LANCASTER UNIVERSITY

Modelling survival in HIV cohorts with applications to data from Zomba, Malawi

by

Emmanuel Mwamlima Singogo

Submitted for the
degree of Doctor of Philosophy

September 2016

Declaration of Authorship

This is to declare that the work in this thesis has been done by Emmanuel Mwamlima Singogo and has not been submitted elsewhere for the award of a higher degree. All published and unpublished resources used in this work have been duly recognised in references section.

Signature:.....

Date:.....

Emmanuel Mwamlima Singogo

A single death is a tragedy; a million deaths is a statistic.

Joseph Stalin

LANCASTER UNIVERSITY

Abstract

Faculty of Health and Medicine

Lancaster Medical School

Doctor of Philosophy

Emmanuel Mwamlima Singogo

The Human Immunodeficiency Virus (HIV) pandemic still remains a major public health concern worldwide. The World Health Organization (WHO) estimates that approximately over 70% of people living with HIV in the world are in sub-Saharan region. Malawi is one of the worst affected countries in sub-Saharan Africa with prevalence reaching up to 16% in some areas. Recent study reports, largely in Africa, comparing outcomes for HIV patients with Kaposi's sarcoma (HIV/KS) and HIV patients without KS indicate poor prognosis and poor health outcomes amongst HIV patients with KS. While efforts are being made to improve the management and care for the HIV/KS patient group, there is also need for continued efforts to better understand the survival patterns in this patients. The work presented in this thesis attempts to investigate the survival patterns in different patient subgroups in HIV cohorts in Malawi by using advanced and novel statistical techniques with an ultimate aim of informing targeted patient treatment and management practices.

In this thesis, we aim to address the following four objectives; (1) to identify risk factors for mortality among HIV patients diagnosed with Kaposi's sarcoma during routine initiation of ART, (2) to model the survival pattern among HIV patients diagnosed with KS, (3) to model local geographical variations in survival among HIV patients on ART, (4) to quantify transition dynamics in HIV and TB co-infection using multi-state modelling.

For the first two objectives, we considered extended Cox models and parametric models. We also used a novel approach of accounting for high attrition in cohorts in which we used a 'gold-standard' data to compare survival in our cohort. Sensitivity analyses indicated consistencies in our approach providing an insight into how model results change when using this comparison approach. Overall We noted an early mortality with most patients dying in the first five months after starting HIV treatment. Patients with TB and the patients who started in the early era of ART were significantly at risk of dying. The model diagnostics indicated that (i) a random effects Cox/Log-Gaussian frailty model and (ii) a flexible parametric proportional hazards model, describe the risk of mortality in the HIV/KS patients well.

For the third objective, spatial survival models were considered. The study showed existence of possible residual spatial variation in survival after adjusting for age, sex, KS status, TB status and unobserved individual frailties. To further aid our understanding, we used the choropleth maps to indicate areas with substantially high probability of mortality risk at different cut-off values. These results highlight the local geographical variations in survival in HIV populations, an element more often ignored in most studies on HIV data.

For the last objective, we considered the homogeneous continuous time multistate Markov models. In this study we found that patients in TB free status had a relatively higher probability of transitioning to being diagnosed with TB compared to dying while in TB free status. However, the cumulative transition hazards for the 'TB free $\rightarrow$ death' transitions compared to the "TB free $\rightarrow$ TB infection" transitions were only higher during the early days of HIV treatment. This result emphasize how early periods after starting HIV treatment is crucial to ensure better prognosis. We also noted significant gender differences in the 'TB-free $\rightarrow$ death' transitions.

It is anticipated that the findings in this thesis will help to inform treatment and management practices of HIV patients. The findings provide clear outcome pathways taken by HIV/TB patients before experiencing a terminal outcome. More

importantly, the findings could help inform policies aimed at improving overall survival in HIV cohorts by establishing targeted patient management and treatment strategies and also formulating a more efficient triage system for care and treatment of particular group of patients.

Acknowledgements

I greatly thank my supervisors Dr Benjamin Taylor, Prof Peter Diggle and Dr Thomas Keegan for their invaluable guidance and support throughout the execution of this work. Many thanks to Dignitas International and Ministry of Health in Zomba, Malawi for allowing me to use the district HIV cohort data for my studies.

I also greatly thank the Economic and Social Research Council(ESRC) UK and North West Doctoral Training Centre (NWDTC) for fully funding my PhD studies at Lancaster University. Without this funding, I could not experience the great academic and social support provided by the faculty members of this vibrant academic environment. Lastly, I also thank the external examiner Prof Paul Lambert (Leicester) and internal examiner Dr Jonathan Read for providing insightful inputs and comments during the viva.

Contents

Declaration of Authorship i

Abstract iii

Acknowledgements vi

List of Figures xii

List of Tables xiv

List of Planned Papers xvi

Dedicatory xvii

I Introduction, Study Setting and Literature Review 1

1 Thesis Introduction 2

2 Research proposal and Data description 5

2.1 Motivation Statement 6

2.2 Research Questions and Objectives 7

2.3 Study Design 8

2.4 Study Outcomes 9

2.5 Study setting and population 9

2.6 Data Description and management 10

2.6.1 Ethics Clearance and Study Oversight 13

References I/2 **15**

3 Literature review of HIV, KS and TB Epidemiology **18**

3.1 HIV Epidemiology 19

3.2 Epidemiology of HIV-related Cancer 22

3.3 HIV and TB Epidemiology 27

3.4 Loss to follow-up in HIV cohorts 29

References I/3 **32**

4 General statistical review of survival analysis methods **39**

4.1 Introduction to survival analysis 40

4.1.1 Right Censoring 41

4.1.2 Density and Distribution Functions 43

4.1.3 Survival function 43

4.1.4 Hazard Function 45

4.2 Models for Survival Data 46

4.2.1 Proportional Hazards Model 47

4.2.1.1 Semi-Parametric PH Models 47

4.2.1.2 Parametric PH Models 49

4.2.2 Accelerated Failure Time Models 54

4.2.2.1 Weibull Model 55

4.2.2.2 Log-logistic Model 56

4.2.2.3 Log normal Model 57

4.2.2.4 Generalized Gamma Model 57

4.3 Extensions to Standard Survival Models 58

4.3.1 Frailty Models 58

4.3.2 Spatial Survival Modelling 60

4.3.2.1 Covariance Functions 62

4.3.3 Competing risk and Multi-state Modelling 64

4.3.3.1 Competing Risk Modelling 65

4.3.3.2 *Cumulative incidence function* 67
4.3.3.3 Fine and Gray Model 68
4.3.3.4 Multistate Modelling 70
4.4 Model Diagnostics and Selection 75
4.4.1 Testing the PH Assumption 76
4.4.2 Residual Analysis 77
4.4.3 Model Selection 78

References I/4 **83**

II Modelling survival in HIV cohorts **89**

5 Estimating survival in HIV cohorts **90**
5.1 Introduction 92
5.2 Objectives 94
5.3 Data Description 95
5.3.1 Inclusion and Exclusion Criteria 99
5.4 Survival Modelling 99
5.4.1 Semi-parametric Models 99
5.4.2 Parametric models 100
5.5 Descriptive Analysis 101
5.5.1 Overall Survival 101
5.5.2 Prevalence of KS at ART initiation 105
5.5.3 Age 106
5.5.4 Sex 107
5.5.5 Period of starting HIV Treatment 110
5.5.6 Distance to HIV clinic 111
5.5.7 TB Status 113
5.5.8 Occupation 114
5.6 Accounting for loss to follow-up 116
5.7 Results from Semi-parametric and Parametric Models 121
5.7.1 Risk factors of loss to follow-up 121

5.7.2 Risk factors of mortality and survival . . . 122
5.7.3 Describing patient heterogeneity . . . 123
5.7.4 Sensitivity Results . . . 125
5.8 Model Diagnostics and Selection . . . 130
5.8.1 Checking the Proportional Hazards Assumption . . . 130
5.8.2 Comparing Cox Models . . . 130
5.8.3 Comparing Parametric Models . . . 132
5.9 Discussion and Conclusions . . . 133

References II/5 **141**

6 Spatial variation in Survival **147**
6.1 Introduction . . . 149
6.2 Objectives . . . 150
6.3 Data description . . . 150
6.4 Methods . . . 152
6.4.1 Model Diagnostics and Comparison . . . 156
6.5 Descriptive Analysis . . . 156
6.5.1 Mapping deaths, active follow-up and lost to follow-up . . . 157
6.5.2 Distribution of patients by ART period . . . 160
6.5.3 Mapping TB Episodes and KS Diagnosis . . . 162
6.6 Spatial Survival Model Results . . . 163
6.6.1 Model Comparison . . . 163
6.6.2 MCMC Diagnostics . . . 165
6.6.3 Estimates of model parameters . . . 168
6.6.4 Exceedance Probability Maps . . . 169
6.7 Discussion . . . 173

References II/6 **177**

7 Multistate Models for TB **179**
7.1 Introduction . . . 181
7.2 Objective . . . 183

7.3 Recap: HIV and TB co-infection . . . 183
7.4 Data Description . . . 184
7.5 Multistate modelling . . . 185
7.6 Model formulation and Parameter Estimation . . . 187
7.7 Descriptive Analysis . . . 190
7.8 Model results and Prediction . . . 194
7.8.1 Transition Intensities . . . 194
7.8.2 Transition Probability . . . 196
7.8.3 Transition Hazards Ratios . . . 203
7.9 Model Assessment . . . 205
7.10 Discussion and Conclusion . . . 208

References II/7 **216**

III General Discussion and Further Research **223**

Complete Bibliography **227**

List of Figures

2.1 Map of study area, Zomba district 10

3.1 Global HIV estimates 19
3.2 Death trends in patients on HIV treatment in Malawi 22
3.3 Death trends among WHO stage 4 patients 26
3.4 National rates of attrition in ART cohorts in Malawi 31

4.1 Competing risks R_i. The direction of arrows denotes "'competing to"' the Event of interest. 65
4.2 The illness-death model 71

5.1 Treatment Initiation at Zomba HIV clinic 96
5.2 Empirical distribution functions 104
5.3 Overall Survival curves 105
5.4 Kaposi's sarcoma prevalence during ART initiation 106
5.5 Plots for mean age by patient status 107
5.6 Density distribution and mean age by sex 108
5.7 Density distribution of age grouped by sex 110
5.8 Survival curves for males and females 110
5.9 Survival curves by period of ART 112
5.10 Survival curves by distance 113
5.11 Survival curves by TB history 114
5.12 Survival curves by TB history 115
5.13 Adjusted K-M survival plots 120
5.14 Sensitivity analysis plots for exponentiated mean β based on 1000 permutations of the data. 129
5.15 Diagnostic plots from Cox models using the Cox-Snell residuals. . . 132

5.16 Comparing Survival functions 133
5.17 Model: Cumulative hazards plots 134

6.1 Quadrat counts of patients 157
6.2 Zoomed Map of Lake Chilwa 158
6.3 Map of deaths, LTFU and active follow-up 160
6.4 Mapping ART patients enrolled in 2004-2007 & 2008-2011 161
6.5 Mapping TB episodes and KS cases 163
6.6 MCMC diagnostic plots 166
6.7 Hazard and correlation trace and autocorrelation plots 167
6.8 Trace and autocorrelation plots for fixed effects 167
6.9 Exceedance Probability Maps 172

7.1 TB flow diagram for different states 187
7.2 Density and Probability plots of surviving TB 193
7.3 Density for age and sex 193
7.4 Transition to TB versus to death from TB free state 201
7.5 Comparing transition probabilities to death 202
7.6 Cumulative transition hazards of death 205
7.7 Comparing the observed and expected number of individuals in each state 207

List of Tables

3.1 Mortality risk factors in HIV+KS patients 27
3.2 Proportion of deaths among LTFU HIV patients 31

4.1 Data preparation for a simple right-censored survival analysis . . . 42
4.2 Data preparation for competing risk modelling 65
4.3 Data preparation for Multistate modelling 71
4.4 Common expressions in Multistate modelling 75

5.1 List of Covariates as defined in section 2.6 98
5.2 Baseline demographic of patient 102
5.3 Results from Log-rank tests 109
5.4 Semi-parametric Estimates for Loss to follow-up 124
5.5 Semi-parametric Estimates for death 125
5.6 Parametric Estimates 126
5.7 Flexible Parametric Estimates 127
5.8 Final adjusted HRs using the comparison method 128
5.9 Test statistics for linear trend using the Schoenfeld residuals in a standard Cox PH model. 131
5.10 Model selection for Cox Models 131
5.11 Model selection ofr parametric models 133

6.1 Comparing Weibull spatial survival models 164
6.2 Weibull Spatial Survival model 168

7.1 Summary of state transitions 191
7.2 List of possible combination of transitions 192
7.3 Estimates of Baseline Transition Rates 195

7.4 Estimates of Transition Probabilities 200
7.5 Probability of next immediate Outcome 202
7.6 Estimates of Transition Hazards Ratios 204
7.7 Model diagnostics using Pearson-type test 208

List of Planned Papers

Paper 1. **Quantifying mortality risk in HIV patients diagnosed with Kaposis's sarcoma in Zomba, Malawi** - *based on chapter 5*

Contribution: lead author, statistical analysis and writing of the paper.

Journal : *International Health*, Accepted, 26^{th}May, 2017

Paper 2. **Local spatial variations in survival among patients routinely started on a lifelong HIV treatment in Zomba, Malawi** - *based* on chapter 6

Contribution: lead author, statistical analysis and writing of the paper.

Target Journal : *Journal of Spatial and Spatio-temporal Epidemiology*

Paper 3. **The contribution of multistate modelling in understanding TB epidemiology in antiretroviral therapy (ART) cohorts** - *based on chapter 7*

Contribution: lead author, statistical analysis and writing of the paper.

Target Journal : *The Journal of Infectious Diseases*

To my wife Wezie and daughter TUSA

Part I

Introduction, Study Setting and Literature Review

Chapter 1

Thesis Introduction

The work in this thesis was motivated by the growing challenges and burden of the HIV pandemic, especially in the resource-limited settings. Though HIV and AIDS is a global disease, but recent program progress reports by The Joint United Nations AIDS Program (UNAIDS) indicate disproportional burden of morbidity and AIDS-related deaths, with the sub-Saharan Africa region being worst affected. In the light of this, our work seeks to contribute to the overall understanding of survival and mortality in HIV patients with application to data from Zomba, Malawi. Furthermore, we also study and quantify the impact of co-infection from other disease such as Kaposi's sarcoma (KS) and Tuberculosis (TB) on HIV patients that are initiated on a life-long HIV treatment (the antiretroviral therapy, ART).

Our work in this thesis is divided into three main parts; Part I - the introductory chapters (research proposal, epidemiological and statistical methods reviews), Part II comprising of three epidemiological studies on survival and mortality in

HIV patients routinely enrolled in HIV treatment program in Zomba, Malawi and Part III which comprises of a summary discussion and conclusions.

In Part I, there are three chapters, this thesis introduction and outline being the first chapter. In Chapter 2, we describe the overall objectives, provide a detailed description of all the datasets used in this thesis including definitions of variables. The introductory part also comprises of detailed literature review on the epidemiology of diseases studied in this work; HIV/AIDS, KS, and TB. In Chapter 4, we end this first part by providing a concise introduction to the all statistical methods used in this thesis. The focus in this review is on the methods used in the analyses in this work and references are provided for further details on mathematical proofs and computation procedures.

In Part II, we present three (3) separate epidemiological studies based on our HIV data and the summary of discussions and conclusions from these three analyses. In Chapter 5, the objective is to quantify the mortality risk among HIV patients on ART using the patient covariates available in our the dataset. To do this, we use standard survival methods to study mortality in a subgroup population of ART patients, focusing on HIV patients that are also diagnosed with KS at the time of starting the HIV treatment. In Chapter 6, we extend the non-spatial frailty models fitted in Chapter 5 by replacing the patient-level frailties with the spatial frailties. The aim in this chapter is to study how survival can vary geographically and to highlight the use of spatial statistics in understanding disease epidemiology. We use spatial survival methods for point pattern data to study the hazards of death in the general HIV cohort in our study region (regardless of KS

and TB). This analysis could potentially help health planners and policy-makers in identifying areas with high risk of death as well as identifying areas requiring extra resources to improve access to health services. Lastly in Chapter 7, the focus is on quantifying the transition probabilities between two pairs of TB states among patients diagnosed with TB in addition to the HIV/AIDS. The data in this analysis are commonly categorised as multistate, and we take advantage of recent analysis tools for analysing these data (especially those available as open-source R-Software packages).

In Part III, we provide a general summary of discussions and conclusions of results presented from Chapter 5 to Chapter 7. Based on our results and experience in this work, we also suggest potential areas of further research that could be considered: both methodological and epidemiology research.

In conclusion, through out this thesis, we have made efforts to address the items listed in the main reporting guidelines for epidemiological studies, with emphasis on cohort studies. We solely used STROBE guidelines and our efforts are documented in the appendix by making reference to the page numbers where a STROBE item was addressed or reported. A STROBE checklist for each chapter is available in the appendix.

Chapter 2

Research proposal and Data description

Summary

This chapter contains the research proposal and a comprehensive review of the literature on HIV and HIV-related diseases with emphasis on cancer and tuberculosis (TB). We also summarize main methods used to address each objective although these methods are separately reviewed and discussed in detail later in chapter 4 and subsequent chapters.

2.1 Motivation Statement

The acquired immunodeficiency syndrome (AIDS) which is a progressive Human Immunodeficiency Virus (HIV), is one of the main leading causes of mortality globally (UNAIDS, 2016). Africa has the highest proportion of HIV-infected populations with over 70% of the global HIV population living in the sub-Saharan region. The Joint United AIDS programme estimates about 1.1 millions AIDs-related deaths globally in 2015 alone, of which 74% were in Africa (UNAIDS, 2016).

Due to the compromised immunity, patients with HIV become vulnerable to many opportunistic illnesses (CDC, 2016). Some of the most common opportunistic illnesses in HIV patients are TB and cancer. With a rapidly growing era of provision of a lifelong HIV treatment the highly active antiretroviral therapy (ART), recent studies have shown that provision of this treatment has a preventive effect on opportunistic diseases (Velásquez et al., 2015; Johansson, Robberstad and Norheim, 2010). As such, it is recommended that all patients suffering from these opportunistic illnesses are tested for HIV and started on HIV treatment for the remainder of their lives.

However, while on treatment there are many factors that may influence prognosis in HIV patients and consequently increase their mortality risk. These factors include patient behavioural and biological characteristics, multiple co-infections, treatment-seeking and adherence to medication prescription, among others. For

example, studies by Takarinda et al., (2015), Vijay et al., (2011), Chu, Mahlangeni et al., (2010) and Mwinjiwa et al., (2013) and Makombe, Harries and al, (2008) highlight the disproportional co-morbidity burden and poor patient outcomes among HIV patients with KS.

In the light of the above, there is a substantial interest in conducting research in HIV programmes in order to constantly evaluate treatment and care practices. The work in this thesis was motivated on this basis: to make a contribution to scientific understanding of these opportunistic diseases in ART cohorts. Our data focuses on the data from the sub-Saharan African country of Malawi.

2.2 Research Questions and Objectives

In this project, we aim to make use of advanced statistical analysis methodologies to help better understand Kaposi's sarcoma (KS) and Tuberculosis (TB) epidemiology in HIV-infected patients. The main research questions to be investigated are;

1. What are the risk factors for mortality amongst HIV/KS patients?

2. Can spatial statistical methods add to our understanding of the epidemiology of KS among HIV patients by characterising the geographical distribution? Do geographical determinants unexplained part of the observed variation in risk?

3. Which state transition paths are common in patients that are diagnosed with TB in addition to HIV?

Specifically, we will address the following objectives in three separate chapters;

1. Understand the relationship between individual risk factors for mortality amongst HIV/KS patients. Also, we will provide an approximate lower bound on the proportion of patients recorded as lost to follow-up that are really deaths.

2. Adjusting for individual-level covariate data, investigate the geographic variation in survival prognoses in the Zomba district.

3. Using multistate survival models, investigate the different infection pathways for individuals with both TB and HIV.

2.3 Study Design

This is a retrospective cohort study. We used the HIV patient data that was routinely captured at an ART clinic in Zomba Malawi from 2004 to 2014. We only considered patients that started ART from 2004 to 2011 but followed them up to end 2014. Additionally, a spatial design (point pattern) for survival outcomes was also considered.

2.4 Study Outcomes

In order to answer the research questions above, the models to be fitted will use the following as response variables (outcomes);

- Survival (dead or alive) amongst HIV patients with KS.
- Time to death amongst HIV patients with KS (<5yrs, >=5yrs after HIV treatment initiation).

2.5 Study setting and population

In 2008, our the study area Zomba district was estimated to be home to a population of approximately 700,000 with over 90% of people living in rural areas (Malawi NSO, 2008). The district has one of the highest adult HIV prevalence (around 16%) in Malawi. Since 2004, the Malawi Ministry of Health has scaled up ART provision and decentralization of most HIV services such as HIV testing, prevention of transmission of HIV virus from pregnant and breastfeeding mother to a child. This has resulted into increase increased early HIV testing and uptake of ART services (Malawi Ministry of Health, 2011).

By the end of 2011, over 18, 000 HIV positive patients were cumulatively enrolled in HIV care in the district from 2004, of which approximately 3% are estimated to either have already, or later develop cancer (with KS being common) (Mwinjiwa et al., 2013; Malawi Ministry of Health, 2011). In this study, we used data from

all adult HIV patients diagnosed with KS enrolled in HIV care between 2004 and September 2011 under routine programme of HIV treatment provision. By the end of 2011, the adult prevalence of tuberculosis (HIV-TB co-infection) in this ART cohort was at around 15% (calculated from the our data). Figure 2.1 shows the map of study district, Zomba district and its location on the map of Malawi.

