7403)
Patient on Cotrimoxazole						
No	(base)		(base)		(base)	
Yes	0.7261	(0.4459,1.0063)	0.723	(0.4426,1.0034)	0.7256	(0.4455,1.0057)
Education level						
None	(base)		(base)		(base)	
Grade 0 - 12	0.2593	(-0.3351,0.8536)	0.258	(-0.33660.8525)	0.2594	(-0.3350,0.8538)
Tertiary	-0.84	(-1.6738,-0.0062)	-0.83	(-1.6702,0.0101)	-0.8369	(-1.6695,-0.0043)
Smoking status						
No	(base)		(base)		(base)	
Yes	0.57923	(0.0448,1.1137)	0.5768	(0.0414,1.112o3)	0.5794	(0.0458,1.1131)
Patient receiving salary						
No	(base)		(base)		(base)	
Yes	-0.6607	(-1.1206,-0.009)	-0.6586	(-1.1215,-0.1958)	-0.66	(-1.1193,-0.2009)

Appendix F

Ethics clearance

F.1 Ethics clearance certificate

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

R14/49 Mr Batsirai Murapa et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M180335

NAME: **(Principal Investigator)**	Mr Batsirai Murapa et al
DEPARTMENT:	School of Public Health Epidemiology and Biostatistics
PROJECT TITLE:	Factors Associated with World Health Organization (WHO) Stage Clinical Characteristics in HIV Positive Patients: An Application of Missing Data Techniques
DATE CONSIDERED:	06/04/2018
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Prof Eustasius Musenge
APPROVED BY:	Professor CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	11/04/2018

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 on the Third Floor, Faculty of Health Sciences, Philip 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 **March** and will therefore be due in the month of **March** 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

Appendix G

Codes used in data analysis

G.1 R codes for Markov multi-state model

setwd("C:/Users/amen/Desktop/MSc Project for manipulations TO ETHICS/Adult Wellness Study") # Importing data from CSV

The easiest way to install the msm package on a computer connected to the Internet is to run the R

```
# loading all necessary packages
# install.packages("msm")
# install.packages("forecast")
# install.packages("markovchain")
# install.packages("dplyr")
# install.packages("tidyverse")
# install.packages("XLConnect") library(msm)
library(forecast)
library(markovchain)
library(dplyr)
# Transion rates Q <- rbind( c(0.25 , 0.25, 0.5, 0.25), c(0.25, 0.25,
0.166, 0.25),c(0.25, 0.25, 0.25, 0.25), c(0.25 , 0.25,
```

```
0.5, 0.25))
    # destring site
# View(adultwellness$site1)
adultwellness$site2[adultwellness$site1=="PHRU"] <- 0
adultwellness$site2[adultwellness$site1=="Tintswalo"] <- 1

    Q.crude <- crudeinits.msm(whostates cumulativetime, studyid, data=adultwellness,
qmatrix=Q)
summary(Q.crude)

    adultwellness.msm <- msm( whostates   cumulativetime, subject=studyid, data =
adultwellness, qmatrix = Q,method = "BFGS", control = list(fnscale = 4000, maxit =
10000))
summary(adultwellness.msm)

    # summarising WHO stage data of individuals (the number of times an individuals
moved from one state to the other)
statetable.msm(whostates, studyid, data = adultwellness)

    # model1 for transitions
adultwellness.msm1 <- msm( whostates   cumulativetime, subject=studyid, data =
adultwellness, qmatrix = Q,method = "BFGS", control = list(fnscale = 4000, maxit =
10000))
adultwellness.msm1

    # Transition probability matrices
pmatrix.msm(adultwellness.msm)
```

# projection(if the patients where followed for 2002 to 2010 as was proposed in the study)
pmatrix.msm(adultwellness.msm, t= 8)

G.2 Stata codes for Adult Wellness data

use medelsdata, clear

*****Variables selection stepwise, pr(.1) lockterm1: ologit whostates (clinicaltotal Gender site1) ARVTreatment CPTTreat edulev Smoked EstAge1 salaried MaritalStatus , or