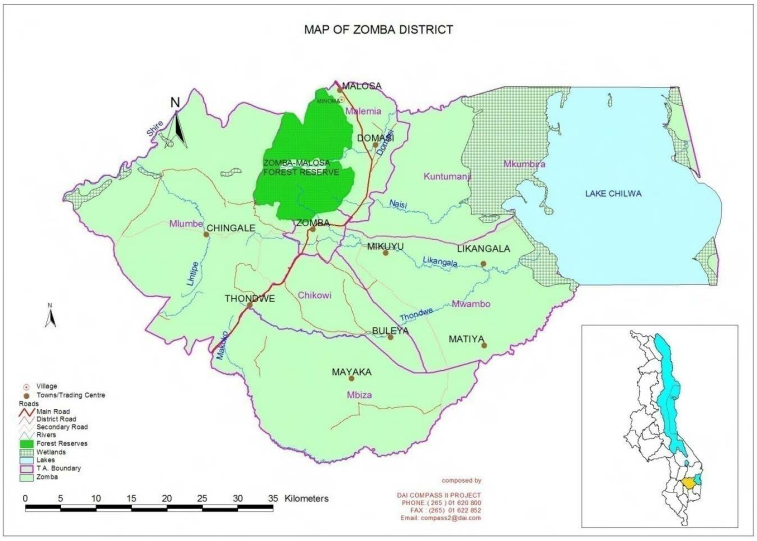

Figure 2.1: The Map of Zomba district. On the bottom right corner is the Map of Malawi, with Zomba district in yellow colour [*Source: Zomba District Socio-Economic Profile Report(2009-2002)*].

2.6 Data Description and management

There are two types of datasets used in this work. The main data was sourced from Dignitas International (an international non-government organisation) in collaboration with the Malawi Ministry of Health, Zomba district office.

The cohort data from Zomba Malawi was routinely collected in an HIV programme from 2004 to 2014. This is the data of all HIV patients that were initiated on ART

after testing positive to HIV. By end of 2011, the main cohort data comprised 18, 275 HIV positive patients (cumulatively) that were enrolled in the HIV clinic care in Zomba from during this period. Three percent (615) of these patients were diagnosed with Kaposi's Sarcoma (skin tumour) during the screening process to determine ART eligibility (Malawi Ministry of Health, 2011). The following variable were available;

- **Recruitment Date**. This is the date at which HIV patients were initiated on a lifelong ART. This variable was categorised in two periods: 2004-2007 and 2008-2011. This is because most ART scale-up activities started after late 2007 and we wanted to compare patient outcomes before and after scale-up of ART in the district.

- **Age**. This is defined as recruitment or baseline age, i.e age (in years) at which patients started HIV ART. With the aging effect, the age of a patients is an important factor and has been used in many studies to adjust any confounding effect on the outcome of interest (Chu, Misinde et al., 2010; Chu, Mahlangeni et al., 2010)

- **Sex**. Sex of patients as declared at recruitment time. Two sexes were recorded as either male or female. Some studies have reported gender differences in survival in patients taking ART (Takarinda et al., 2015).

- **Name of village/traditional authority (TA)**. The name of a village where patients were residing at the time of ART initiation. For confidentiality issues, geo-coordinates for special features near their residences were used

such as home of village head person within each village, churches, market places, schools, dams, hospitals, military and police stations. The data on the geo-coordinates were obtained from the National Statistical Association of Malawi and were matched to each patient's address as indicated on their clinic form.

- **Distance**. Using the village coordinates, we calculated the euclidean distance from the village point to Zomba HIV clinic. This variable is interesting in two perspectives; the issue of accessibility to the clinic as a predictor of survival, and spatial variation in survival outcomes.

- **Occupation status**. The type of occupation at ART initiation time. This had the following categories; student, self-employed (includes farmers and business persons), employed (teachers, health care workers, security personnel etc) and other (occupation not specified at baseline). We hypothesised that different occupation status have different exposure to risk of HIV progression and we wanted to quantify this risk in our cohort.

- **TB status**. This information was updated at every HIV clinic visit. The possible status categories recorded were; TB-free/ no TB, with TB and on TB treatment, TB treatment completed but awaiting ascertainment of TB status and also death. The patients that are categorised as "'TB completed"' are those patients whose the TB test results after completing TB treatment was inconclusive. We considered all forms of TB whether pulmonary TB (PTB), Extra pulmonary TB (EPTB), or multi-drug resistant (MDR) TB. Data on TB diagnosis tool (whether smear or X-ray etc) was not captured

in data collection tools which were used in HIV clinics but this information was available in TB clinics. The challenge was that it was impossible to link and merge data from TB and HIV clinics as they were collected separately. There is documented evidence on the link between TB and HIV/AIDS and this was included in the models as a confounder (Fenner et al., 2013; Collins et al., 2010).

- **ART outcome**. This information was updated at each HIV clinic at the end of every three months. The possible outcomes recorded were; alive (active follow-up), dead and lost to follow-up.

- **Time**. This was defined as time from the recruitment date to the date of ART outcome in chapter 5 and chapter 6. This definition was changed in chapter 7 as time to the different TB status as considered.

2.6.1 Ethics Clearance and Study Oversight

We obtained ethics clearance to use patient data from the the Faculty of Health and Medicine Research Ethics Committee (FHMREC) at Lancaster University and also from the National Health Sciences Research Committee (NHSRC - 1278), Lilongwe Malawi. The data were stored in password protected folders and access was restricted to the researcher only.

As Principal Investigator, the research student had primary responsibilities for the study with support from three academic supervisors in the department of

Lancaster Medical School at Lancaster University: Dr Benjamin Taylor, Professor Peter Diggle and Dr Thomas Keegan.

References I/2

CDC (2016) *Living with HIV: Opportunistic Infections*, http://www.cdc.gov/hiv/basics/livingwithhiv/opportunisticinfections.html.

Chu, K.M., Mahlangeni, G. et al. (2010) 'AIDS-associated Kaposi's sarcoma is linked to advanced disease and high mortality in a primary care HIV programme in South Africa.', *Journal of the International AIDS Society* 13, p. 23.

Chu, K., Misinde, D. et al. (2010) 'Risk factors for mortality in AIDS-associated Kaposi sarcoma in a primary care antiretroviral treatment program in Malawi.', *International health* 2 (2), pp. 99–102.

Collins, J. et al. (2010) 'Effect of antiretroviral therapy on survival of HIV-infected tuberculosis patients in Peru', *Rev Panam Infectol* 12 (2), pp. 37–43.

Fenner, L. et al. (2013) 'Tuberculosis and the risk of opportunistic infections and cancers in HIV-infected patients starting ART in Southern Africa.', *Tropical medicine & international health : TM & IH* 18 (2), pp. 194–8.

Johansson, K.A., Robberstad, B. and Norheim, O.F. (2010) 'Further benefits by early start of HIV treatment in low income countries: Survival estimates of early versus deferred antiretroviral therapy', *AIDS Research and Therapy* 7, pp. 1–5.

Makombe, S.D., Harries, A.D. and al, E. (2008) 'Outcomes of patients with Kaposi's sarcoma who start antiretroviral treatment under routine program conditions in Malawi', *Tropical Doctor* 38, pp. 5–7.

Malawi Ministry of Health (2011) *Guidelines for the use of Antiretroviral Therapy in Malawi*, tech. rep., Malawi Ministry of Health, Lilongwe http://www.who.int/hiv/pub/guidelines/malawi_art.pdf, Accessed: 2013-11-25.

Malawi NSO (2008) *National Population Census 2008*, tech. rep., Malawi National Statistical Office(NSO), Zomba http://www.nsomalawi.mw/index.php?option=com_content&view=article&id=106&Itemid=6, Accessed: 2013-12-10.

Mwinjiwa, E. et al. (2013) 'Burden, characteristics, management and outcomes of HIV-infected patients with Kaposi's sarcoma in Zomba, Malawi', *Public Health Action* 3, pp. 180–185.

Takarinda, K.C. et al. (2015) 'Gender-related differences in outcomes and attrition on antiretroviral treatment among an HIV-infected patient cohort in Zimbabwe: 2007-2010.', *International Journal of Infectious Diseases* 30, pp. 98–105.

UNAIDS (2016) *FACT SHEET 2016*, tech. rep., Switzerland: The Joint United Nations Programme on HIV/AIDS http://www.unaids.org/en/resources/fact-sheet, Accessed; 2016-04-15.

Velásquez, G.E. et al. (2015) 'Impact of HIV on mortality among patients treated for tuberculosis in Lima, Peru: a prospective cohort study', *BMC Infectious Diseases* 16 (1), p. 45.

Vijay, S. et al. (2011) 'Treatment Outcome and Mortality at One and Half Year Follow-Up of HIV Infected TB Patients Under TB Control Programme in a District of South India', *PLoS ONE* 6 (7), ed. by M. Pai, e21008.

Chapter 3

Literature review of HIV, KS and TB Epidemiology

Summary

In this chapter, we present a comprehensive review of the epidemiology of HIV, KS and TB. The relationship between these diseases are also explained with the use of available literature. The review is focussed on studies and program reports from Africa especially the sub-Saharan region because our data is from this region.

3.1 HIV Epidemiology

The HIV and AIDS pandemic still remains a major public health problem worldwide. Global HIV prevalence has remained at the current rate of at 0.8 % [0.7 - 0.9] for the past five years. There are large regional variations in HIV prevalence (WHO, 2014b). Figure 3.1 shows disproportional burden of HIV with African being worst hit by the epidemic. Other continents with high burden of HIV are America and Europe.

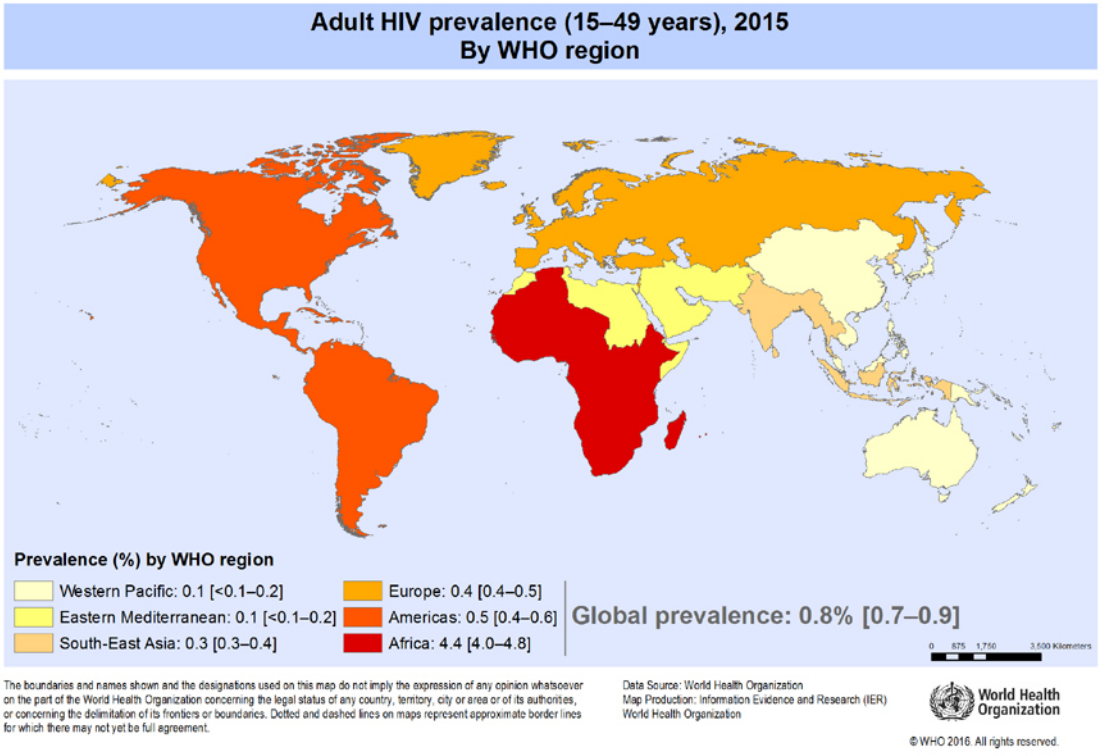

Figure 3.1: Regional variations of adult HIV prevalence[Source: http://www.who.int/gho/hiv/epidemic_status/prevalence/en/, accessed March 2014]

The World Health Organization (WHO) estimates that approximately over 70% of people living with HIV in the world are in the sub-Saharan (SSA) region (World Health Organisation, 2016). Within the SSA region, HIV prevalence is highest in

females in Swaziland and Lesotho up to 37% and 31% respectively . With heavy disease burden and limited resources, most countries in SSA region have weak and fragile health systems. However, in the last decade the region has benefited from the increased funding in HIV programmes targeting low and middle-income countries, currently estimated at over $19 billion (UNAIDS, 2016). Thanks to this huge funding, free HIV treatment ART is now provided in many developing countries but treatment coverage is still below the 2020 treatment goals of 90-90-90 set by the Joint United Nation AIDS Programme in 2015; 90% of HIV positive people tested, 90% of HIV positive people started on ART, 90% of HIV patients on HIV treatment having suppressed viral load (UNAIDS, 2014).

Malawi is one of the countries in southern Africa with high HIV prevalence. In 2008, the Malawi national adult HIV prevalence was estimated at 12% and as high as 23% in rural areas (Malawi NSO, 2008). However, current research by the Joint United Nation AIDS Programme estimates the prevalence in adults 15-49 years to be at 9.1% [8.4% - 9.9%] (UNAIDS, 2015). The most common ways in which the HIV is transmitted in Malawi is through sexual intercourse and through vertical transmission from the mother to a child. Approximately 10% of HIV population in the Malawi are children (Malawi Ministry of Health, 2011).

The provision of free highly active ART in Malawi dates back to 2004, when tertiary government hospitals started offering free treatment the HIV positive patients (Malawi Ministry of Health, 2008). Following the provision of HIV treatment, most HIV related services including HIV testing and counselling, prevention of vertical

transmission of HIV from mothers to babies (pregnant and breast feeding), tuberculosis prevention and treatment, were greatly expanded. The aim of expansion and decentralisation of HIV services was to ensure that HIV infected patients had timely access to treatment and care. A standardised treatment screening tool is used in order to establish patients that are eligible to start ART. For example between 2004 and 2011, HIV patients were started on ART if their CD4 count was less than 250 cells/μL, or if they had TB or one of a list of stage 4 clinical conditions as defined by WHO HIV clinical Staging (Malawi Ministry of Health, 2011; World Health Organisation, 2016).

HIV treatment provides protection against opportunistic illnesses and also boosts patients' immunity. For instance, Makombe, Harries, Yu et al., (2007) reported improved treatment outcomes for HIV patients co-infected with TB. Johansson, Robberstad and Norheim, (2010) reported significant differences in terms of survival between patients that started ART immediately and those that ART was delayed, with patients that started early having an expected net benefit of 14.5 life years per patient. The national HIV programme in Malawi reports a declines in HIV/AIDS-related deaths since the introduction of free ART in 2004. Figure 3.2 shows the gradual decline of all-cause deaths in HIV population in Malawi. However, population level age-specific AIDS mortality rates are not available in Malawi.

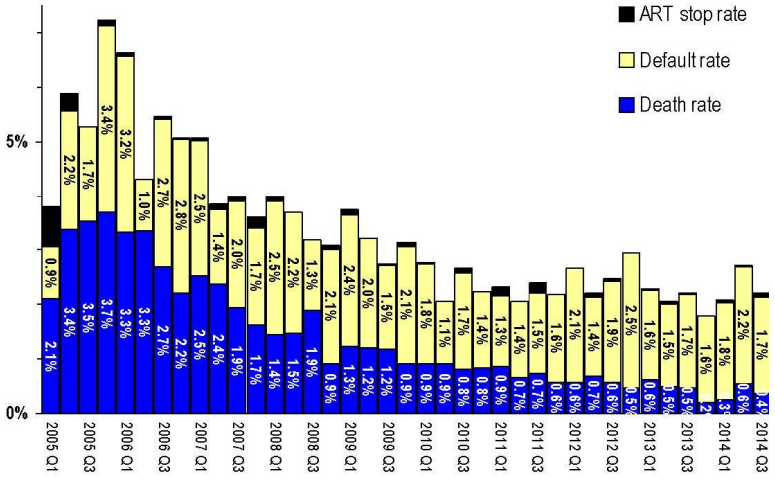

Figure 3.2: National death trends in patients on HIV treatment[Source: (Malawi Minstry of Health, 2014a)]

3.2 Epidemiology of HIV-related Cancer

Cancer is one of the leading causes of death globally, with 13% of deaths attributed to cancer and 63% of these deaths occurring in developing countries. In addition, 56% of global new cases are registered in developing nations (World Health Organization, 2011b). The World Health Organization projects that by 2030, new cases of cancer will increase by 69% from 12.7 million in 2011 to 21.4 million in 2030. In addition, the number of deaths will increase by 72% from 7.6 million to 13.2 million deaths. Despite these overwhelming figures, cancer diseases have received low priority from health care services in Sub-Saharan Africa (Sitas, Parkin et al., 2006). One of the reasons is undoubtedly the overwhelming burden of infectious diseases such as HIV and Malaria which have received undivided attention by governments and NGOs. The continued implementation of expanded HIV/AIDS programs in the SSA has resulted in a considerable drop in HIV/AIDS related deaths (World Health Organization, 2008). With decreases in HIV/AIDS related deaths, there is a steady shift of burden in the SSA region from infectious diseases

to non-communicable diseases such as cardiovascular disease and cancer. This decrease could be due to demographic transition, as there access to health and social services is improved over time. Amongst the common cancer types in Africa are; Kaposi's sarcoma (KS), cervical cancer, lymphomas, oesophageal cancer, liver cancer, breast cancer, lung cancer, and prostate cancer, (Kendiga et al., 2013; Wabing et al., 2000; Chokunonga et al., 1999; Sitas, Bezwoda et al., 1997). The patterns of these cancers vary greatly in different regions within Africa. However due to limited finances and resources, very few African governments have made efforts to curb these diseases as the majority do not have national cancer programs and population-based cancer data . As a result many people die or are diagnosed late when the disease has already progressed (World Health Organization, 2011a).

At the time of this study, population-based cancer data were not available in Malawi which could be used to ascertain national prevalence for different cancer types. Also in a recent nationwide cross-sectional health facility-based survey by Msyamboza et al., (2012) in Malawi, approximately 18, 946 new cancer cases were recorded between 2007 and 2010. Of 18, 946 cases, it was found that the top five common cancers in Malawi were KS (34%), cancer of cervix (25%), oesophagus (12%), non-Hodgkin lymphoma (6%), and urinary bladder (3%). Approximately half (52%) of all cancer cases identified were from southern region, 34% from central region and 13% from northern region (regional population of 45%, 42% and 13% respectively). These reported regional statistics correspond to HIV prevalence figures in these three regions, which are 18%, 10% and 8% in northern, central and southern regions respectively (since majority of cancer types in Malawi are

HIV-related) (Malawi NSO, 2008). Kaposi's sarcoma and cervical cancer were found to be more prevalent in HIV infected populations, consistent with the literature that indicate that with the rise of HIV/AIDS in Africa, KS has become the most frequently reported cancer in many sub-Saharan countries (Rohner et al., 2014). By 2010, there were a total of 84 facilities (including public, research and non-governmental facilities) providing either cancer laboratory diagnosis, or treatment, palliative care, or referral services. Surgery, where undertaken was completed in all referral public hospitals which are: Queen Elizabeth Central Hospital and Zomba Central Hospital in the southern region, Kamuzu Central Hospital in central region, and Mzuzu Central Hospital in the northern region (Msyamboza et al., 2012). Until recently, the majority of cancer treatment in Malawi has been palliative (chemotherapy) and poor patient outcomes for cancer have been reported in sub-Saharan Africa (Kendiga et al., 2013; Sankaranarayanan et al., 2010). For the rest of this section, we review literature on KS epidemiology with a focus on its occurrence in HIV populations.

Kaposi's sarcoma (KS) is a tumour caused by Human Herpes virus 8 (KSHHV) (Minhas and Wood, 2014). The virus is common amongst the people whose immune system have been severely weakened, for example people infected with HIV. Studies have shown that highly active antiretroviral therapy (HAART), a treatment given to HIV positive patients, has preventive effects on KS and it is recommended being the first line treatment for patients (David, 2014). The most common transmission modes for Kaposi's sarcoma Human Herpes-virus (KSHHV) are sexual contract, blood transfusion, and saliva contact (Minhas and Wood, 2014).

The epidemiology of KSHHV varies greatly geographically with more endemic regions being sub-Saharan Africa, the Mediterranean, and north-west China. However, the United States of America and Western Europe have a low prevalence (Minhas and Wood, 2014; Sankaranarayanan et al., 2010; Wu et al., 2014). In the regions where KSHHV is not endemic, the virus is more prevalent among men who have sex with men (MSM). This being the case, there are no comparable studies from this area looking at the KS in the general HIV population because MSM is largely unrecognisable in endemic KS areas. A systematic review proposed by Rohner et al., (2014) is one of the studies reporting the global trend of KSHHV.

Since the introduction of HAART, the incidence of KS has been declining in the general HIV populations and fewer people with advanced conditions are starting HIV treatment; this trend is also observed elsewhere in HIV populations on HAART (Gbabe et al., 2014). The most common treatment for KS in Africa is chemotherapy as opposed to radiotherapy which is more often available in developed countries (Dedicoat, Vaithilingum and Newton, 2003). In their systematic review of African studies, Gbabe et al., (2014) reaffirms the role of HAART and chemotherapy in delaying disease progression in HIV patients with KS. Figure 3.3 shows the decline in the percentages of deaths in patients who are staged in WHO HIV clinical stage 4 in Malawi. Kaposi's sarcoma is one of the conditions in this stage 4. From this figure, we also note a significant declining trend of proportion of patients starting treatment in advanced stage from about 23% in 2004 to 5% in 2014.

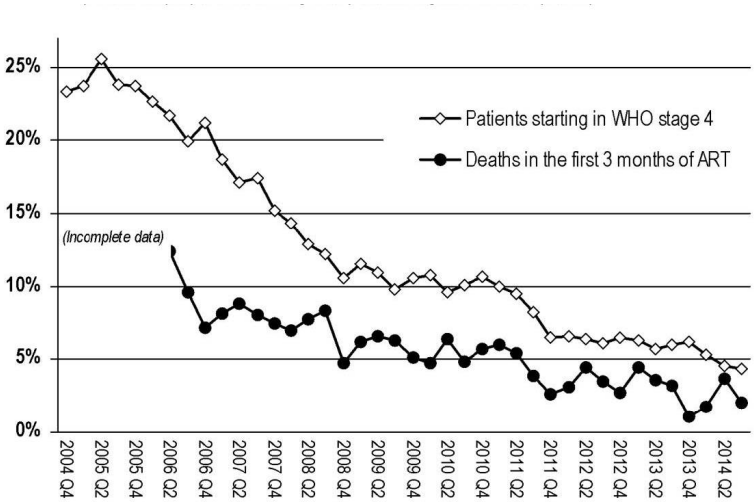

Figure 3.3: Trends in all-cause mortality amongst patients starting HIV treatment in advanced HIV stage (WHO Stage 4)[Source: (Malawi Minstry of Health, 2014a)]

Table 3.1 shows some of the risk factors of mortality in HIV populations with KS. These factors are discussed in detail in studies by Chu, Misinde et al., (2010), Chu, Mahlangeni et al., (2010) and Ziegler et al., (2003). Other factors such as sex and occupation have not been found to be significant although the number of deaths is disproportional in these subgroups (Makombe, Harries and al, 2008; Chu, Misinde et al., 2010).

Table 3.1: Examples of risk factors for mortality in HIV patients with KS.

Risk factors	Description
AIDS	HIV patients with KS progress to AIDS increasing their risk of death (Chu, Mahlangeni et al., 2010)
Advanced KS stage	the risk of mortality is increased with KS stage (Chu, Misinde et al., 2010)
Poor adherence	patients with irregular uptake of treatment (HAART and chemotherapy) have increased risk of mortality (Ngarina et al., 2013)
Low CD4 counts	high mortality in patients with low CD4 counts (Johansson, Robberstad and Norheim, 2010)
Age	older patients tend to have higher mortality risk (Belayneh, Giday and Lemma, 2015)

3.3 HIV and TB Epidemiology

Tuberculosis is common in people with impaired immunity such as HIV positive persons, diabetic, malnourished and tobacco smokers and is one of very the infectious bacterial diseases. The disease is spread through air when an infected person coughs, speaks or sings. It is one of the most common opportunistic infections in persons living with HIV. The recent TB fact sheet by the World Health Organisation indicates that people who are HIV positive are 20 to 30 times more

likely to develop active TB (World Health Organization, 2016). For HIV positive patients, being diagnosed with TB is a sign of progression to AIDS (CDC, 2016). The World Health Organisation reports that TB is a leading killer of HIV patients, with one in every three deaths among HIV patients was due to TB (World Health Organization, 2016).

The global TB treatment success rates range from 30 to 83% but it is much lower in HIV cohorts. HIV positive positive persons receiving ART have reduced risk of developing TB as several studies have reported the protective effective of ART on opportunistic illness, with a reduction in risk of death ranging from 70 - 90 % (Belayneh, Giday and Lemma, 2015; Vijay et al., 2011; Collins et al., 2010; Johansson, Robberstad and Norheim, 2010).

In an effort to combat TB occurrence, several campaigns have been initiated all aiming at eliminating the TB disease. These campaigns include adoption of universal ART provision, WHO End TB strategy which targets 35% reduction in TB cases by 2020 and the UN zero campaigns targeting TB, AIDS, Malaria and poverty among others. In Malawi the TB treatment success rate is estimated at 89% which is higher than the WHO target of 85%. The main reason is for this success is due to the increase in funding the country has received from international grants such as the Global Fund (WHO, 2014a).

3.4 Loss to follow-up in HIV cohorts

Like in other cohort studies, loss to follow-up (LTFU) is still a challenge in HIV studies(Freeman, Semeere and Wenger, 2016). Being lost to follow-up is defined as failing to attend scheduled clinic visits after a defined period of time. This definition varies by country, but in Malawi patients are deemed lost to follow-up if they do not show up after 120 days from a scheduled clinic visit date (Malawi Ministry of Health, 2008; Chi et al., 2011). As a result, there is a growing interest in addressing the impact of loss to follow-up data in estimation of survival in HIV studies. Up to now the methods proposed to address the impact of loss to follow-up (LTFU) on survival estimates have largely depended on the proportion of deaths based on a sample of LTFU patients.

Without a clear picture of mortality in the LTFU patients, analyses ignoring this phenomenon may underestimate survival and mortality rates in such cohorts because the deaths in LTFU patients are not included in the analysis. In absence of active patient tracing, different approaches have been used to improve estimation of mortality in HIV cohorts. These methods include; tracing a random sample of LTFU patients (*tracing method)* and *meta method* (Yiannoutsos, An et al., 2008; Egger et al., 2011; Henriques et al., 2012; Kiragga et al., 2013). These proposed methods are have been largely developed by a South African research initiative on International Epidemiological Databases to Evaluate AIDS (IeDEA) (http://www.iedea-sa.org/). For the tracing approach, a random sample of LTFU patients is usually selected and traced to ascertain their true

outcomes and the final data for analysis is adjusted to reflect an estimate proportion of deaths among LFTU patients. For the meta method, a meta-analysis of LTFU studies is conducted and the resulting estimates of mortality in LTFU patients is used in a developed formula to estimate mortality (see the formula at http://www.iedea-sa.org/index.php?id=2785, accessed; 2016-02-20)

We compiled some of the studies that have been conducted using active tracing of patients and the reported proportions of mortality among LTFU patients in HIV cohorts varied across studies (see Table 3.2 below). It can be noted that studies that traced between 13 and 45 % of LTFU patients reported proportion of deaths ranging between 20% and 60% among traced LTFU patients. On the other hand, studies that had traced at least 50% of lost patients, reported proportion of deaths in this group ranging between 20 - 87%. These statistics point to the fact that in some HIV cohorts, mortality is higher than others thereby limiting the generalization of the *tracing* and *meta method*. We will use these figure later in Chapter 5 to inform our approach of addressing this issue of LTFU in our analysis.

In Malawi despite a decline in HIV deaths, the national LTFU rates have consistently remained around 2% since the start of ART programme in 2004, as shown in Figure 3.4 (here, the LTFU rate is labelled as "'Default rate"').

Table 3.2: Proportion of deaths among LTFU HIV patients

Study	Country	#patients Traced (%)	#Deaths (%)
Kiragga AN, Castelnuovo B, Musomba R et al (2013).	Uganda	406(50.6)	107(26.3)
Julie Henriques, Mar Pujades-Rodriguez, et al (2012)	Malawi	202(34.6)	121(60) vs 10% in non-LTFU
Brinkhof MW, Pujades-Rodriguez M, Egger M (2009)	South Africa (5 studies), Malawi (3 studies), Uganda (2 studies), Zambia, Botswana, Ethiopia, Kenya, Tanzania, and Mali (one study each)	185(73.1) 70(45.4) 173(64.8) 789 46(67.6) 111(13.4) 621(17.6)	127(69) 19(27) 83(48) 359(46) 40(87) 32(29) 124(20)
Yiannoutsos CT, An MW, Frangakis CE et al (2008)	Kenya	621 (54.3)	126(20.1)

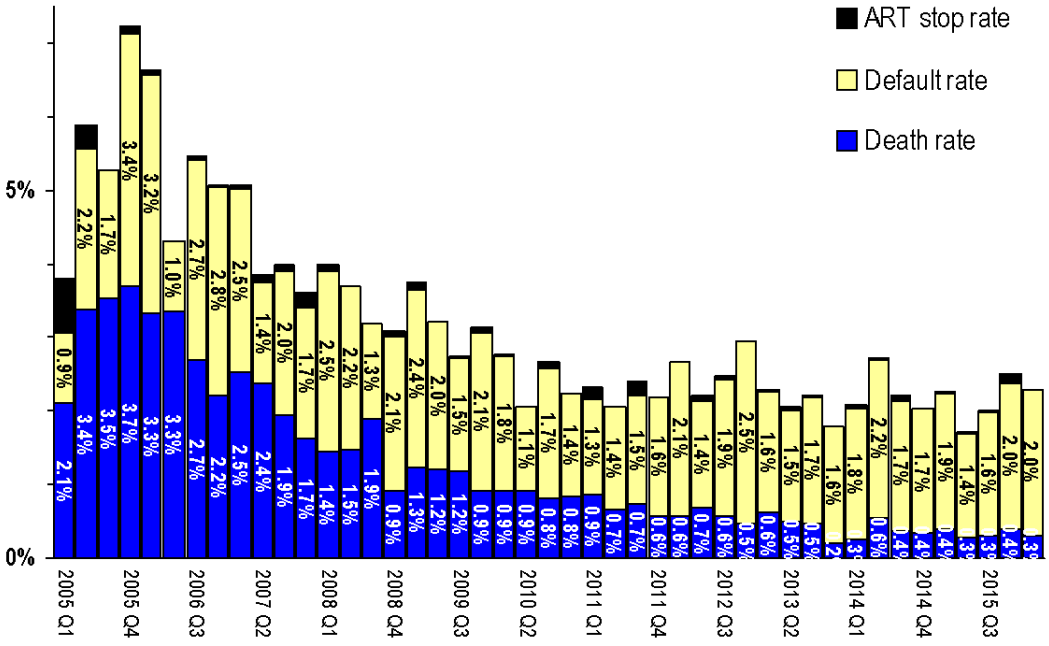

Figure 3.4: Quarterly rates of attrition from ART cohort as composed from those who stop taking ART, patients loss to follow-up and those who die while on ART treatment. This includes funding for different types of TB [Source: (Malawi Minstry of Health, 2016)]

References I/3

Belayneh, M., Giday, K. and Lemma, H. (2015) 'Treatment outcome of human immunodeficiency virus and tuberculosis co-infected patients in public hospitals of eastern and southern zone of Tigray region, Ethiopia.', *The Brazilian journal of infectious diseases* 19 (1), pp. 47–51.

CDC (2016) *Living with HIV: Opportunistic Infections*, http://www.cdc.gov/hiv/basics/livingwithhiv/opportunisticinfections.html.

Chi, B.H. et al. (2011) 'Universal definition of loss to follow-up in HIV treatment programs: a statistical analysis of 111 facilities in Africa, Asia, and Latin America.', *PLoS medicine* 8 (10), e1001111.

Chokunonga, E. et al. (1999) 'AIDS and Cancer in Africa: the evolving epidemic in Zimbabwe', *AIDS* 13, pp. 2583–2588.

Chu, K.M., Mahlangeni, G. et al. (2010) 'AIDS-associated Kaposi's sarcoma is linked to advanced disease and high mortality in a primary care HIV programme in South Africa.', *Journal of the International AIDS Society* 13, p. 23.

Chu, K., Misinde, D. et al. (2010) 'Risk factors for mortality in AIDS-associated Kaposi sarcoma in a primary care antiretroviral treatment program in Malawi.', *International health* 2 (2), pp. 99–102.

Collins, J. et al. (2010) 'Effect of antiretroviral therapy on survival of HIV-infected tuberculosis patients in Peru', *Rev Panam Infectol* 12 (2), pp. 37–43.

David, T.D. (2014) *HIV/AIDS, and Kaposi's Sarcoma*, http://www.webmd.com/hiv-aids/guide/aids-hiv-opportunistic-infections-kaposis-sarcoma_.

Dedicoat, M., Vaithilingum, M. and Newton, R. (2003) 'Treatment of Kaposi's sarcoma in HIV-1 infected individuals with emphasis on resource poor settings', *Cochrane Database Systematic Reviews*.

Egger, M. et al. (2011) 'Correcting Mortality for Loss to Follow-Up: A Nomogram Applied to Antiretroviral Treatment Programmes in Sub-Saharan Africa', *PLoS Medicine* 8 (1), ed. by G.P. Bisson, e1000390.

Freeman, E., Semeere, A. and Wenger, M.e.a. (2016) 'Pitfalls of practicing cancer epidemiology in resource-limited settings: the case of survival and loss to follow-up after a diagnosis of Kaposi's sarcoma in five countries across sub-Saharan Africa.', *BMC Cancer* 16, p. 65.

Gbabe, O.F. et al. (2014) 'Treatment of severe or progressive Kaposi's sarcoma in HIV-infected adults.', *The Cochrane database of systematic reviews* Issue 9. Art. No.: CD003256.

Henriques, J. et al. (2012) 'Comparison of Methods to Correct Survival Estimates and Survival Regression Analysis on a Large HIV African Cohort', *PLoS ONE* 7 (2), ed. by M.P. Fox, e31706.

Johansson, K.A., Robberstad, B. and Norheim, O.F. (2010) 'Further benefits by early start of HIV treatment in low income countries: Survival estimates of early versus deferred antiretroviral therapy', *AIDS Research and Therapy* 7, pp. 1–5.

Kendiga, C.E. et al. (2013) 'Cancer treatment in Malawi: A Disease of Palliation', *World Jouranl of Oncology* 4, pp. 142–146.

Kiragga, A. et al. (2013) 'Comparison of Methods for Correction of Mortality Estimates for Loss to Follow-Up after ART Initiation: A Case of the Infectious Diseases Institute, Uganda', *PLoS ONE* 8 (12), ed. by L. Kaderali, e83524.

Makombe, S.D., Harries, A.D. and al, E. (2008) 'Outcomes of patients with Kaposi's sarcoma who start antiretroviral treatment under routine program conditions in Malawi', *Tropical Doctor* 38, pp. 5–7.

Makombe, S.D., Harries, A.D., Yu, J.K.-L. et al. (2007) 'Outcomes of tuberculosis patients who start antiretroviral therapy under routine program conditions in Malawi', *Int J Tuberc Lung Dis* 38, pp. 412–416.

Malawi Ministry of Health (2008) *Guidelines for the use of Antiretroviral Therapy in Malawi*, tech. rep., Malawi Ministry of Health, Lilongwe http://apps.who.int/medicinedocs/documents/s18803en/s18803en.pdf; Accessed: 2013-12-04.

Malawi Ministry of Health (2011) *Guidelines for the use of Antiretroviral Therapy in Malawi*, tech. rep., Malawi Ministry of Health, Lilongwe http://www.who.int/hiv/pub/guidelines/malawi_art.pdf, Accessed: 2013-11-25.

Malawi Minstry of Health (2014a) *ART Progress Report-Quarter 3*, Malawi Ministry of Health, Lilongwe.

Malawi Minstry of Health (2016) *Malawi Integrated HIV Program Report, January-March 2016*, Malawi Ministry of Health, Lilongwe.

Malawi NSO (2008) *National Population Census 2008*, tech. rep., Malawi National Statistical Office(NSO), Zomba http://www.nsomalawi.mw/index.php?option=com_content&view=article&id=106&Itemid=6, Accessed: 2013-12-10.

Minhas, V. and Wood, C. (2014) 'Epidemiology and transmission of Kaposi's sarcoma-associated herpesvirus', *Viruses* 6, pp. 4178–4194.

Msyamboza, K.P. et al. (2012) 'Burden of Cancer in Malawi; Common Types, incidence and Trends: National population-based cancer registry', *BMC Research Notes* 5, pp. 1–8.

Ngarina, M. et al. (2013) 'Reasons for poor adherence to antiretroviral therapy postnatally in HIV-1 infected women treated for their own health: experiences from Mitra Plus study in Tanzania', *BMC Public Health* 450 (13).

Rohner, E. et al. (2014) 'Incidence rate of Kaposi sarcoma in HIV-infected patients on antiretroviral therapy in Southern Africa: a prospective multicohort study.', *Journal of acquired immune deficiency syndromes (1999)* 67 (5), pp. 547–54.

Sankaranarayanan, R. et al. (2010) 'Cancer survival in Africa, Asia, and Central America: a population-based study', *Lancet Oncology* 11, pp. 165–173.

Sitas, F., Bezwoda, W.R. et al. (1997) 'Association between human immunodeficiency virus type 1 infection and cancer in the black population of Johannesburg and Soweto, South Africa', *British Journal of Cancer* 75, pp. 1704–1707.

Sitas, F., Parkin, M. et al. (2006) *Disease and Mortality in Sub-Saharan Africa*, 2nd, The International Bank for Reconstruction and Development/The World Bank, Chapter 20.

UNAIDS (2014) *90-90-90- An ambitious treatment target to help end the AIDS epidemic*, http://www.unaids.org/en/resources/documents/2014/90-90-90.

UNAIDS (2015) *Malawi HIV and AIDS Estimates*, tech. rep., Switzerland: The Joint United Nations Programme on HIV/AIDS http://www.unaids.org/en/regionscountries/countries/malawi, Accessed: 2016-04-15.

UNAIDS (2016) *FACT SHEET 2016*, tech. rep., Switzerland: The Joint United Nations Programme on HIV/AIDS http://www.unaids.org/en/resources/fact-sheet, Accessed; 2016-04-15.

Vijay, S. et al. (2011) 'Treatment Outcome and Mortality at One and Half Year Follow-Up of HIV Infected TB Patients Under TB Control Programme in a District of South India', *PLoS ONE* 6 (7), ed. by M. Pai, e21008.

Wabing, H.R. et al. (2000) 'Trends in cancer incidence in Kyandondo County, Uganda, 1969-1977', *British Journal of Cancer* 82, pp. 1585–1592.

WHO (2014a) *Malawi Tuberculosis Financing Profile* , http://www.who.int/tb/country/data/profiles/en/.

WHO (2014b) *Prevalence of HIV among adults aged 15-49 (%)* , http://www.who.int/gho/hiv/epidemic_status/prevalence/en/.

World Health Organisation (2016) *Global Health Observatory Data: HIV/AIDS*, Available at http://www.who.int/gho/hiv/en/.

World Health Organization (2008) *Malawi Country Cooperation Strategy at glance*, tech. rep., World Health Organization, http://www.who.int/countryfocus/cooperation_strategy/ccsbrief_malawi_en.pdf?ua=1, Accessed: 2014-09-30.

World Health Organization (2011a) *Cancer Fact Sheet no 297*, tech. rep., World Health Organization, Geneva, http://www.who.int/mediacentre/factsheets/fs297/en/.

World Health Organization (2011b) *HIV/AIDS Fact Sheet*, tech. rep., World Health Organization, Geneva, http://www.who.int/hiv/topics/tb/hiv_tb_factsheet_june_2011.pdf, Accessed; 2014-01-15.

World Health Organization (2016) *Tuberculosis: Fact Sheet Fact sheet N°*104, tech. rep., World Health Organization, Geneva, http://www.who.int/mediacentre/factsheets/fs104/en/, Accessed: 2016-05-10.

Wu, X. et al. (2014) 'One hundred and five Kaposi's sarcoma patients: a clinical stud in Xinjiang, Northwest China', *Journal of European Academy of Dematology and Venereology* 28, pp. 1545–1552.

Yiannoutsos, C.T., An, M.-W. et al. (2008) 'Sampling-Based Approaches to Improve Estimation of Mortality among Patient Dropouts: Experience from a Large PEPFAR-Funded Program in Western Kenya', *PLoS ONE* 3 (12), ed. by B. Grinsztejn, e3843.

Ziegler, J. et al. (2003) 'Risk factors for Kaposi's sarcoma: A case-control study of HIV-seronegative people in Uganda', *International Journal of Cancer* 103 (2), pp. 233–240.

Chapter 4

General statistical review of survival analysis methods

Summary

In this chapter, we provide a general review of statistical methods for analysing time-to-event data. The main focus in this chapter is on methods for analysing right censored time to event data in which not all individuals experience the event of interest by the end of the study. We also review and discuss some of the extensions of standard survival models such as; flexible parametric survival models, competing risks model and multistate models and spatial methods for survival data and their application to disease epidemiology. Inferential approaches are also discussed for each model and efforts are made to link these methods to future studies in the next three chapters.

4.1 Introduction to survival analysis

Survival analysis techniques are used to analyse time to event data. These methods have been extensively used in many application fields such as engineering sciences, social science, medicine and epidemiology. For example, in medicine, survival methods are used in clinical trials and epidemiological studies. For instance, survival analysis is used for analysing time to death, time to cancer remission and length of stay in hospital (Kleinbaum and Klein, 2012; Diva, Banerjee and Dey, 2007).

In many studies not all individuals necessarily experience the event of interest by the end of study. In cohort studies, individuals may also be lost to follow-up during the course of a study. For individuals that did not experience the event of interest by the end of study, their observed times are referred to as right censored times. An analysis based on the complete event times only (by omitting the incomplete times) will more often over estimate the parameters of interest (see Figure 5.2 on page 104). Any analysis tool to be applied to such types of data ought therefore to account the fact that complete time was not observed for some individuals. Regression models that take into account this problem are popularly known as survival or duration models.

In analysing survival data, interest often focusses on modelling the instantaneous risk of experiencing the event or the estimated time before an individual experiences the event of interest. As a result, the main functions used in survival models are the hazard function and survival function. These functions are formally defined

and discussed in detail in subsection 4.1.3 and subsection 4.1.4 below. A detailed tutorial on survival analysis can be found in many textbooks such as Ibrahim, Chen and Sinha, (2010), Kleinbaum and Klein, (2012), Kalbfleisch and Prentice, (2002) and Hosmer and Lemeshow, (1999), and Cox, (1972).

4.1.1 Right Censoring

In studies in which not all individuals experience the event of interest, the investigators only have knowledge about the survival time up to a certain point but not the complete true observed time. In many cases, incomplete time is observed when either

- the individual does not experience the event by the end the study or
- the individual is lost to follow-up or
- the individual voluntarily withdraws from the study (the reason could be on health ground).

When the event time for an individual in the study has not been observed, then such times are referred to as *censored times*. In this thesis, we consider random censoring which is a common assumption in many application fields such as medicine and epidemiology. Random censoring means that survival times are independent of the underlying censoring mechanism (Kleinbaum and Klein, 2012). There are three types of censoring; right-censoring, left-censoring and interval censoring.

Observed times are said to be left-censored if the event of interest has already occurred before date of enrolment. Interval censoring is when the event time is known only to have happened in an interval.

However, in practice most survival data are right-censored (Kleinbaum and Klein, 2012, p. 7). Data are right-censored when an individual has not been observed to experience an event of interest by the end of study. This means that the event of interest in known to have happened after some time t but the exact time is not known. Henceforth, we will assume that the data of interest are right-censored.

In most studies, the study period is determined in advance (fixed or Type I censoring), and in this case the number of events is random. In some studies, however, the number of events is determined in advance (random or Type II censoring) meaning the study period is random, with the study ending when the required number events has been observed.

Table 4.1 shows the general design of data for a simple right-censored survival analysis. As we will notice from section 4.3, this design may change depending on the complexity of the problem under study.