***overal model for both sites
xtset studyid VVDDate

***Multilevel fixed effects ordered logistic regression
xtologit whostates i.clinicaltotal i.Gender i.site1 i.ARVTreatment i.CPTTreat i.edulev i.Smoked i.salaried
quadchk
xtologit whostates i.clinicaltotal i.Gender i.site1 i.ARVTreatment i.CPTTreat i.edulev i.Smoked i.salaried , or
****
****using variables the researcher selected from above fit a bayesian model

bayes, or saving(/*bayesimputeddata*/) mcmcsize(7500) burnin(2500) thinning(7) rseed(1976444) block(whostates: i.clinicaltotal) block(whostates: i.Gender) block(whostates: i.site1) block(whostates: i.ARVTreatment) block(whostates: i.CPTTreat) block(whostates: i.edulev) block(whostates: i.Smoked) block(whostates: i.salaried) blocksummary: meologit whostates

```
i.clinicaltotal i.Gender i.site1 i.ARVTreatment i.CPTTreat i.edulev i.Smoked i.salaried ||
studyid: bayesgraph diagnostics all
bayesstats ess

    ************************************imputing dataset
mi set mlong
mi register imputed Gender ARVTreatment CPTTreat edulev Smoked salaried
mi impute chained (logit) Smoked CPTTreat ARVTreatment Gender salaried (ologit)
edulev , add(5) rseed(1976444)

    *******saving imputed dataset
*save imputeddatasetTNTPHRU, replace
use imputeddatasetTNTPHRU,clear

    ***running imputed regression model for 5 imputations
mi xtset studyid VVDDate
mi estimate, or cmdok: xtologit whostates i.clinicaltotal i.Gender i.site1 i.ARVTreatment
i.CPTTreat i.edulev i.Smoked i.salaried
```

G.3 Stata Codes for simulated data

```
set obs 15000
gen studyid=int(1+(1000-1)*runiform())
sort studyid

    gen y=int(1+(5-1)*runiform())
gen w=int(1+(3-1)*runiform())
```

```
replace w = 0 if w == 1
replace w = 1 if w == 2

/*Create body mass index*/
gen d=float(18+(30-18)*runiform())
gen m=int(1+(5-1)*runiform())
**************simulated models complete case
xtset studyid

save completedatasimu
use completedatasimu, clear

***Multilevel fixed effects ordered logistic regression with complete data
xtologit y d i.w i.m , or

************************************************* **********generating values i will use
to create missing data
gen v=int(1+(5-1)*runiform())
gen f=int(1+(5-1)*runiform())
gen r=int(1+(5-1)*runiform())

*label drop all
*****generating missing values
replace w =. if v == 3
replace m =. if f == 1
replace d =. if r == 2

save missingdatasimu
```

```
use missingdatasimu25, clear

tabmiss w m d

***Multilevel fixed effects ordered logistic regression with missing data
**************************************** ****using variables the researcher selected from
above a bayesian model bayes, or saving(bayesimputeddatasetsimulated) mcmcsize(7500)
burnin(2500) thinning(7) rseed(19764442) block(y: i.w) block(y: i.m) block(y: d)blocksummary:
meologit y i.w i.m d || studyid: bayesgraph diagnostics all
****************************************************
***imputing dataset mi set mlong mi register imputed w m d
mi impute chained (regress) d (logit) w (mlogit) m, add(5) rseed(19764442)
*******saving imputed dataset save simulatimputeddataset, replace
use simulatimputeddataset,clear

***running imputed regression model for 5 imputations
mi xtset studyid
mi estimate, or cmdok: xtologit y i.w i.m d, or

***************************************************************************** *creating dataset
with a missingness of about 50 percent use completedatasimu, clear
gen v=int(1+(3-1)*runiform())
gen f=int(1+(3-1)*runiform())
gen r=int(1+(3-1)*runiform())

*label drop all
*****generating missing values
replace w =. if v == 1
```

replace m =. if f == 2

replace d =. if r == 1

*save missingdatasimu50

use missingdatasimu50, clear

tabmiss w m d

bayes, or saving(bayesimputeddatasetsimulated1)

mcmcsize(7000) burnin(2500) thinning(7) rseed(19764442)

block(y: d) block(y: i.w) block(y: i.m) blocksummary:

meologit y d i.w i.m || studyid: bayesgraph diagnostics all

** ****imputing dataset

mi set mlong

mi register imputed w m d

mi impute chained (regress) d (logit) w (mlogit) m, add(5) rseed(19764442)

*******saving imputed dataset

*save simulatimputeddataset, replace

use simulatimputeddataset,clear

***running imputed regression model for 5 imputations

mi xtset studyid

mi estimate, or cmdok: xtologit y d i.w i.m, or