Table 4.1: Data preparation for a simple right-censored survival analysis

ID#	age	sex	times	status	
1	18	F	119	1	← event
2	40	F	75	0	← censored
3	22	M	30	1	
4	38	F	19	0	
5	30	M	101	1	
6	20	F	80	1	

4.1.2 Density and Distribution Functions

The probability density function $f(t)$ for a random variable T, denoting survival time, is defined as;

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t} \tag{4.1}$$

where Δt represents a small change in t and $t \geq 0$. And the cumulative distribution function of T is

$$F(T) = P(T < t) = \int_0^t f(u)du \tag{4.2}$$

where $f(t)$ is given as above. The distribution $F(t)$ gives the probability that the survival time does not exceed some given value t (Collett, 2014, pp. 11).

4.1.3 Survival function

The survival function, denoted as $S(t)$ is related to the distribution function of t as follows;

$$S(t) = P(T > t) = 1 - F(t) \tag{4.3}$$

The $S(t)$ function gives the probability of an individual surviving beyond time t (Cox, 1972).

The survival function can be estimated by both non-parametric and parametric methods. Some parametric methods are described in detail under subsection 4.2.2. For non-parametric estimators, the two most popular methods used are the Kaplan-Meier (K-M) and the life table estimators. The only difference

between these two non-parametric estimators is that the life table method is a grouped analogue of K-M method and is usually used for larger samples of data or population-level data. Again these methods are well described in main text books on survival analysis (Ibrahim, Chen and Sinha, 2010; Kleinbaum and Klein, 2012; Kalbfleisch and Prentice, 2002; Hosmer and Lemeshow, 1999; Cox, 1972). However, in this thesis we concentrate on the use of Kaplan-Meier estimator of the survival function developed by Kaplan and Meier, (1958) because it is the most popular estimator. The K-M estimator is given by

$$\hat{S}(t) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j} \tag{4.4}$$

where n_j and d_j are the number of individuals at risk and the number of deaths respectively in the time interval (t_k, t_{k+1}), with $k = 1, 2, 3, \ldots r$ events occurring at times $t_1 \leq t < t_2 \leq \cdots t_r$. Survival curves based on the K-M estimator in Equation 4.4 are popularly referred to as Kaplan-Meier (K-M) curves, named after the authors. The K-M estimator has the following assumptions (Kaplan and Meier, 1958);

- that censoring is not related to the prognosis of patients.
- the survival probability is the same for patients recruited early and late in the study.
- that events actually occurred at the reported event times.

In survival studies, we are often interested in measuring the effect of covariates on the time of occurrence of an event. In this case, the use of the K-M estimator is limited only to inclusion of one categorical covariate or factor in the model. To compare the K-M curves amongst different groups we can use the log-rank test. The log-rank test can be approximated by Chi-Square statistic;

$$U_L = \sum_{j=1}^{r} d_{ij} - e_{ij} \tag{4.5}$$

where d_{ij} is the number of deaths in group i at time j, and $e_{ij} = n_{ij}d_j/n_j$ is the expected number of individuals who die at time t_j in groug i (Collett, 2014).

The statistic for comparing survival between two groups is

$$W_L = U_L/\sqrt{V_L}, \quad \text{with} \quad V_L = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)} \tag{4.6}$$

where V_L is the variance and $W_L \sim \chi_1^2$.

4.1.4 Hazard Function

The hazard function denoted as $h(t)$ is a function that measures the probability of failure in an infinitesimally small time period $(t, t + \Delta t)$, given that the individual has survived up to time t (Cox, 1972). Mathematically, this is defined as

$$h(t) = \lim_{\Delta t \to 0} \left\{ \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} \right\}, \quad t \geq 0 \tag{4.7}$$

where T is as defined above. The hazard, survival and the density functions are mutually related. The density function $f(t)$ is just the product of the hazard function and the survival function i.e. $f(t) = h(t) \times S(t)$. From the relationship of $f(t)$ and $S(t)$ in Equation 4.3, this means that the hazard function can also be expressed as;

$$h(t) = f(t)/S(t) = -\frac{\partial}{\partial t} \log S(t) \tag{4.8}$$

4.2 Models for Survival Data

In many cases especially in medical and epidemiological studies, there is interest in adjusting for the effects of covariates on the risk under study. In order to do this, regression models are used. These models fall under two main frameworks: proportional hazards (PH) models and accelerated failure time (AFT) models. We review these two modelling frameworks in the sequel. However, our review here is limited to time-constant covariates, but extensions for handling time-varying covariates are available in standard survival textbooks and software.

For the rest of this review we will use the following notation for simplicity. Let n be the number of individuals and p be the number of covariates for each individual. Also, let X be a $n \times p$ design matrix and β be a $p \times 1$ vector of covariate effects to be estimated from the data.

4.2.1 Proportional Hazards Model

Proportional Hazards (PH) models are regression models in which the hazards are modelled using the covariates and possibly random effects. The main assumption of PH models is that the hazard ratios between individuals in two different groups are proportional over time. In addition, they also assume that the covariates included in the model have a multiplicative effect on the hazard ratios (Kleinbaum and Klein, 2012). If the hazard ratio is greater than one, the exposure effect lowers the survival rate (increases risk of an event occurring) and vice versa.

A proportional hazards model can either be semi-parametric or parametric. We discuss these different types of models in turn under subsubsection 4.2.1.1 and in subsubsection 4.2.1.2.

4.2.1.1 Semi-Parametric PH Models

A semi-parametric model is a model that has both parametric and non-parametric components. An example of a semi-parametric model is the widely applied Cox PH regression model. The Cox PH model is a semi-parametric model because the distribution of the baseline hazard function is left unspecified. The Cox PH model is defined as

$$h(t) = h_0(t) \exp(X\beta) \tag{4.9}$$

where $h_0(t)$ is the baseline hazard (distribution-free). The cumulative baseline hazard function $H_0(t)$ is defined as

$$H_0(t) = \int_0^t h_0(u)du \tag{4.10}$$

The hazard function is related to the survival function in the way that

$$S(t) = \exp\left\{-\int_0^t h(u)du\right\} = \exp\left\{-H(t)\right\}, \tag{4.11}$$

where $H(t)$ is called the cumulative hazard function. This relation makes easy to generate one function from the other. For instance, if the survival function is available, one can obtain the cumulative hazard function by taking the natural logarithm of the survival function i.e. $H(t) = -\log S(t)$ from which $h(t)$ follows (Collett, 2014).

The parameters β can be estimated from the data (t_i, δ_i, X) (time, censoring indicator, covariate) for $i = 1, 2, \ldots, n$ individuals by maximizing the logarithm of the partial likelihood(PL) function introduced by Cox, (1972) as follows;

$$PL(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp\left(X\beta\right)}{\sum_{l \in R(t_i)} \exp\left(X_l\beta\right)} \right\}^{\delta_i} \tag{4.12}$$

where δ_i is an indicator of whether t_i is a right-censored time or not, $R(t_i)$ is the set of individuals at risk of experiencing the event at time t_i and X_i is the vector of covariates for individual i. This likelihood assumes that there are no ties in event times. However, if ties exist in the data, then the modified partial likelihood

introduced by Breslow, (1975) or by Efron, (1977) can be used to estimate the parameters in the likelihood. Breslow's estimator is;

$$PL(\beta) = \prod_{i=1}^{n} \frac{\exp(s_i\beta)}{\left\{\sum_{l\in R(t_i)} \exp(X_l\beta)\right\}^{d_i}} \tag{4.13}$$

where $s_i = \sum_{k\in D_i} X_k$, at distinct time i and d_i is the number of deaths at time t. The Efron's estimator is;

$$PL(\beta) = \prod_{i=1}^{D} \frac{\exp(s_i\beta)}{\prod_{j=1}^{d_i}\left\{\sum_{k\in R(t_i)} \exp(X_k\beta) - \frac{j-1}{d_i}\sum_{k\in D_i} \exp(X_k\beta)\right\}} \tag{4.14}$$

where D_i is a set of individuals that died at time i and all other terms as defined above.

4.2.1.2 Parametric PH Models

As noted above, a functional form for the baseline hazard $h_0(t)$ is not specified in the Cox PH model. However, it is also possible to specify a form for $h_0(t)$. When the form of the hazard function is specified and depends on some parameters, the resultant models are called parametric.

Parametric modelling has some advantages over the Cox PH modelling. Among others, parametric models have better prediction and extrapolation, better performance in modelling time-dependent effects and often produce reliable and consistent estimates under asymptotic results (Nardi and Schemper, 2003; Efron,

1977). The main caveat with parametric models is that they put strong assumptions on the distribution of the baseline hazard function. Due to uncertainty of misspecification of the baseline hazard distribution, many analysts opt for Cox PH over parametric model because the distribution of baseline hazards is left unspecified. However, the use of cubic splines in flexible parametric modelling of log of the cumulative hazard function provide similar estimates to the Cox PH and is increasingly being applied in medicine (see page 53 on discussion of flexible modelling).

In this section we limit our discussion to the case when the form of the baseline hazard function is derived either from the Exponential , Weibull, or Gompertz distribution. We also consider:

i) **Exponential PH model**

The Exponential distribution is a one parameter (rate) distribution that assumes that the rate is constant over time. So the hazard function is written as;

$$h_0(t) = \lambda$$

where λ is the rate parameter. The resultant Exponential PH model is

$$h(t) = \lambda \exp(X\beta) \tag{4.15}$$

Also the survival function in this case is $S(t) = \exp\{-\exp(X\beta)\lambda t\}$ (see Equation 4.11), and the density function of T becomes $f(t) = \lambda \exp(X\beta) \exp\{-\exp(X\beta)\lambda t\}$.

The likelihood function is expressed as

$$L(\beta, \lambda) = \prod_{i=1}^{n} \{\lambda \exp(X\beta)\}^{\delta_i} \exp\{-\exp(X\beta)\lambda t\} \tag{4.16}$$

ii) **Weibull PH model**

The Weibull distribution yields a more flexible baseline hazard than the Exponential in the sense that it has two parameters, a scale and shape parameter and includes the Exponential as a special case. The shape parameter controls the shape of the hazard function allowing it to either decrease or increase monotonically over time. The baseline hazard function $h_0(t)$ and the Weibull PH model respectively are ;

$$\begin{aligned} h_0(t) &= \lambda\alpha t^{\alpha-1} \\ h(t) &= \lambda\alpha t^{\alpha-1} \exp(X\beta) \end{aligned} \tag{4.17}$$

where λ is the scale parameter and α is the shape parameter, both λ,$\alpha > 0$. If $\alpha = 1$, then Equation 4.17 becomes an Exponential model (special case of Weibull model) and λ is interpreted as a rate parameter. If $\alpha > 1$, then the hazards increase over time and if $\alpha < 1$, the hazards decrease over time. The Weibull survival function is $S(t) = \exp\{-\exp(X\beta)\lambda t^{\alpha}\})$ and the density function for T is $f(t) = \lambda\alpha t^{\alpha-1} \exp(X\beta) \exp\{-\exp(X\beta)\lambda t^{\alpha}\})$. The likelihood function is expressed as

$$L(\beta, \alpha, \lambda) = \prod_{i=1}^{n} \left\{\lambda\alpha t^{\alpha-1} \exp(X\beta)\right\}^{\delta_i} \exp\{-\exp(X\beta)\lambda t^{\alpha}\} \tag{4.18}$$

iii) **Gompertz PH model**

The Gompertz model is also a two-parameter model with the hazard and survival functions defined as follows;

$$\begin{aligned} h_0(t) &= \lambda \exp(\gamma t) \\ S(t) &= \exp\left\{-\lambda\gamma^{-1}(\exp(\gamma t)\right\} \end{aligned}$$

where λ and γ are unknown parameters to be estimated from the data. The Gompertz PH model is

$$h(t) = \lambda \exp(\gamma t) \exp(X\beta) \tag{4.19}$$

A positive value of γ means increasing hazard rates with time, a negative value means a decreasing hazard rate and if γ is zero, then the model reduces to an exponential model. The β are obtained by maximising the log-likelihood;

$$L(\beta, \gamma, \lambda) = \prod_{i=1}^{n} \{\lambda \exp(\gamma t) \exp(X\beta)\}^{\delta_i} \exp\left\{-\exp(X\beta)\lambda\gamma^{-1} \exp(\gamma t)\right\} \tag{4.20}$$

iv) **Flexible PH Model**

This is one of the flexible parametric models proposed by Royston and Parmar, (2002). Royston and Parmar, (2002) proposed a flexible method for estimating the baseline hazard function $h_0(t)$. They proposed to model the log of cumulative hazard function using a natural cubic spline constrained to be linear beyond the maximum survival time. Cubic splines are basically piecewise polynomials passing through a set of control points, also known as *knots*. The flexible PH model is

formulated as follows;

$$\log H(t) = s(\log(t); \gamma) + X\beta \tag{4.21}$$

where $H(t)$ is the cumulative hazard function, $s(.)$ is natural cubic spline with γ as spline coefficients and $s(.)$ is defined as

$$s(x; \gamma) = \gamma_0 + \gamma_1 x + \gamma_2 v_1(x) + \ldots + \gamma_{m+1} v_m(x), \quad \text{with}$$

$$v_j(x) = (x - k_j)_+^3 - \lambda_j (x - k_{\min})_+^3 - (1 - \lambda_j)(x - k_{\max})_+^3$$

where $v_j(x)$ is the j^{th} basis function, $k_1 < \ldots k_m$ being the internal knots, and λ_j is a j^{th} location parameter defined as $\lambda_j = (k_{\max} - k_j)/(k_{\max} - k_{\min})$

It is usually optimal to choose the internal knots around more dense regions to minimise variance. As such, knots around the median or quantiles are preferred, suffice to say that the placement of knots does not critically affect the goodness of fit of the model (Hinchliffe and Lambert, 2013).

iv) **Flexible Proportional Odds Model**

Similar to Equation 4.21, the flexible parametric proportional odds model is formulated as

$$\log(S(t)^{-1} - 1) = s(\log(t); \gamma) + X\beta \tag{4.22}$$

where $S(t)$ is the survival function, $s(.)$ is natural cubic spline with γ as spline coefficients, all other parameters defined as in Equation 4.21. A detailed description of these flexible models can be found in Royston and Parmar, (2002).

4.2.2 Accelerated Failure Time Models

The AFT models describes the direct relationship between failure time and covariates. While in Cox PH model the key assumption is that the hazards are proportional, in AFT models the key assumption is the accelerated failure time. An acceleration factor is a measure of association in AFT models just as hazards ratio in Cox PH models. This factor describes the stretching out or contraction of survival functions when comparing one group to another for any fixed value of the $S(t)$. An acceleration factor of greater than one indicate that the covariate effect is beneficial (or stretches out survival time) and vice versa. The frequently used distribution families in AFT modelling are; Exponential, Weibull, Log-logistic, Log-normal and Gamma distribution(Kleinbaum and Klein, 2012). The Exponential and Weibull models are special cases of parametric survival models that satisfy both PH and AFT assumptions. The advantage of parametric PH models over Cox PH models is that they are consistent with theoretical survival function $S(t)$ (Kleinbaum and Klein, 2012).

The AFT models put individuals with different covariates on different time scales. Also, the AFT models assume that the log of failure time T, $\log T$, is in a linear relationship with a mean μ, the covariates (and their parameters) and an error term W, where W takes a particular distribution. In other words, AFT models are additive on log scale but multiplicative with respect to T.

$$\log T = \mu + X\beta + \sigma W$$

where σ is a scaling parameter. The type of model to be fitted depend on the choice of the distribution for the errors W. As already mentioned in preceding paragraphs, the available distribution for survival models in analysis software are Exponential, Weibull, Gompertz, generalized Gamma, Log-normal and Log-logistic.

In practice, AFT models are rarely used in application analyses, with PH models more popular in medicine and epidemiology. However, AFT models offer a good alternative to PH models in that they can be used when the PH assumption is violated. The PH assumption may be violated in the instances when the treatment under study tends to delay the onset of the event as opposed to increasing or reducing occurrence of the event (Patel, Kay and Rowel, 2006).

In the next sections, we briefly summarize the behaviour of the hazard and survival functions for each of these distributions.

In AFT framework and using Section 4.2.2, the parametric models have the survival function in the form of $S(t) = S_0\left\{t/\exp(X\beta)\right\}$, where S_0 is the baseline survivor function defined as, $S_0(t) = \mathrm{P}\left\{\exp(\mu + \sigma W \geq t)\right\}$ (Collett, 2014, pp. 207).

4.2.2.1 Weibull Model

The AFT form of the Weibull model is parametrised in terms of survival time (Kleinbaum and Klein, 2012, pp. 208). We have discussed that the Weibull PH

model is

$$h(t) = \lambda\alpha t^{\alpha-1}\exp(X\beta) \tag{4.23}$$

$$S(t) = \exp\left\{-\lambda t^{\alpha}\exp(X\beta)\right\} \tag{4.24}$$

The parameters are obtained by maximising the log-likelihood function;

$$L(\beta,\alpha,\lambda) = \prod\left[\lambda\alpha t^{\alpha-1}\exp(X\beta)\right]^{\delta_i}\exp\left\{-\exp(X\beta)\lambda t^{\alpha}\right\} \tag{4.25}$$

The limitation of the Weibull model is that it is a monotonic function of time, can either increase or decrease.

4.2.2.2 Log-logistic Model

The survival function for the log-logistic model is

$$S(t) = \left\{1+\exp\left(\frac{\log t-\mu-X\beta}{\sigma}\right)\right\} \tag{4.26}$$

where μ and σ are as defined in subsection 4.2.2 and are estimated from the data. The model can be used in situations where the hazard is thought to increase for a short time and decrease overtime, such as in studies on risk of death in patients after a heart transplant. If the proportional odds assumption holds for the data, then the Log-logistic model can also be fitted as a survival odds model; $\frac{S(t)}{1-S(t)}$ (Kleinbaum and Klein, 2012).

4.2.2.3 Log normal Model

The log normal and log logistic models tend to be similar since the logistic and normal distribution are similar. These models are parametrised only as AFT. In both models, the natural logarithm follows a normal and logistic distribution respectively. These models are usually suitable for scenarios where the hazard rates are non monotonic, specifically increasing in the first instance and then decreasing.

The survival function for the log normal model is

$$S(t) = 1 - \Phi\left\{\frac{\log(t) - \mu - X\beta}{\sigma}\right\} \tag{4.27}$$

where $\Phi(z)$ is a normal cumulative distribution function, μ and σ are estimated from the data (Collett, 2014)

4.2.2.4 Generalized Gamma Model

Finally, we consider a generalized gamma model which is also parametrised as an AFT model only and offers more flexibility in modelling hazard functions. The survival function is given by

$$S(t) = 1 - \Gamma_{(\mu t)^\theta}(\rho)$$

where $\Gamma_{(\mu t)^\theta}(\rho)_0 = \frac{1}{\Gamma(\rho)} \int^{(\lambda t)^\theta} \mu^{\rho-1} \exp(-\mu)\mathrm{du}$. The flexibility of the hazard function of the generalized gamma distribution is that it has the Gamma, Weibull, Exponential and Log normal models as special cases. If both ρ and θ are equal to

1, then it becomes exponential model, if $\rho = 1$ it becomes Weibull, and reduces to log normal if $\rho \to \infty$ (Collett, 2014).

4.3 Extensions to Standard Survival Models

The modelling techniques discussed so far can be considered as standard. However, extensions to these models are available which are aimed at modelling more complex datasets. Here, we present and discuss frailty models, spatial models and multistate models (which include the competing risk models).

4.3.1 Frailty Models

In survival modelling, sometimes interest lies in modelling heterogeneity at the individual or at a group level not captured by the available covariates. To accomplish this, all of the survival models discussed in the preceding paragraphs can be extended to incorporate a term that specifically measures heterogeneity, popularly referred to as a *frailty term* (Collett, 2014; Therneau and Grambsch, 2000; Wienke, 2011). For example, the Equation 4.9 can be extended to incorporate a frailty term Z to become

$$h(t) = Zh_0(t)\exp(X\beta) \tag{4.28}$$

Here, Z has a known distribution form and is a surrogate for unobservable heterogeneity, say, amongst individual patients or clusters: Z can be thought of as a

multiplicative random effect. Mathematically, Z is defined as $Z \overset{iid}{\sim} f(.;\sigma)$ where f is a positive-valued probability density function of given form (e.g log-normal or gamma). The likelihood function for the parameters conditional on $Z_1, Z_2, \ldots, Z_n$ is given by

$$L(\beta, \sigma|z) = \prod_{i=1}^{n} \{z_i h_0(t) \exp(X\beta)\}^{\delta_i} \exp\{-z_i H_0(t_i) \exp(X\beta)\} \tag{4.29}$$

where δ_i is the censoring indicator, β and σ are parameter estimates for covariate effect and the heterogeneity component respectively, and $H_0(t_i)$ is the cumulative baseline hazard function. Larger values of this variance indicate greater heterogeneity among the sample units. The Gamma and the Log-normal distributions are the popular choices for the distribution of Z.

For a gamma-distributed Z, its density is given as

$$f(z) = \frac{1}{\Gamma(k)} \lambda^k z^{k-1} \exp(-\lambda z) \tag{4.30}$$

where its variance σ is indirectly calculated as $Var(z) = \lambda^2 k$. The interest is in the value of this variance

The density for a log-normal distributed Z is given by

$$f(z) = \frac{1}{\sqrt{2\pi\sigma z}} \exp\left(-\frac{(\ln(z))^2}{2\sigma^2}\right) \tag{4.31}$$

where σ is the variance, and again this is reported in the fitted models.

4.3.2 Spatial Survival Modelling

Spatial survival models are another extension of the standard survival models. These models are an extension of frailty models in which the frailties are allowed to be spatially correlated and this correlation depends on the distance apart in space or depends on the spatial neighbourhood effect. In these models, the assumption of independence in the residuals is relaxed. In the context of epidemiology, spatial methods have extensively been used to study disease pattern and identifying the source of a particular problem (Adebayo and Fahrmeir, 2005; Henderson, Shimakura and Gorst, 2002; Li and Ryan, 2002; Banerjee and Dey, 2005). Spatial methods have a particular appeal in epidemiology as they can help understand underlying environmental factors influencing disease occurrence or progression. Fitting such models can help in planning and allocation of resources to tackle the disease. With survival analysis being popular in epidemiology and health research, spatial survival analysis have become popular with the growing methodological work (including computation methods) in this area in the past two decades.

One of the main reasons to conduct a spatial survival analysis is to identify regions in space in which the risk of experiencing the event is unusually large. Other factors being equal, we might expect individuals closer in space to have similar environmental exposures compared with individuals far apart.

In subsection 4.3.1, we discussed frailty models as an extension of the Cox PH model. A similar approach is used in spatial survival models to include the spatial frailties. These spatial frailties describe the spatial variation in survival or hazards

of death. The general formulation of the proportional hazards spatial frailty model denoted by re-writing Equation 4.28 as,

$$h(t;x) = h_0(t)\exp\{X\beta + S(x)\}, \quad x \in \mathbb{R}^2 \tag{4.32}$$

where βs as defined above, $S(x)$ is the value of a continuous stationary Gaussian process (SGP) at x in space. Values of $S(x)$ are more correlated between points that come close compared to those that are far away. Different choices of $S(x)$ are available. For example, Henderson, Shimakura and Gorst, (2002) used a slightly different parameterisation and a gamma random field. It is also possible to have assume models that assume discrete spatial variation:

$$h(t,s) = h_0(t)\exp\{X\beta + S_i\} \tag{4.33}$$

where S_i are defined on discrete regions and possibly have a conditional autoregressive correlation structure.

Our focus in this thesis is on spatial proportion hazards (PH) models defined in Equation 4.32 and these models have been discussed by Henderson, Shimakura and Gorst, (2002), Li and Ryan, (2002), and Banerjee and Carlin, (2003). Methodology work on other modelling frameworks is also available in the literature: Banerjee and Carlin, (2004) on spatial proportional odds models; Zhang and Lawson, (2011) on spatial AFT models; Banerjee, Wall and Carlin, (2003). Also, some work has been done on spatial multistate modelling by Nathoo and Dean, (2008) and Brezger, Kneib and Lang, (2005).

Due to complexity of spatial survival models, the Bayesian approaches to modelling are common than the Frequentist approach. Taylor and Rowlingson, (2014) is an example of recent methodological work aimed at improving the computation time of spatial PH models and the authors provide a concise review of other available computation approaches. In Chapter 6, we employ the computation approaches as proposed by Taylor and Rowlingson, (2014). Their methods are implemented an R-package `spatsurv`.

4.3.2.1 Covariance Functions

The list of available covariance functions used for this purpose among others include; Gaussian, Spherical, Wave, Power law and the Matern family and details about these functions can be found in textbooks such as Diggle and Ribeiro, (2007, p. 54-56) and Banerjee, Carlin and Alan E, (2004, pp. 32-35).

However, in many applications due to its flexibility, the Matern family functions are favoured in model selection process (Banerjee, Carlin and Alan E, 2004; Henderson, Shimakura and Gorst, 2002). The Matern family has an Exponential as a special case.

Here, we present a summary of some of popularly used correlation functions due to Matern, (1986, p. 18) and Diggle, Tawn and Moyeed, (1998) and are given as follows;

1. Exponential function:

$$\rho(\mu) = \exp(-\mu/\phi), \quad \phi > 0 \tag{4.34}$$

where $\mu = \|x - x'\|$ is the distance between two arbitrary points $x, x' \in \mathbb{R}^2$, and ϕ is the spatial decay parameter with dimensions of distance.

2. Powered exponential (Diggle, Tawn and Moyeed, 1998):

$$\rho(\mu) = \exp[-(\mu/\phi)^\kappa], \quad \phi > 0 \tag{4.35}$$

where κ (also called the *order*) is a shape parameter determining how smooth the underlying $S(x)$ is, and ϕ is as defined above.

3. Matern function (Matern, 1986):

$$\rho(\mu) = [2^{\kappa-1}\Gamma(\kappa)]^{-1}(\mu/\phi)^\kappa K_\kappa(\mu/\phi), \quad \kappa, \phi > 0 \tag{4.36}$$

in which κ and ϕ are as defined above, and $K_\kappa(.)$ is a modified Bessel function of order κ. The Exponential function is a special case of the Matern function, with $\kappa = 0.5$.

The correct choice for the underlying spatial correlation function is important because it informs extent to and manner in which the correlation decays in space for the disease. Misspecification of this function may lead to incorrectly specifying how spread disease is in space and the closeness of the spatial units.

4.3.3 Competing risk and Multi-state Modelling

The methods presented so far assumed standard time-to-event data in which only one event (all-cause death) is of interest and all other events are censored. However, in medical research there are situations in which we may wish to model failure from a number of possible causes. We call these causes competing risks. For example, if our interest is to study time until death due to breast cancer, then if some women die from stroke, this will impede the occurrence of death due to breast cancer. In this case, we regard stroke as a competing cause of death.

It should be noted that all competing events have the same initial state (being alive and free of breast-cancer as in our example). Sometimes, interest is in estimating transitions into intermediate states/events, for instance from being diseased to disease-free (healing/recovery) or to relapse. In this case, we can use multistate models which are a generalization of the competing risks models. In these models, subjects are usually assumed to belong to one of a finite number of states, including death or censored. The rest of this section is devoted to reviewing the mathematical theory and their assumptions of the competing risk and multistate models.

We hasten to mention to the reader that although there are extensions of multistate models to include individual frailties and spatial frailties are available, we restrict our review in this section to the non-spatial multistate and competing risk models. For models with frailties, see some of the methodological work by Alvaro-Meca et al., (2012) and Nathoo and Dean, (2008) and Brezger, Kneib and Lang, (2005).

4.3.3.1 Competing Risk Modelling

The data set up in competing risk framework is different from the standard time-to-event data with one cause of death. In competing risk data, in addition to censored times (0) and cause of interest(1), other competing causes are assigned values $2, 3 \cdots$ instead of being censored to indicate they are also outcomes of interest. Table 4.2 shows how data are commonly set up before analysing them. In this example, individuals with IDs 1 and 5 had the same event 1, the individuals 3 experienced the event 2 while individuals 2 and 4 did not experience any event by the end of the study and were censored (0).

Table 4.2: Data preparation for competing risk modelling

ID#	age	sex	times	status
1	18	F	119	1
2	40	F	75	0
3	22	M	30	2
4	38	F	19	0
5	30	M	101	1

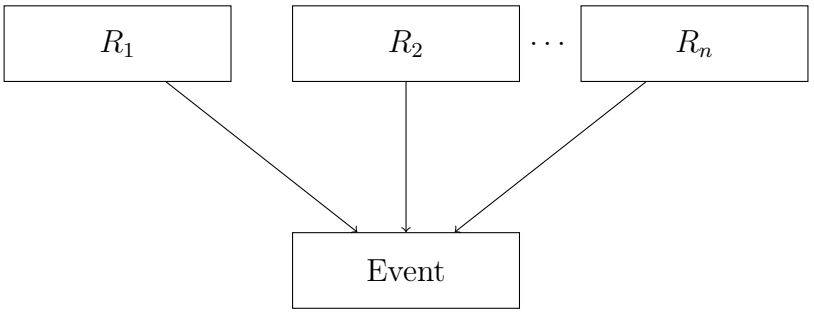

Figure 4.1: Competing risks R_i. The direction of arrows denotes "'competing to"' the Event of interest.

The all-cause hazard rate was defined in Equation 4.7. Under the competing risk framework, the instantaneous risks of dying from cause j (cause-specific hazard),

for $j = 1, 2, \ldots\ldots m$ is expressed as;

$$h_j(t) = \lim_{\Delta t \to 0} \left\{ \frac{P(t \leq T \leq t + \Delta t, C = j | T \geq t)}{\Delta t} \right\} \tag{4.37}$$

with t and Δt as defined above (Andersen, Per Kragh and Keiding, 2006; Collett, 2014, pp. 405-21).

The overall survival function is defined as the probability of an individual surviving all types of failures to time t. This is defined as;

$$\begin{aligned} S(t) &= \prod_{j=1}^{m} \exp\left\{ - \int_0^t h_j(u) du \right\} \\ &= \prod_{j=1}^{m} S_j(t) \end{aligned} \tag{4.38}$$

The survival function $S_j(t) = \exp\{-H_j(t)\}$, with $H_j(t)$ as the cumulative hazard for cause j does not have the same interpretation as $S(t)$ in Equation 4.38 as $m > 1$.

The assumption for $S_j(t)$ is that an individual will survive one cause only. But in practice, $S_j(t)$ is not observable since survival is defined as surviving all causes, not just one cause (Collett, 2014). Because of its awkward interpretation, $S_j(t)$ is rarely reported in the survival analyses.

Under the assumption of independent censoring, the density and partial likelihood functions for a specific cause j follow from Equation 4.8 and Equation 4.12, and are expressed as;

$$f_j(t) = h_j(t) S(t) \tag{4.39}$$

$$L = \prod_{i=1}^{n} \prod_{j=1}^{m} \left\{ \frac{\exp(X\beta_j)}{\sum_{l \in R(t_i)} \exp(X_l \beta_j)} \right\}^{\delta_{ij}} \tag{4.40}$$

In order to fit the regression models for competing risk data, separate cause-specific models are fitted by treating all other causes as censored. For instance, a Cox regression for the cause-specific hazard model is a modification of Equation 4.9 and is written as;

$$h_j(t) = h_{0j}(t) \exp(X\beta_j) \tag{4.41}$$

where the baseline hazard is now cause-specific baseline hazard and the covariate coefficients β_j are for a cause j. These models can be used to determine association between covariates and the specific cause j.

4.3.3.2 *Cumulative incidence function*

The survivor function is a useful summary for the standard survival data with no competing risks. However, in the presence of the competing risks, the cause-specific cumulative incidence function (also called the *subdistribution function*) is used as a summary tool for these types of data (Beyersmann, Allignol and Schumacher, 2012; Collett, 2014, pp. 405-21). A cumulative incidence function is the probability of an individual dying before time t from cause j in the presence of all other competing risks. It is expressed as;

$$\mathrm{F}_j(\mathrm{t}) = \mathrm{P}(\mathrm{T} < t, C = j), \quad \text{for} \quad j = 1, 2 \ldots\ldots m \tag{4.42}$$

Following the expression of $f_j(t)$ in Equation 4.39, the cause-specific cumulative incidence $F_j(t)$ is just the integrated function and can be estimated as a product of two estimates as follows;

$$\hat{F}_j(t) = \sum_{i:t_i \leq t} (\frac{\delta_{ij}}{n_i}) \hat{S}(t_{i-1}) \tag{4.43}$$

where $\frac{\delta_{ij}}{n_i}$ is the Nelson-Aalen estimator of the hazard function for the j^{th} cause, δ_{ij} is an event indicator for cause j, n_i is the number of individuals just before time t that are alive and uncensored and $\hat{S}$ is the K-M estimate of the overall survival function ignoring different competing causes. This is because Equation 4.43 uses information on death times for all causes. Suffice to say that in time-to-event data where $m = 1$, it is possible to obtain the cumulative incidence function from the hazard function (pp. 247-66 Kalbfleisch and Prentice, 2002; Collett, 2014, pp. 405-21).

4.3.3.3 Fine and Gray Model

In order to obtain an interpretable cause-specific cumulative incidence function, Fine and Gray, (1999) proposed a modified cause-specific hazard (Equation 4.37) to include individuals that died from a cause other than cause j (as still event-free). The modified cause-specific is called the *subdistribution function* and is the probability of failure due to cause j at a moment time. It is defined as;

$$\lambda_{ij}(t) = \lim_{\delta t \downarrow 0} \left\{ \frac{P(t \leq T \leq t + \delta t, C = j | T \geq t \quad \text{or} \quad \{T \leq t \text{and} \quad C \neq j\})}{\delta t} \right\} \tag{4.44}$$

Using this hazard function, it is now possible to obtain an interpretable corresponding cause-specific cumulative incidence function. This can be modelled either using a Cox regression or parametric models. In a Cox regression, the partial likelihood function for the j^{th} cause is expressed as follows;

$$L^* = \prod_{h=1}^{r_j} \left\{ \frac{\exp(X_h \beta_j)}{\sum_{l \in R(t_{(h)})} w_{hl} \exp(X_l \beta_j)} \right\} \tag{4.45}$$

in which the product is taken over r_j individuals who died from cause j at ordered times $t_{(1)}, t_{(2)}, \ldots\ldots, t_{(r_j)}$, X_h is a vector of covariates for individuals who died at time $t_{(h)}$, $h = 1, 2, \ldots\ldots, t_j$ and β_j is as defined above. The risk set $R(t_{(h)})$ contains individuals who have not yet died by time $t_{(h)}$.

The formulation of the Fine and Gray hazard model is similar to Equation 4.41 but the main differences are in the calculation of the risk set and how the product in the likelihood function is calculated (Collett, 2014, pp. 405-21). Also, the cumulative incidence function can be estimated as

$$\hat{F}_{ij}(t) = 1 - \exp\left\{ -\hat{H}_{ij}(t) \right\} \tag{4.46}$$

where $\hat{H}_{ij}(t)$ is an estimate of the cumulative subdistribution hazard function estimated as

$$\lambda_{ij}(t; X_i) = \lambda_{0j}(t) \exp(X_i \beta_j)$$

Researchers are often faced with a decision to choose between the use of Cox PH for a cause-specific hazard and a Fine & Gray model. There three main issues

that ought to be considered. First, The estimates from cause-specific hazard function do not have an informative interpretation because they rely heavily on the assumption that censoring is independent from competing events (Fine and Gray, 1999). Secondly, because of the issue of possible dependent censoring, the use of cause-specific K-M estimate would be difficult to interpret as it ignores the issue of dependent censoring among competing events, which is very difficult to verify. As such, by assuming the overall function (see Equation (4.43), the CIF bypasses this dilemma. Lastly, although the cause-specific hazard ratios and subdistribution hazard ratios provide an informative summary of covariate effects, the proportionality assumption cannot be satisfied at the same time for both models (Fine and Gray, 1999).

The subdistribution function is analogous to Cox PH model. However, in subdistribution function, the CIF is used as the hazard function and its interpretation can be tricky to non-statisticians. Due to awkwardness in interpretation of the subdistribution function, cause-specific cumulative incidence function (CIF) are rather popular in analysis of competing risks data (Zhang and Fine, 2008).

4.3.3.4 Multistate Modelling

In the competing risk modelling framework, times of occurrence to competing events from the initial state is the main concern. However, sometimes the times of occurrence between non-fatal states other than initial state or death are of interest and competing risk models cannot be used in this case. This is where multistate models come in.

Multistate models are used to analyse event history data collected over a period of time and all events occurring for the individuals are recorded. Before analysing data using the multistate models, it ought to be prepared in the right format. The format of the data depends on whether the sequence or the order in which the occurrence of states matter. For example, if the order of visits to the states matters, then the data is put in a counting process format. Table 4.3 is an example of data in the counting process format.

Table 4.3: Data preparation for Multistate modelling

ID#	baseAge	sex	timein	timeout	state
1	23	F	0	200	0
1	23	F	200	247	0
1	23	F	247	283	2
2	34	M	0	121	0
3	30	M	0	197	1
3	30	M	197	233	0
4	17	F	0	75	0
4	17	F	75	177	0
4	17	F	177	190	0

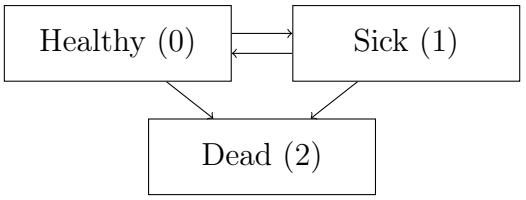

Figure 4.2: The illness-death model

For example, the individual with ID number 1 died after 283 days in *state* 2 and the individual 4 remained in the initial *state* 0 by the end of the study.

Figure 4.2 is a popular illness-death model in multistate modelling. The individual starts in initial state of *being healthy* and can move (*transition*) to other states

either *being sick* or *die*. Depending on the type of the disease, it is also possible to move back from being sick to being healthy again (recovery). But once an individual is in *'dead'*state , no further transitions are possible. States *'healthy'* and *'sick'* are called as *transient states* whereas state *dead* is called an *absorbing state*.

A comprehensive review of methods for multistate models can be found in textbooks such as Collett, (2014), Klein et al., (2013) and Beyersmann, Allignol and Schumacher, (2012), and Cox and Miller, (1965). In this overview, only *Markvov processes* are discussed. We start by giving some definitions and introducing notation to be used in the sequel.

A stochastic process (SP) X is a collection of random variables (rvs) $\{X(t) : t \in \mathcal{T}\}$ with T being some index set $0, 1, 2, 3, \ldots$. Let $\mathcal{S}$ be the *state space* (the set of possible values) of a stochastic process of $X(t)$, with $\mathcal{S} = \{1, \ldots, p\}$. Let $\pi_k(0) = \text{Prob}(X(0) = k)$ $k \in \mathcal{S}$ be the *initial distribution*.

A stochastic process is a Markov process if $P[X(t) = j|\mathcal{H}_-] = P[X(t) = j|X(t_1) = i_1]$, for $t > t_1 > t_2 > \ldots$ and $\mathcal{H}_- = \{X(t_1) = i_1, X(t_2) = i_2, \ldots\}$ is the history of the process. This means that the current state does not depend on the whole history $\mathcal{H}_-$ of the the process but just on the immediate previous state $X(t_{i-1})$. This assumption is often criticised as too simplistic especially in medical research in which patient's medical history play an important role in predicting patient's prognosis. Nevertheless, all the transition functions in this thesis will assume the Markov property. Details of statistical methods for non-Markovian processes and

hidden-Markovian processes can be found in the referenced textbooks.

The derivation of functions in multistate models is similar to the way the hazard and the survival function are derived except that in the multistate modelling the emphasis is on obtaining the dynamics of transitions from one state to another. The transition probability of state j to k ($j \to k$) is

$$P_{jk}(s,t) = \text{Prob}(X(t) = k | X(s) = j), \quad j,k \in \mathcal{S} \tag{4.47}$$

The transition probabilities can be summarized in a matrix called *transition probability matrix*. In our example, the probability matrix at time t is

$$P(0,t) = \begin{Bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ 0 & 0 & 1 \end{Bmatrix}$$

where p_{11} and p_{22} are probabilities of remaining in state 1 and 2 respectively, p_{12} is the probability of moving from state 1 to state 2, p_{13} is the probability of moving from state 1 to state 3, p_{21} is the probability of moving from state 2 back to state 1, p_{23} is the probability of moving from state 2 to state 3. State 3 is an absorbing state so transitions out of this state are not possible, therefore there is 100% chance ($p_{33} = 1$) of remaining in this state at any time t. Also note that the sum of each row must equal to 1 (law of total probability).

The transition intensity is the transition rate describing the rate at which a Markov chain moves between states i.e. from state j to state k, and is mathematically

defined as

$$q_{jk}(t) = \lim_{\delta t \to 0} \left\{ \frac{P_{jk}(t, t+\delta t)}{\delta t} \right\} \tag{4.48}$$

The transition intensity kernel (matrix) for the multistate models contains non-zero elements for all possible transitions, and zero otherwise. As in the transition probability matrix, the transitions are read from left to right. For example, in illness-death example with three states, the transition matrix is

$$Q = \begin{Bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ 0 & 0 & 0 \end{Bmatrix}$$

where q_{11} and q_{22} is intensities of remaining in *healthy* state and in *sickness* state respectively, q_{12} is the transition health $\rightarrow$ sick, q_{13} is the transition *health* $\rightarrow$ *death* state, q_{21} is the recovery process *sick* $\rightarrow$ *health*, q_{23} is the transition *sick* $\rightarrow$ *death*. The last row of Q has all entries zeros because transitions out of this state are not possible once an individual has entered the *death* state (absorbing state). The row sum of the entries in the matrix Q are *zero* and the diagonal entries are defined as the negative sum of off-diagonal entries in the same row, i.e. $q_{11} = -(q_{12} + q_{13})$ and $q_{22} = -(q_{21} + q_{23})$.

The likelihood is formulated as follows. Let T^i_{jk} be the i^{th} individual's time of transition from state j to state k and let $N_{jk}(t) = \sum_i^n N^i_{jk}(t)$ be counting process of the occurrences of the transition of interest among all the individuals. Also, let $Y_j(t) = \sum_i^n Y^i_{jk}(t)$ be the number of individuals at risk in state j at time $t-$, $j, k \in \mathcal{S}$. Then the likelihood function is (pp. 181-82 Hougaard, 2000; Aalen,

Borgan and Gjessing, 2008, p. 210)

$$L = \prod_{i}^{n} \prod_{j \neq k} \prod_{h=1}^{N_{jk}^{i}(\tau_i)} q_{jk}^{i}(T_{jk}^{ih}) \exp\left(- \int_{0}^{\tau_i} q_{jk}^{i}(t) Y_{j}^{i}(t) dt \right) \tag{4.49}$$

where covariates information is contained in $q_{jk}^{i}(.)$ and $q_{jk}^{i}(.)$ may well depend on whole history $\mathcal{H}_{t}^{i}$ of the process.

Table 4.4 provide a summary of other common functions of interest in multistate modelling.

Table 4.4: Common expressions in Multistate modelling

Name	Expression
State Probability	$\pi_j(t) = \texttt{Prob}(X(t) = k)$ $= \sum_{j \in \mathcal{S}} \pi_j(0) P_{jk}(0,t), \quad j,k \in \mathcal{S}$
Total number of observed direct $j \to k$ transitions in interval [0,t]	$N_{jk}(t) = \sum_{i}^{n} N_{jk}^{i}(t), \quad j \neq k$ where $N_{jk}^{i}(t) = \#$direct transitions $j \to k$ in [0,t] for i
Total length of length in *state* k	$L_k = \int_{t_1}^{t_2} P_{jk}(t) dt$

4.4 Model Diagnostics and Selection

In this section we provide a summary review of general methods for assessing fitness of the survival model to the data at hand. For proportional hazard models, we discuss how the assumption of proportionality is assessed. Several other model diagnostics methods including graphical diagnostic plots based on deviance and martingale residuals and a statistical test using Schoenfeld residuals are well documented in survival textbooks such as Collett, (2014), Therneau and Grambsch,

(2000) and Hosmer and Lemeshow, (1999), and Kleinbaum and Klein, (2012). We also look at available methods for comparing and choosing a best model from a set of competitive models.

4.4.1 Testing the PH Assumption

It is important to check the proportionality assumption in PH models. The assumption is checked for every variable included in the model. To check this assumption, usually graphical plots and Schoenfeld residual test are used.

The model-based graphical plots used for this purpose are; plots of complementary log-log survival curves against survival time, and plot of the observed K-M estimates $\hat{S}(t)$ and the expected survival estimates $S(t;\hat{\theta})$ with a covariate included.

The scaled Schoenfeld residuals are calculated for each covariate in the model. This is different from other types of residuals which are just a single value (Therneau and Grambsch, 2000). The assumptions for this test is that Schoenfeld residuals for each covariate are independent of each other. The test assesses if the residuals are correlated to some function of time $g(t)$. Different forms of $g(t)$ are available such as K-M estimator, ranks and $\log(t)$ (Therneau and Grambsch, 2000, pp. 127-152).

The Schoenfeld residuals are defined as;

$$r_i^* = r \text{ var}(\hat{\beta})\, r_{ji} \tag{4.50}$$

where r is number of events, var$(\hat{\beta})$ is the variance-covariance matrix of the parameter estimates in the fitted Cox PH model, and $r_{ji} = \delta_i \{x_{ji} - a_{ji}\}$, and $\hat{a}_{ji} = \frac{\sum_l x_{jl} \exp(\beta' x_l)}{\sum_l \exp(\beta' x_l)}$, x_{ji} is the value for j^{th} covariate for i^{th} individual in the study, and l is number of individual at risk at time t_i.

If the PH assumption is violated by a model (or indeed for a particular covariate in the model), then some of the options available are to consider fitting a stratified Cox PH model or a parametric model.

4.4.2 Residual Analysis

As mentioned in preceding paragraphs, the parametric models assigns a distribution for the survival function (and consequently the hazard function). In analyzing whether this assumed distribution best describes the data at hand, we opt to use two residuals; the modified Cox-Snell and Deviance residuals. Firstly, we define the general Cox-Snell residuals (Collett, 2014) as;

$$r_i = -\log[\hat{S}(t_i)] \tag{4.51}$$

where $\hat{S}(t_i)$ is the survivor function of the i^{th} individual at time t_i , i=1,.....,n individuals. The modified Cox-Snell residuals

$$rcs_i = \begin{cases} r_i & \text{if } T \leq \text{C} \\ r_i + 0.693 & \text{if } T > \text{C} \end{cases}$$

where T is the survival time, C is the censoring time, i ranges from 1 to n individuals.

The modified Cox-Snell residuals (rcs) have a property that if the model fits the data well, then these residuals follow the standard exponential distribution with mean one (Collett, 2014, p. 233). However, for $T > C$, the median(rather than the mean) is chosen for the unit exponential survival function for the excess survival time: $t(50) = \log 2 = 0.693$ (Collett, 2014, p. 233) as defined in equation for rcs_i above. The plot of rcs_i against the cumulative hazard of the residuals should closely follow the $x = y$ line if the model fits the data well.

Secondly, Deviance residuals are defined as;

$$D_i = \text{sign}[M_i] \left\{-2[M_i + \delta_i \ln(\delta_i - M_i)]^{1/2}\right\} \tag{4.52}$$

where $M_i = \delta_i - r_i$ are Martingale residuals, $\delta_i \in (0, 1)$ and $r_i = \hat{H}_0(t_i) \exp(\sum X\beta)$, $i = 1, ..., n$ and $\hat{H}_0$ is a Breslow estimator of the baseline hazard. Using both residuals, if the assumed distribution fits the data well, then the plot should show a linear trend with intercept zero (i.e a line of y=x).

4.4.3 Model Selection

The estimation of parameters in survival analysis is accomplished by maximizing either the partial likelihood, the profile likelihood, marginal likelihood or the full likelihood functions or through Bayesian estimation. The proportional hazards

model developed by Cox, (1972) and its variants described earlier on make use of the partial likelihood function. The process of selecting a parsimonious model makes use of these likelihood quantities.

There are several ways of choosing the best fit model. Informally, the likelihood values can be compared and the best model is chosen as the one that maximizes the likelihood. Semi-parametric and parametric models can not be compared to each other using the likelihood statistic because these two models use different forms of likelihood functions in estimating the model parameters. However, the more formal ways of comparing models are through the use of (i) likelihood ratio test (LRT) for nested models, (ii) Alkaike Information Criterion (AIC), (iii) Bayesian Information Criterion (BIC), and (iv) Deviance Information Criterion (DIC). The goal of these measures is to find the parsimonious model from a set of possible models by penalizing any additional parameters added to the model. These methods are described in many standard survival texts such as Hosmer and Lemeshow, (1999) and Ando, (2010) and in a recent review by Gelman, Hwang and Vehtari, (2014). Here, we only present their definition and we will mainly use LRT, AIC and DIC (for Bayesian models) for model selection because they are widely.

The LRT is a measure of how the log-likelihood values for two models being compared deviate from each other. It is calculated as;

$$\text{LRT} = -2(\log L_S - \log L_C) \tag{4.53}$$

where $\log L_S$ is the log-likelihood value of the simpler model and $\log L_C$ is the log-likelihood value for the complex model(model with additional parameters). The LRT can also be thought as a difference of deviances from two models (deviance $= -2(\log L)$. In our study we will use partial log-likelihood instead of full likelihood because models to compared are semi-parametric. The test of difference between the simper model and complex model is approximated by assuming that the LRT is a Chi-Square statistic;

$$\text{LRT} \sim \chi^2_{df_C - df_S}$$

where df_C and df_S is the number of degrees of freedom from the complex and the simpler model respectively.

The AIC values are calculated as follows ;

$$\text{AIC} = -2\log(L) + 2(p) \tag{4.54}$$

where L is the likelihood and p is the number of parameters estimated in the model. The best model for the data tend to minimize the AIC value (Gelman, Hwang and Vehtari, 2014).

The BIC is similar to AIC only that BIC more heavily penalizes additional parameters in the model by $\log(n)$, n is the sample size of the data. Its formula is

$$\text{BIC} = -2\log(L) + p\log(n) \tag{4.55}$$

where n is the number of uncensored observations (events). The model with a lower BIC values is the one that is chosen a best fit model.

Both AIC and BIC have an advantage that they can be easy computed and are widely understood in the conventional literature. However, the disadvantage for these criteria is that they work better in linear models and survival data with large sample sizes but not for censored data with small sample size (Hurvich and Tsai, 1986). To address this problem in censored data, some methodological work has been done including the work of Hurvich and Tsai, (1986), Faraggi and Simon, (1998) and Volinsky and Raftery, (2000) and Liang and Zou, (2008).

The DIC value from which a generalization of hierarchical modelling of the AIC is calculated as (Gelman, Carlin et al., 2010, p. 182);

$$\text{DIC} = D(\bar{\theta}) + 2p_D \tag{4.56}$$

where $D(\hat{\theta}) = -2\log(p(y|\hat{\theta})) + C$ is the deviance, (C is a constant that cancels out when comparing the models), $p_D = \bar{D} - D(\bar{\theta})$ is the effective number of parameters, $\bar{\theta} = E[\theta]$ is the posterior mean and $\bar{D} = E[D(\theta)]$ is the posterior mean deviance. Again, like AIC and BIC, the model with the lowest DIC value is selected among a set competitive models.

A relatively new widely applicable information criterion (WAIC) was proposed by Watanabe, (2010). The WAIC is defined as;

$$\text{WAIC} = -2(\text{lppd} + p_{WAIC}) \tag{4.57}$$

where lppd $= \sum_{i=1}^{n} \mathbb{E} \log p_{post}(\bar{y}_i)$ is the expected point-wise predictive density and p_{WAIC} is the effective number of parameters calculated same as p_D.

References I/4

Aalen, O., Borgan, O. and Gjessing, H. (2008) *Survival and Event History Analysis: A Process Point of View*, 1st, New York.

Adebayo, S.B. and Fahrmeir, L. (2005) 'Analysing child mortality in Nigeria with geoadditive discrete-time survival models.', *Statistics in medicine* 24 (5), pp. 709–28.

Alvaro-Meca, A. et al. (2012) 'Impact of comorbidities and surgery on health related transitions in pancreatic cancer admissions: a multi state model.', *Cancer epidemiology* 36 (2), e142–6.

Andersen, Per Kragh and Keiding, N. (2006) *Survival and Event History Analysis*, England: John Wiley & Sons Ltd.

Ando, T. (2010) *Bayesian Model Selection and Statistical Modeling*, Florida: CRC Press. Chapter 7.

Banerjee, S. and Carlin, B.P. (2004) 'Parametric spatial cure rate models for interval-censored time-to-relapse data', *Biometrics* 60 (1), pp. 268–275.

Banerjee, S. and Dey, D.K. (2005) 'Semiparametric proportional odds models for spatially correlated survival data', *Lifetime Data Analysis* 11, pp. 175–191.

Banerjee, S. and Carlin, B.P. (2003) 'Semiparametric spatio-temporal frailty modeling', *Environmetrics* 14, pp. 523–535.

Banerjee, S., Carlin, B.P. and Alan E, G. (2004) *Hierarchical Modelling and Analysis for Spatial Data*, 2nd ed., Florida: CHAPMAN & HALL/RCR.

Banerjee, S., Wall, M.W. and Carlin, B.P. (2003) 'Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota', *Biostatistics* 4, pp. 123–142.

Beyersmann, J., Allignol, A. and Schumacher, M. (2012) *Competing Risks and Multistate Models with R*, New York, NY: Springer New York.

Breslow, N.E. (1975) 'Analysis of Survival Data under the Proportional Hazards Model', *International Statistics Review* 45, pp. 45–58.

Brezger, A., Kneib, T. and Lang, S. (2005) 'BayesX: Analyzing Bayesian Structured Additive Regression Models', *Journal of Statistical Software* 14 (11), pp. 1–22.

Collett, D. (2014) *Modelling Survival Data in Medical Research*, Third Edit, Boca Raton.

Cox, D.R.(R. and Miller, H.D.(D. (1965) *The theory of stochastic processes.* Methuen, p. 398.

Cox, R.D. (1972) 'Regression Models and Life-Tables', *Journal of the Royal Statistical Society, Series B* 34, pp. 187–220.

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998) 'Model-Based Geostatistics', *RSS Applied Statistics* 47, pp. 299–350.

Diggle, P.J. and Ribeiro, P.J. (2007) *Modelbased Geostatistics*, New York: Springer.

Diva, U., Banerjee, S. and Dey, D.K. (2007) 'Modelling spatially correlated survival data for individuals with multiple cancers', *Statistical Modelling* 7, pp. 191–213.

Efron, B. (1977) 'The Efficiency of Cox's Likelihood Function for Censored Data', *Journal of Maerican Statistical Association* 72, pp. 557–565.

Faraggi, D. and Simon, R. (1998) 'Bayesian variable selection method for censored survival data', *Biometrics* 54, pp. 1475–85.

Fine, J.P. and Gray, R.J. (1999) 'A Proportional Hazards Model for the Subdistribution of a Competing Risk', *Journal of the American Statistical Association* 94 (446), p. 496.

Gelman, A., Carlin, J.B. et al. (2010) *Bayesian Data Analysis*, Second, Florida: CRC Press.

Gelman, A., Hwang, J. and Vehtari, A. (2014) 'Understanding predictive information criteria for Bayesian models', *Statistical Computing* 24, pp. 997–1016.

Henderson, R., Shimakura, S. and Gorst, D. (2002) 'Modeling Spatial Variation in Leukemia Survival Data', *American Statistical Association* 97, pp. 965–972.

Hinchliffe, S.R. and Lambert, P.C. (2013) 'Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions', *BMC Medical Research Methodology* 13(13).

Hosmer, D. and Lemeshow, S. (1999) *Applied Survival Analysis*, 2nd, New York: John Wiley & Sons.

Hougaard, P. (2000) *Analysis of Multivariate Survival Data*, New York.

Hurvich, C. and Tsai, C. (1986) 'Regression and time series model selection in small samples', *Biometrika* 76, pp. 297–307.

Ibrahim, J.G., Chen, M.-H. and Sinha, D. (2010) *Bayesian Survival Analysis*, New York: Springer.

Kalbfleisch, J.D. and Prentice, R.L. (2002) *The Statistical Analysis of Failure Time Data*, 2nd, Wiley Series in Probability and Statistics, Hoboken, NJ, USA: John Wiley & Sons, Inc.

Kaplan, E.L. and Meier, P. (1958) 'Nonparametric Estimation from Incomplete Observations', *Journal of American Statistical Association* 53, pp. 457–481.

Klein, J.P. et al. (2013) *Handbook of survival analysis*, p. 640.

Kleinbaum, D.G. and Klein, M. (2012) *Survival Analysis-A Self-Learning Text*, 2nd, USA: Springer.

Li, Y. and Ryan, L. (2002) 'Modeling Spatial Survival Data using Semiparametric Frailty Models', *International Biometric Society* 58, pp. 287–297.

Liang, H. and Zou, G. (2008) 'Improved AIC Selection Strategy for Survival Analysis', *Comput Stat Data Anal* 52(5), pp. 2538–48.

Matern, B. (1986) *Spatial Variation*, Berlin: Springer-Verlag.

Nardi, A. and Schemper, M. (2003) 'Comparing Cox and parametric models in clinical studies', *Statistics in medicine* 22, pp. 3597–3610.

Nathoo, F.S. and Dean, C.B. (2008) 'Spatial Multistate Transitional Models for Longitudinal Event Data', *Source: Biometrics* 64 (1), pp. 271–279.

Patel, K., Kay, R. and Rowel, L. (2006) 'Comparing proportional hazards and accelerated failure time models: an application in influenza', *Pharm Stat* 5(3), pp. 213–24.

Royston, P. and Parmar, M. (2002) 'Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects', *Statistics in Medicine* 21, pp. 2175–2197.

Taylor, B.M. and Rowlingson, B.S. (2014) 'spatsurv: an R package for Bayesian inference with spatial survival models', *Available at* https://arxiv.org/abs/1501.01665.

Therneau, T.M. and Grambsch, P.M. (2000) *Modeling Survival Data: Extending the Cox Model*, New York: Springer.

Volinsky, C. and Raftery, A. (2000) 'Bayesian information criterion for censored survival models', *Biometrics* 56, pp. 256–62.

Watanabe, S. (2010) 'Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory.', *Journal of Machine Learning Research* 11, pp. 3571–94.

Wienke, A. (2011) *Frailty models in survival analysis*, CRC Press.

Zhang, J. and Lawson, A.B. (2011) 'Bayesian Parametric Accelerated Failure Time Spatial Model and its Application to Prostate Cancer', *Journal of Applied Statistics* 38(2), pp. 591–603.

Zhang, M.-J. and Fine, J. (2008) 'Summarizing differences in cumulative incidence functions', *Statistics in Medicine* 27(24), pp. 4939–4949.

Part II

Modelling survival in HIV cohorts

Chapter 5

Quantifying mortality risk and survival in HIV patients diagnosed with KS during treatment initiation

---

Summary

In this analysis, we look at estimating survival in our HIV cohort. The main focus is on a subgroup of patients who are clinically diagnosed with Kaposi's sarcoma (HIV/KS) at the time of starting a lifelong HIV treatment (ART). With compromised immunity, this group of HIV patients is reported to have poor diagnosis compared to the general HIV population. However, within-group differences in prognosis are rarely reported for this ART group. Our aim is to understand more about the survival in this group and how it differs with patient covariates. To do

this, we consider both semi-parametric and parametric models. We conclude with assessing how the problem of loss to follow-up (LTFU) could affect the estimation of survival by conducting the sensitivity analysis on the approximate bound for the proportion of deaths among LTFU patients. Using this approximate bound, our model diagnostics and selection indicate that (i) a random effects Cox Model with Log-Gaussian frailties and (ii) a flexible parametric proportional hazards model, describe the risk of mortality in the HIV/KS patients well. This subgroup analysis can be used to inform targeted patient management treatment strategies, rather than the 'blanket-method' of treating all patients in the same way, and thus can provide the basis for formulating a more efficient triage system for care and treatment of patients.

5.1 Introduction

Malawi, like many other sub-Saharan countries, is one of the developing nations with a poorly resourced public health system. The country is characterized by high infant mortality and high mortality in 15-49 age group (Malawi NSO, 2008). It also has a heavy disease burden ranging from Human Immunodeficiency Virus (HIV), malaria, tuberculosis (TB) and an increasing trend of non-communicable disease cases such as cancer, cardiovascular diseases, respiratory disease and diabetes. (World Health Organization, 2008). A 2014 Country Report by the World Health Organization estimates that there are up to 55,000 new HIV infections per year in Malawi, with high national HIV prevalence among women (13%) than men (8%). Non-communicable disease co-morbidities such as cancer are common in HIV population, particularly Kaposi's sarcoma, breast cancer and cervical cancer (Gbabe et al., 2014; Msyamboza et al., 2012; Makombe, Harries and al, 2008).

In 2004, in the face of a growing HIV epidemic, the Malawi Ministry of Health with the help of development partners, introduced free HIV treatment and care services to all individuals diagnosed with the HIV virus (Malawi Ministry of Health, 2003). Consequently, HIV care and treatment services were expanded to include HIV testing and counselling services, services aimed at prevention of HIV transmission from mother to child during pregnancy and breastfeeding, provision of tuberculosis (TB) and antiretroviral therapy (ART) treatment. The purpose of HIV treatment is to help boost the immunity system of HIV-positive patients, and

if taken correctly, it can prolong one's time to death due to Acquired Immunodeficiency Syndrome (AIDS). This has resulted in a significant decrease in the number of deaths in general HIV population (Malawi Minstry of Health, 2014b).

In Malawi, previous studies by Mwinjiwa et al., (2013), Msyamboza et al., (2012) and Makombe, Harries and al, (2008) have reported on the incidence of Kaposi's sarcoma (KS) and on the differences in treatment outcomes among patients with KS. However, studies highlighting the survival prognosis from sub-group analysis of HIV patients such as patients with both HIV and KS (HIV/KS) are lacking. In this chapter, we attempt to fill this gap by studying the within-group differences in survival of HIV/KS patients in Zomba, Malawi. We argue that such analyses are necessary to help clinicians and public health specialist in their efforts to triage patients based on their prognosis. The Zomba district has one of the highest HIV prevalences in Malawi, estimated at 16% in 2010 (Malawi NSO, 2008). As of 2008, the life expectancy in the district was higher in women (51 years) compared to males (47 years). This life expectancy is lower than what was reported in Europe and Asia around same period of time (Adetunji and Bos, 2006). High HIV mortality and high disease morbidity are the main factors for such low life expectancy in SSA region.

Also, we address the issue of loss to follow-up (LTFU). Recently, there has been a growing interest in addressing the impact of loss to follow-up data in the estimation of survival in HIV cohorts (Freeman, Semeere, Wenger et al., 2015; Kiragga et al., 2013; Henriques et al., 2012). The methods proposed to address the impact of LTFU on survival estimates have largely depended on the number of deaths in

a sample of LTFU patients (Yiannoutsos, Johnson et al., 2012; Yiannoutsos, An et al., 2008). Others have proposed treating LTFU as a competing event to the occurrence of death (Graham et al., 2013). In the current analysis, we consider the use of a sensitivity analysis by comparing the overall survival to cohort data that is treated as an approximate upper bound of the survival curve. With this approach, it is also possible to get an approximate lower bound for the proportion of LTFU patients that are actually deaths. We apply this novel approach by retrospectively analysing routinely collected HIV data. In the presence of high loss to follow-up, our aim is to improve the estimation of survival in KS patients and further improve our understanding of KS epidemiology in HIV population.

5.2 Objectives

In this chapter, we address the first two objectives of this thesis work as listed in section 2.2. The two objectives are;

1. Understand the relationship between individual risk factors for mortality amongst HIV/KS patients.

2. Provide an approximate lower bound on the proportion of patients recorded as lost to follow-up that are really deaths.

5.3 Data Description

We consider cohort data from HIV patients that were additionally diagnosed with Kaposi's sarcoma at the time of treatment initiation. The data are taken from an HIV clinic at Zomba Central Hospital (ZCH), in the eastern region of Malawi. The main outcome was all-cause death. However, we also fitted models with being lost to follow-up as a competing outcome to the occurrence of death.

During the study period (2004 -2011), all persons in the district who tested HIV positive at HIV testing centres were referred for an ART initiation eligibility assessment either to the HIV clinic at Zomba Central Hospital or referred to other peripheral HIV clinics. The ART eligibility assessment was based the World Health Organisation (WHO) HIV clinical staging and using CD4 count measurements (Malawi Ministry of Health, 2003). HIV positive persons were started on ART immediately if they were categorized to be in WHO stage 4 or they were in Stage 3 with TB episodes. Also, if CD4 count was used as an assessment criterion, then patients were started on ART if their CD4 count was less than 250 cells/μL (Malawi Ministry of Health, 2003). The CD4 threshold changed over time: from 2004-2006: 200; 2006-2011: 250; 2011-2014: 350; from 2014 onwards: 500. Overall, the majority of patients in this cohort were started on ART largely based on WHO clinical staging criteria. The reason for this is being that the laboratory diagnostic services were not generally available during this period (Malawi Ministry of Health, 2003).

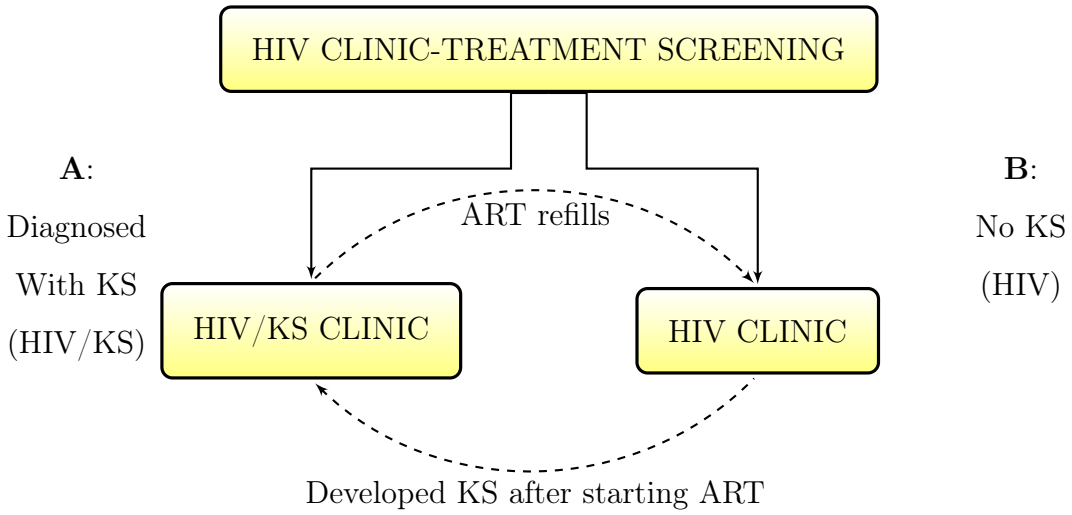

Figure 5.1: Treatment initiation and follow up for HIV and HIV/KS patients at Zomba HIV Clinic. The starting point is after patients are deemed eligible for treatment.

In addition to their HIV treatment, HIV patients diagnosed with Kaposi's sarcoma (HIV/KS) were also given chemotherapy to treat KS unless the extent, location, and symptomatology were not severe. Figure 5.1 is a flowchart showing how patients were enrolled at HIV and HIV/KS clinic for follow-up and clinical management.

The data comprise approximately 14, 884 adult HIV-positive patients that were enrolled in HIV clinic care in Zomba from 2004 to 2011. Three percent (615) of these adult HIV patients were diagnosed with Kaposi's Sarcoma (skin tumour) during ART eligibility screening stage (Malawi Ministry of Health, 2011). Baseline and follow-up demographic and clinical information including the following were collected using a standardized national ART clinic form; age, sex, address (village name and traditional authority), treatment start date, occupation, TB history, KS status, outcome, outcome date. We also obtained geo-coordinates data from Malawi National Statistical Office and matched the coordinates to the patients'

address by matching the names of the locations in both data sets. We calculated the Euclidean distance (the straight-line distance) from each patients' village or reported residency location to the HIV clinic at ZCH. The locations of patients were represented at the village level.

$$||(x, y) - (a_i, b_i)|| = \sqrt{(x - a_i)^2 + (y - b_i)^2} \tag{5.1}$$

where ||.|| denotes distance, (x, y) are the coordinates of the clinic and (a_i, b_i) are coordinates associated with patient i, with i from 1 to n individuals. Geo-coordinates for special features in a village were used. These features included church, market place, school, dam, hospital, military, and police.

The issue of accessibility to the HIV clinic makes the Euclidean distance interesting as a predictor of survival. However, the disadvantage is that this is measured as a straight-line distance which in many cases is not reflective of how patients travel to the clinic, hence underestimating the travel distance by patients. Distance was categorised into two groups: $\leq$ 8 km and >8 km, with the category based on radius of the catchment area for the clinic (based on the unpublished interviews with the clinicians in Zomba).

Table 5.1 is a summary of definitions for all the covariates considered in the study. In order to come up with a lower bound of proportion of deaths in LTFU as proposed in Section 5.6, a reference dataset with complete outcome data was needed. Unfortunately, such complete datasets are rare in Africa as the issue of LTFU is a one of major challenges in cohort studies, mainly due to lack of

Table 5.1: List of Covariates as defined in section 2.6

Variable	Categories
Age (yrs)	Continuous
Sex	1= Male, 2= Female
ART period	1= 2004-2007, 2= 2008-2011
TB status	1= TB(current/past), 0= No TB
Occupation	1= Employed (includes teachers & health care workers), 2= Self-employed(includes business persons & farmers), 3= Student, 4= Other(occupation not specified)
Distance (Km)	1= $\leq$8 Km, 2= >8 Km
KS status	1= With KS, 0= NO KS
Outcome	1= dead, 2= active follow-up, 3= lost to follow up
Times	Continuous

national vital registration systems. As such, we obtained permission to extract KS data from the SEER database of the National Cancer Institute, United States, was obtained from http://seer.cancer.gov/. These data date back from 1973 to 2010. We restricted the period and age to time between 2004 and 2010 and maximum age of 68 years respectively in order to match the age range of the Malawi data. The following variables were selected;

- **Age at diagnosis**. It is recorded as `AGE_DX` in the database.
- **Status**. In the database, it is `STAT_REC` and the categories were 1 for "Alive" and 4 for "Dead".
- **Time**. This is the time of follow-up in months and is recorded as `SRV_TIME_MON` in the SEER database.

The total number of patients included in the SEER data were 2,019 and the ages ranged from 19 to 68. The median follow-up time was 28 months and maximum follow-up time was curtailed at 80 months in order to match with our data.

5.3.1 Inclusion and Exclusion Criteria

We included data for all adult HIV/KS patients and data for non-KS HIV patients for comparison. We excluded patients aged below 14 years because KS in children is not a big problem and its diagnosis is not well developed as in adults (Bruin and Stefan, 2013). We also excluded patients referred from neighbouring districts because our study is limited to Zomba district. In addition, patients with missing date of initiation or sex were excluded. Consequently, we excluded 9% (n=56) and leaving 559 HIV/KS patients in the final analysis.

5.4 Survival Modelling

This section contains details of the analysis methods and models fitted the right-censored data used in this analysis. We first present the semi-parametric models; the standard Cox PH model and the two Cox frailty models. The last subsection describes the AFT models considered in this analysis.

5.4.1 Semi-parametric Models

We started by conducting a descriptive analysis of covariates included in the analysis. We also plotted K-M curves based on Equation 4.4 and used the log-rank test (see Equation 4.5) to compare survival amongst different groups of patients. In order to assess risk factors of mortality, we fitted different proportional hazards regression models with the aim of choosing the best-fit model.

The following are the variants of the semi-parametric proportional hazards (PH) models used in this analysis (see Equation 4.9 and Equation 4.28 for definitions);

- Standard Cox PH as in Equation 4.9
- Cox/Gamma frailty model, where Z in Equation 4.28 has a gamma distribution (see Equation 4.30)
- Cox/Log-Gaussian frailty model, where Z takes a log-Gaussian distribution (see Equation 4.31)

We compared results from these three semi-parametric PH models and the best-fit model was selected based on the model with the minimum value of the partial likelihood value (PLik) since these models do not use a full likelihood function.

5.4.2 Parametric models

In addition, we fitted and compared results of six parametric models; four accelerated failure time (AFT) models and a flexible proportion hazards model (FPH) and a flexible proportion odds model (FPO). The two latter models were based on Royston & Parmar spline model (Royston and Parmar, 2002). The AFT models take the form as defined in subsection 4.2.2 and we considered: Weibull, Log-Normal (LogN), Log-logistic (LogLog) and the Gamma models. The flexible models formulated as defined in Equation 4.21 and Equation 4.22. The best fit model from a set of parametric models fitted was selected using Alkaike Information Criteria

(AIC) as defined in chapter 4 in Equation 4.54. The best model for the data minimizes the AIC value.

All the analyses were conducted in R-Software using the `survival` and `flexsurv` packages. We reported the corresponding 95% confidence intervals (95% CI) for each estimated $\hat{\beta}$.

5.5 Descriptive Analysis

In this section, we provide a detailed descriptive analysis of the data by presenting the overall survival among patients and a detailed description of each covariate to be included in the survival modelling.

5.5.1 Overall Survival

A total of 105 all-cause deaths were recorded among 559 HIV/KS patients representing a seven-year death rate of 19% (Table 5.2). The death rate among HIV patients without KS at baseline was 7.8%. The majority of all patients were aged between 30 and 45 years and 48% of all recorded deaths were also seen in this age group. For all patients, the age ranged from 15 to 68 years, with on average male patients being older than females (Male: 36.3, Female: 32.7, not shown in Table 5.2).

Table 5.2: Baseline demographic, clinical characteristics and outcomes of HIV/KS patients.

Covariate	Total Patients n(%total)	Patients alive n(%total)	LTFU Patients n(%total)	No. of Deaths n(%total)	Median Event time† (25%,75% quantiles)
Total	559 (100)	231(41.3[a])	223(39.9[b])	105(18.8[c])	5.1(2.0, 11.8)
Age (yrs)					34*(28, 42)
15-29	178(31.8)	63(27.3)	81(36.3)	34(32.4)	3.8(1.8, 10.05)
30-44	303(54.2)	137(59.3)	116(52.0)	50(47.6)	5.9(2.4, 10.8)
45+	78(14.0)	31(13.4)	26(11.7)	21(20.0)	6.4(1.9, 15.4)
Sex					
Female	252(45.1)	111(48.1)	92(41.3)	49(46.7)	4.9(1.7, 12.9)
Male	307(54.9)	120(51.9)	131(58.7)	56(53.3)	5.8(2.3, 10.6)
ART Period					
2004-2007	259(46.3)	80(34.6)	110(49.3)	69(65.7)	6.2(2.3, 14.8)
2008-2011	300(53.7)	151(65.4)	113(50.7)	36(34.3)	4.2(1.9, 6.8)
Distance					7.4*(3.7, 14.6)
$\leq$*8 kms*	288(51.5)	125(54.1)	114(51.1)	49(46.7)	5.5(2.9, 10.3)
$>$*8 kms*	271(48.5)	106(45.9)	109(48.9)	56(53.3)	5.0(1.9, 13.3)
TB Status					
No TB	528(94.5)	219(94.8)	213(95.5)	96(91.4)	5.3(1.9, 12.9)
With TB	31(5.5)	12(5.2)	10(4.5)	9(8.6)	4.4(1.9, 6.7)
Occupation					
Student	10 (1.8)	5(2.2)	3(1.3)	2(1.9)	9.4(6.6, 12.3)
Employed	38 (6.8)	12(5.2)	17(7.6)	9(8.6)	5.1(3.9, 7.2)
Self-Employed	371(66.4)	157(68.0)	149(66.8)	65(61.9)	
Other	140(25.0)	57(24.7)	54(24.2)	29(27.6)	4.4(1.9, 10.9)

*`median age and distance`, †`Time in months`

a, b, c `these are row percentages`

Half of the 105 deaths occurred within five months after starting ART. Younger patients (age $<$ 30 years) had shorter crude median time to death (3.8 months) compared to other age groups. A total of 223 (39.9%) patients were categorised as a loss to follow-up (Table 5.2). A larger proportion of these LTFU patients was recorded among patients who were either middle-aged (30-44), had no TB or were self-employed.

In Table 5.2, we also note that nearly half of the patients (46%) in our cohort entered into treatment programme between 2004 and 2007. The percentage of deaths was higher in patients who started treatment in the earlier period – 65% of deaths were in those entering treatment 2004-2007. Contrastingly, the crude median time to death was greater (but not significantly so) among patients who entered into treatment 2004-2007.

We also compared two different empirical cumulative distributions (ECDF) and survival functions (see Equation 4.1) between the HIV/KS patients and the 14,325 non-KS HIV patients. At a glance, it can be seen in Figure 5.2 (a) that HIV/KS tend to have a lower ECDF than HIV patients when we consider event times only. However, if we include the censored times, the difference is no longer obvious as shown in Figure 5.2 (b), although arguably HIV/KS still have relatively lower ECDF. However, these graphs highlight how the use of the ECDF may not be the proper distribution to explain the differences in survival because it does not account for censoring. In section 5.4, we have outlined appropriate analysis methods that handle the influence of censored times including the use of survival and hazard functions.

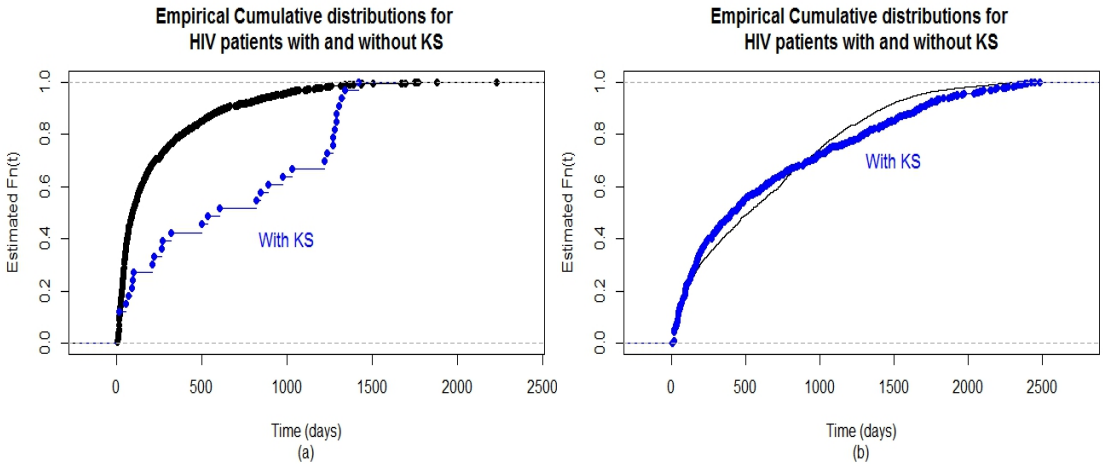

Figure 5.2: Comparison of empirical distribution for HIV/KS (blue) and non-KS patients (black). Plot (a) is a plot for event times only while plot (b) includes both event times and censored times.

Using the K-M estimator given in Equation 4.4, we compared the estimated the crude K-M survival curves between HIV and HIV/KS patients. The survival curves are presented in Figure 5.3. Among the non-KS patients, the survival is higher compared to HIV/KS patients. There seems to be a rapid deterioration of prognosis amongst HIV/KS patients compared with non-KS patients. However, the overall survival probability in both groups remained high, approximately above 70% although the uncertainty (95% CI) in the survival increased with increasing follow-up time.

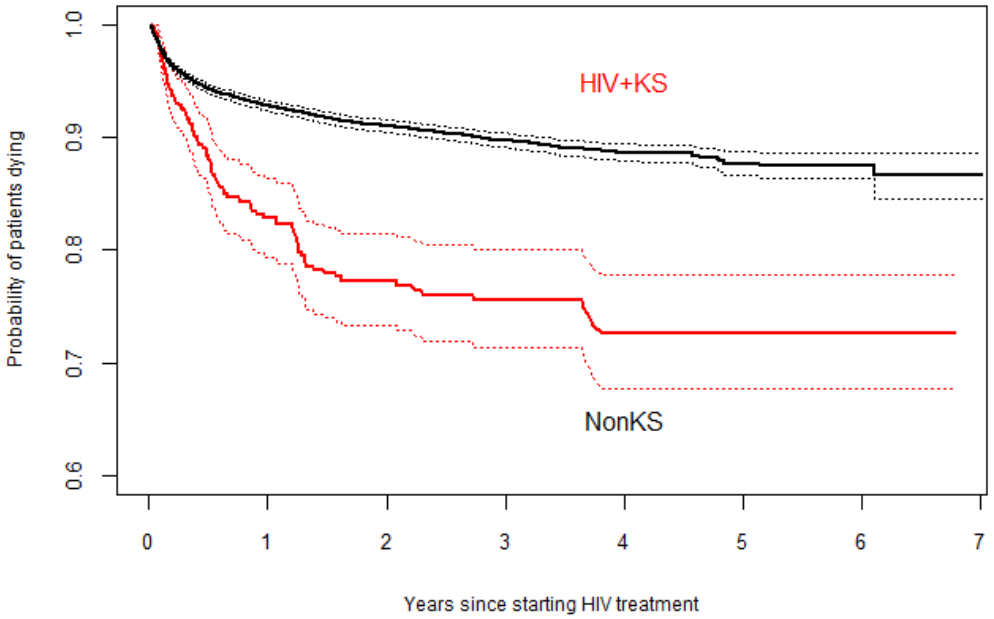

Figure 5.3: Survival curves (95% CI - dotted lines) based on K-M estimates. Higher survival can be noticed in HIV group than in HIV/KS patients.

5.5.2 Prevalence of KS at ART initiation

Overall, the percentage of patients starting ART with KS decreased over time as shown in Figure 5.4. The highest prevalence was recorded in 2005 when 70 (11% out of 615) patients had KS at the time of starting ART. It can be noted that thereafter the annual prevalence fell steadily to about 2% by 2011. This decline could be due to the expansion of HIV treatment services in the district since Kaposi's sarcoma is associated with advanced disease progression in HIV populations.

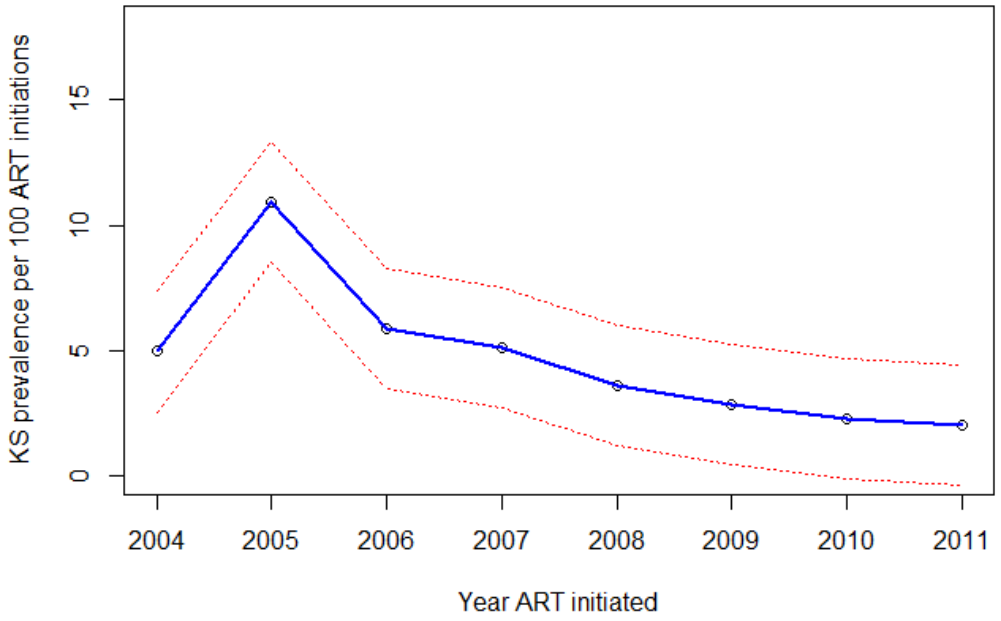

Figure 5.4: Prevalence of Kaposi's sarcoma during initiation of a lifelong HIV treatment (ART) over a 7 –year period. The dotted lines are the 95% CI, with the significant decrease in prevalence diminishing by year 2008.

5.5.3 Age

The ages for the HIV/KS patients ranged from 15 to 68 years, with a median age of approximately 35 years. Amongst the 105 deaths, half of the patients were aged 34 years or below. The plot in Figure 5.5 highlights the fact that the difference in median age between patients that died and patients with censored times is coincidental as the confidence intervals for mean age overlap for these two outcomes. Table 5.2 shows that three-quarters (75%) of patients that died were aged below 45 years. This is the same age group 15-49 that is considered to be sexually active and the majority of patients in HIV care are from this age group, as shown in the density plot on the left panel of Figure 5.6. Age is an important

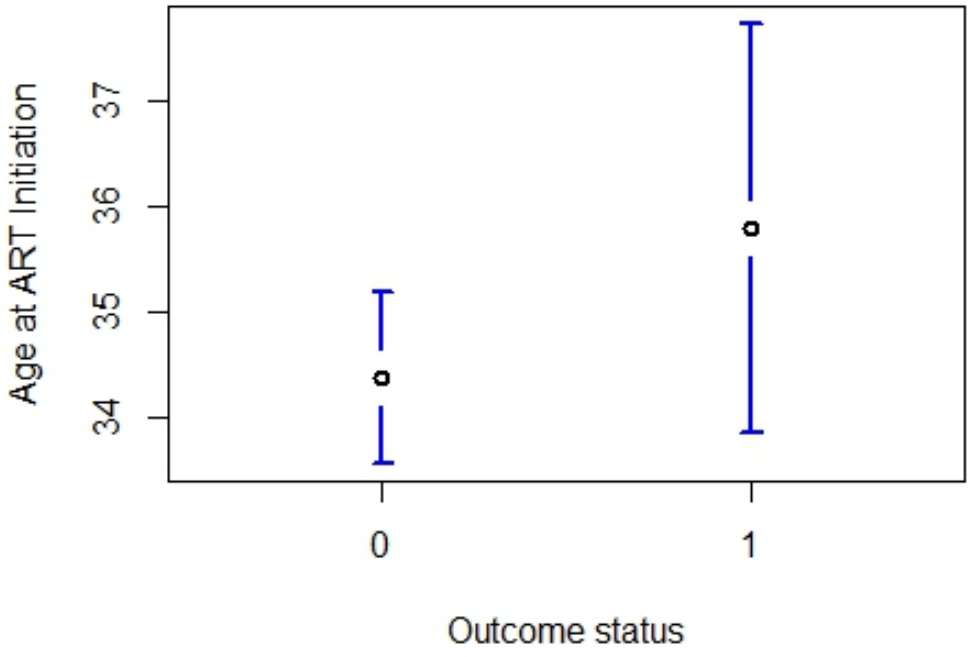

Figure 5.5: Mean age (with 95% confidence interval) of patients that experienced the event (labelled 1 on x-axis) and those with censored times (labelled 0 on x-axis). Note the overlapping confidence intervals.

factor in predicting survival because of the ageing effects, that is, the older one becomes, the more frail they become.

5.5.4 Sex

There was a total of 252 females and 308 male HIV/KS patients. Among the 252 (45%) females, 49 (19% of all females) died and their crude median event time was 4.9 months with 25% and 75% quantile range of approximately 2 and 13 months respectively Table 5.2. On the other hand, males tended to have longer crude median event time compared to females (5.8 vs 4.9). In addition, 60(20%) of 308

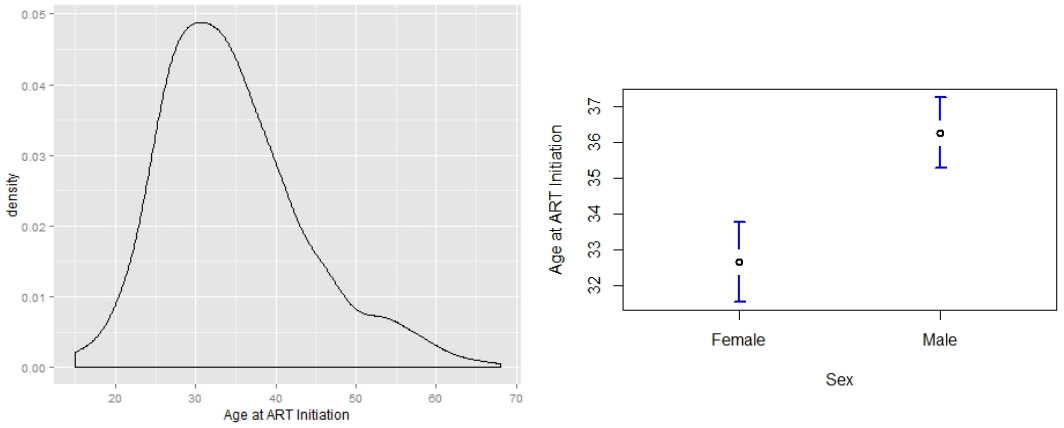

Figure 5.6: The plot on the *left* shows age distribution of HIV/KS patients and the plot on the *right* is a comparison of mean age of male and female HIV/KS patients. On averege, male patient seems to be older than females.

males died and half of them died within 5.8 months after starting HIV treatment. On average, male patients were older than females, as shown in the right panel of Figure 5.6 and the left panel of Figure 5.7.

To compare the number of observed and expected deaths in males and females, we used the log-rank statistic, which is approximately a Chi-Square statistic with one degree of freedom using the following;

$$U_L = \sum_{j=1}^{r} d_{ij} - e_{ij} \tag{5.2}$$

where d_{ij} is the number of deaths in group i at time j, and $e_{ij} = n_{ij}d_j/n_j$ is the expected number of individuals who die at time t_j in groug i (Collett, 2014, p. 233).

The statistic for comparing survival between two groups is

$$W_L = U_L/\sqrt{V_L}, \quad \text{with} \quad V_L = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)} \tag{5.3}$$

where V_L is the variance and $W_L \sim \chi_1^2$. Table 5.3 provides summary results from the log-rank tests for different covariates. At 5% significance level, the results for comparing number of deaths among male and female patients indicate that non-significant differences in observed and expected number of deaths in both males and females ($p = 0.949$), although more deaths were observed in male patients than expected. This observation is emphasized by visual output using Figure 5.8 which is a plot of survival curves male and female patients. We also note there is a steep decline in survival probability in both sexes during the early days of ART treatment.

Table 5.3: Using Log-rank test to compare the observed and expected number of deaths by sex, Treatment Period, TB history and Occupation.

Covariate	#patients	#observed deaths	#Expected deaths	p-value
Sex				
Females	252	49	49.3	0.949
Males	307	56	55.7	
Period				
2004-2007	259	69	54.2	0.003
2008-2011	300	36	50.8	
TB Status				
No TB	528	96	100.4	0.0547
TB	31	9	4.86	
Occupation				
Employed	38	9	7.62	0.76
Self-Employed	371	65	70.07	
Student	10	2	1.52	
Other	140	29	25.79	

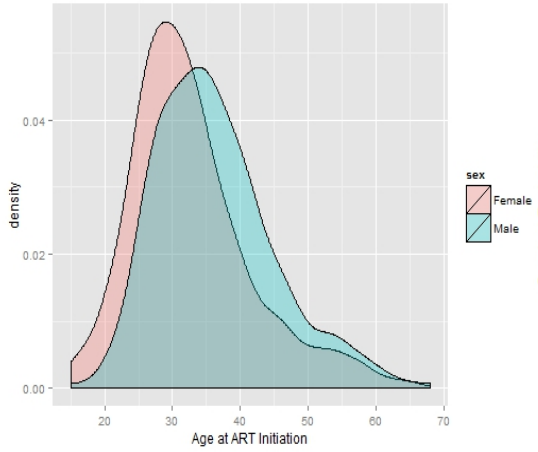

Figure 5.7: Density distribution of age grouped by sex

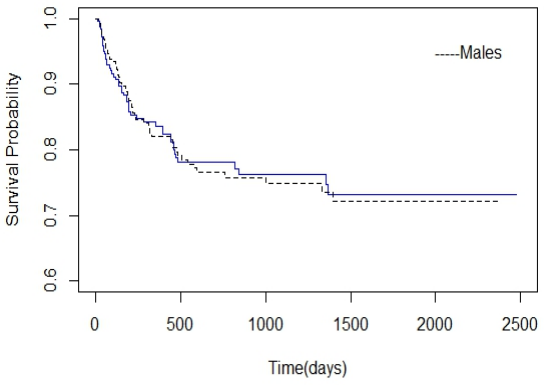

Figure 5.8: Survival curves for males and females

5.5.5 Period of starting HIV Treatment

In 2008, the Ministry of Health in Malawi revised the HIV treatment guidelines to improve the way patients were managed (Malawi Ministry of Health, 2008). Following the provision and expansion of HIV services, one of the questions of interest for clinicians and public health specialists, would be to find out if patients' prognosis has improved as treatment and routine clinical management and care of the HIV infected patients continues to improve over time. In this regard, we were interested in comparing the statistics and survival between the periods 2004 - 2007 and 2008 - 2011.

From Table 5.2, over half of HIV/KS patients started treatment between 2008 and 2011. Among the 105 patients that died, over 60% had started treatment between 2004 and 2007, of which half of them died during the first six months on treatment. Interestingly, among the 36 patients that died during the 2008 - 2011

period, half of them died during the first four months after starting HIV treatment, two months faster compared to those who started treatment before 2008. As seen from Table 5.3, log-rank tests results indicate a significant increase in the number of expected deaths amongst those that started treatment before 2008 (p-value =0.003). This difference in survival was more pronounced just after one year, as shown in Figure 5.9. Table 5.3 provides summary results from the log-rank tests for different covariates. At 5% significance level, the results for comparing the number of deaths among male and female patients indicate that non-significant differences in observed and expected number of deaths in both males and females ($p = 0.949$), although more deaths were observed in male patients than expected. This observation is emphasized by visual output using Figure 5.8 which is a plot of survival curves male and female patients. The curves cross each other suggesting a non-significant difference in survival. We also note there is a steep decline in survival probability in both sexes during the early days of ART treatment.

5.5.6 Distance to HIV clinic

At the time of this study, the HIV clinic at Zomba Hospital was the only clinic in the district that provided clinical care to patients diagnosed with both HIV and KS. As a result, all HIV/KS patients were required to seek and access treatment services at Zomba HIV clinic.

We computed the Euclidean distance from patient's village-level point to Zomba HIV clinic using Equation 5.1. Most urban settlements around the Zomba town

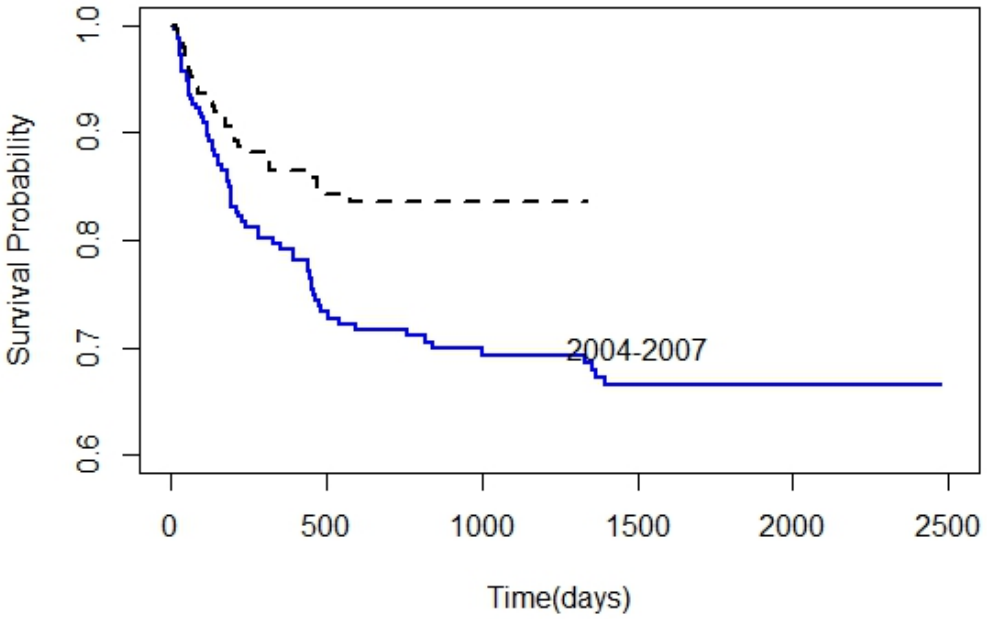

Figure 5.9: Survival curves for patients that started in 2004 - 2007 period and 2008 - 2011 period. Higher survival is noticed amongst those that started in latter period.

lie within an 8 kilometre-radius. Hence, we categorized the distance into two groups; within 8 km and over 8 km. Our main interest was to assess whether this covariate relates to patient outcomes considering that in most low-income countries patients travel long distances to access health services. However, using PubMed and Google Scholar engines, we did not find any literature relating distance to clinic and survival outcomes. Nevertheless, we hypothesized that distance could influence attendance to clinical appointments, with more patients living far from the clinic being prone to missing appointments. This could lead to poor treatment adherence and increase the risk of death.

The distances travelled by patients to the clinic ranged from <1 km to 31.4 km. In Table 5.2, it is seen that approximately 53% of patients that died were coming from

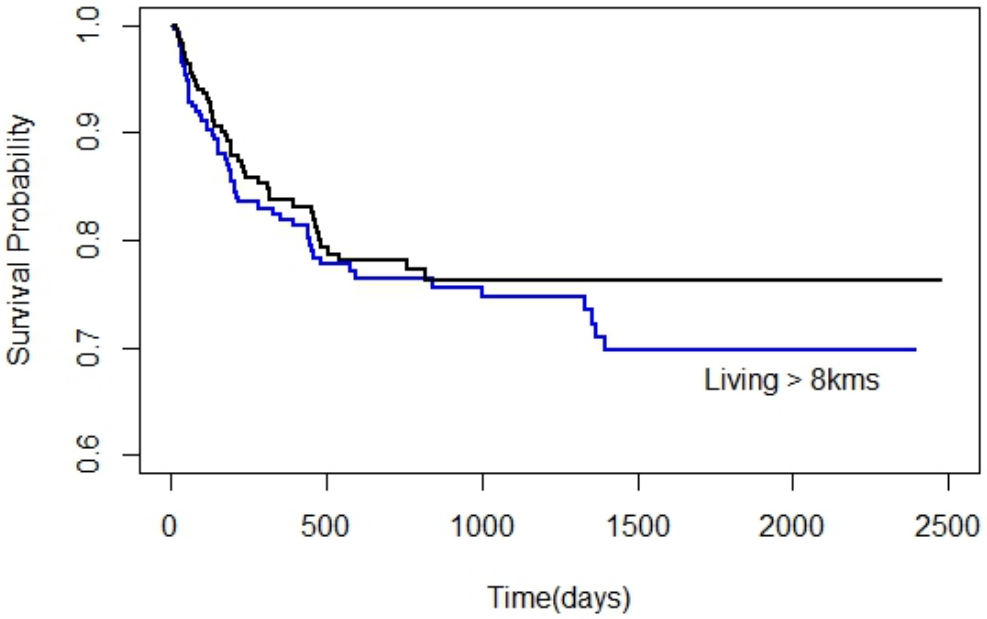

Figure 5.10: Survival curves for HIV/KS patients by distance to the clinic. The plots cross each at least at three separate occasions.

a distance of 8 km or more. The overall median distance for all HIV/KS patients was 7.4 km. Despite the differences seen in the survival curves in Figure 5.10, the log-rank test provided a non-significant p-value of 0.38, indicating no difference between the observed and expected number of deaths both distance categories. Like other factor variables, at this point we did worry about the proportion of deaths amongst LTFUs by distance and just treated all LTFUs as censored.

5.5.7 TB Status

Among the HIV/KS patients, there were only 31 (9%) patients with TB, of which 9 (29%) died during the study period. Mortality was lower amongst patients without TB, about 18%. Furthermore, patients with TB had a shorter crude median time

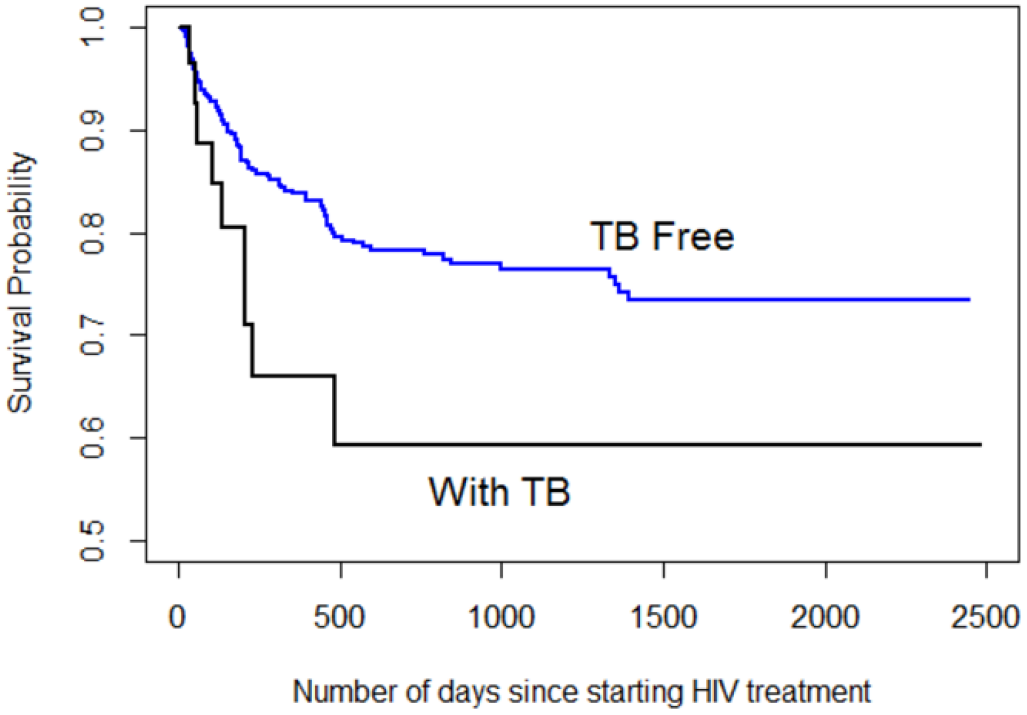

Figure 5.11: Survival curves for patients with and without TB. Having TB is associated with a lower survival.

to death, shorter by one month compared to the non-TB patients as shown in Table 5.2.

A log-rank test in Table 5.3 also indicated that the number of observed deaths amongst the TB group was marginally insignificant at 5% level (p-value = 0.0547). Figure 5.11 shows the K-M survival plots by TB history, and higher survival is observed in TB free patient group.

5.5.8 Occupation

The majority (66%) of HIV/KS patients were self-employed, of which 65 (18%) died. Half of self-employed patients died within the first 6 months of starting

treatment. There were 38 teachers and health care workers (categorized as employed) with HIV and KS and approximately 24% died during the study period. More deaths than expected were observed among patients that were either employed, were students or those categorised as "Other" (Table 5.3). Despite this observation, the differences were not statistically significant at 5% (p-value$= 0.76$).

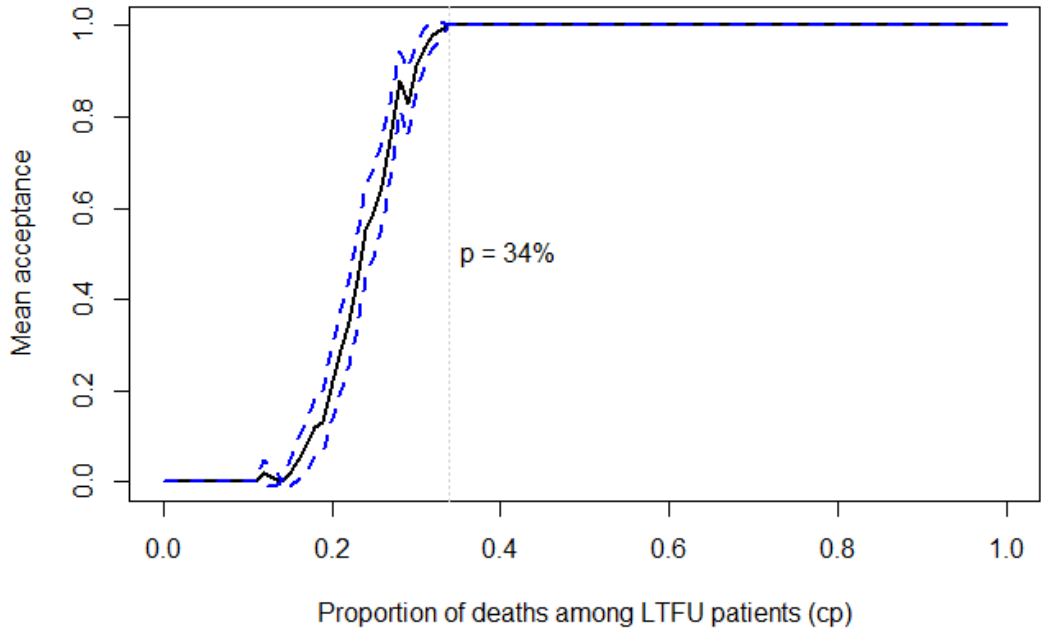

Figure 5.12: Proportion of times the Zomba curve was below California curve. The blue dotted lines are the 95% lower and upper bounds of the proportion satisfying $\hat{S}$(80in Malawi) $\leq \hat{S}$(80in US).

5.6 Accounting for loss to follow-up and determining a lower bound for the proportion of LTFUs that are deaths

In the Zomba KS dataset, 223 (39.9%) patients were lost to follow up meaning that their vital status was not ascertained during the study period. Since some of the LTFUs may be deaths, we used a novel way of understanding and accounting for potential deaths among the patients who were lost to follow-up. Other studies in sub-Saharan Africa have also reported high loss to follow-up and many of these studies have reported that in fact many of these LTFUs may be deaths (see Section 3.4).

We began by assuming that all LTFUs were censored observations. However, when we compared the resulting survival curve to that of US population (described in Section 2.6), we found that the Malawi cohort was doing much better. This was surprising since we would expect survival in the Malawi cohort to be worse compared to the US cohort, so we consider the US data as providing a sort of the "'upper bound"' for the survival curve. However, we do not know the proportion of LTFUs that are really deaths in our cohort.

Suppose we have a candidate proportion (c_p) of deaths among LTFUs. For any c_p, we can randomly change the $223 \times c\_p$ patient's status to death and the remainder be censored. Using this modified data, we would get estimates of $\hat{\beta}_1$ from the model. Repeating this process with a different $223 \times c\_p$ patient's

statuses modified, we would get another $\hat{\beta}_2$. Repeating this, say k times, we get;

$$\hat{\beta}_1, \hat{\beta}_2, \ldots, \ldots, \hat{\beta}_k$$

and we can combine these to get an overall $\hat{\beta}$ for this given c_p.

For each c_p and simulated dataset, we compared the 80 month survival to that from the US cohort. The time point of 80 months was chosen as an example based on the maximum time point when the K-M estimate become constant in the Malawi cohort. We found an approximate lower bound for c_p as the minimum value of c_p such that the 80 month survival in all simulated datasets was below that of the US cohort.

In summary, our *Comparison approach* of identifying an approximate lower bound for mortality rate among LTFUs was conducted in 6 main steps;

Step 1: We created a list of candidate proportions (c_ps) of deaths amongst LTFU patients from 0 to 1 increasing by 0.01 units. As a result 101 c_ps were used.

Step 2: For each selected c_p, we generated 100 different datasets in which $223 \times c\_p$ of the LTFUs follow-up times were changed to event-time otherwise they were censored. This means the number of deaths increased gradually with increasing proportion (c_p).

Step 3: A Cox model was fit using this updated dataset.

Step 4: We also created a 100 x 101 matrix whose entries took the value of 1 if the survival probability $\hat{S}(80)$ based on the K-M curve at 80 months (S(t=80)) for Zomba cohort was less than the survival probability for California cohort i.e. the matrix entries were defined by an indicator function as;

$$I(S_{ZA}(t=80) < S_{CAL}(t=80)) = \begin{cases} 1, & \text{if TRUE} \\ 0, & \text{Otherwise} \end{cases}$$

Step 5: We computed the proportion of times the K-M estimate (at month 80) for Zomba patients was less than California patients.

Step 6: We plotted the candidate proportion against the proportion calculated in Step 5. The cut-off value of c_p is the value of c_p corresponding to the point when the mean acceptance first records a 100% acceptance. This is our lower bound for c_p.

Using this selected value, we conducted sensitivity analyses for the two final selected models (semi-parametric and parametric). We sampled 1000 times with c_p% of patients assigned as 'deaths' during each sampling. This is equivalent to fitting 1000 models and saving the estimated β for each covariate in a matrix, with covariates as rows and estimates as columns ($p \times 1000$). The final $\hat{\beta}$ were estimated as $\mathbb{E}(\hat{\beta}) = \mathbb{E}[\mathbb{E}(\hat{\beta}|Y)]$ and the variance of $\hat{\beta}$ using the law of total variance defined as

$$\text{Var}(\hat{\beta}) = \underbrace{\text{Var}\left\{\mathbb{E}(\beta|Y)\right\}}_{\text{Var}(\beta)} + \underbrace{\mathbb{E}_Y\left\{\text{Var}(\beta|Y)\right\}}_{\text{mean}\{\text{Var}(\beta)\}} \tag{5.4}$$

where Y denotes the random sample data. Using $\mathbb{E}(\hat{\beta})$ and $Var(\hat{\beta})$ we were able to construct the confidence intervals.

We compared the crude survival estimates of our HIV/KS cohort data to the KS cohort from USA, California described in Section 2.6. We followed the steps as described above. The USA dataset was used because it had complete and updated patient outcomes, of which it difficult to find datasets of such quality in Africa which have high rates of loss to follow-up (Freeman, Semeere, Wenger et al., 2015). The main limitation of using these USA data is that these come from two different healthy setting with the USA data expected have more healthier population than the Malawi data. However, using these data gives us an approximate upper bound on survival on which we can compare with our data.

Figure 5.12 shows the value of the candidate proportion (c_p) against the proportion of datasets for which the simulated $\hat{S}(80\text{in Malawi}) \leq \hat{S}(80\text{in US})$. In order to help make the survival curves from the USA data and the Malawi data comparable, we ensured that the following were the same: same age range of 15 - 68; same study period, 2004 - 2010; and same maximum follow-up time of months.

We truncated the follow-up period at 80 months and also restricted age to less than 69 years to ensure comparable $\hat{S}$ from the two cohort data.

The comparison results showed that a minimum of 34% of deaths should be assumed among LTFU patients (Figure 5.12). Any proportion $>$ 34% resulted in the Zomba curve to be consistently below the California curve as measured at month 80.

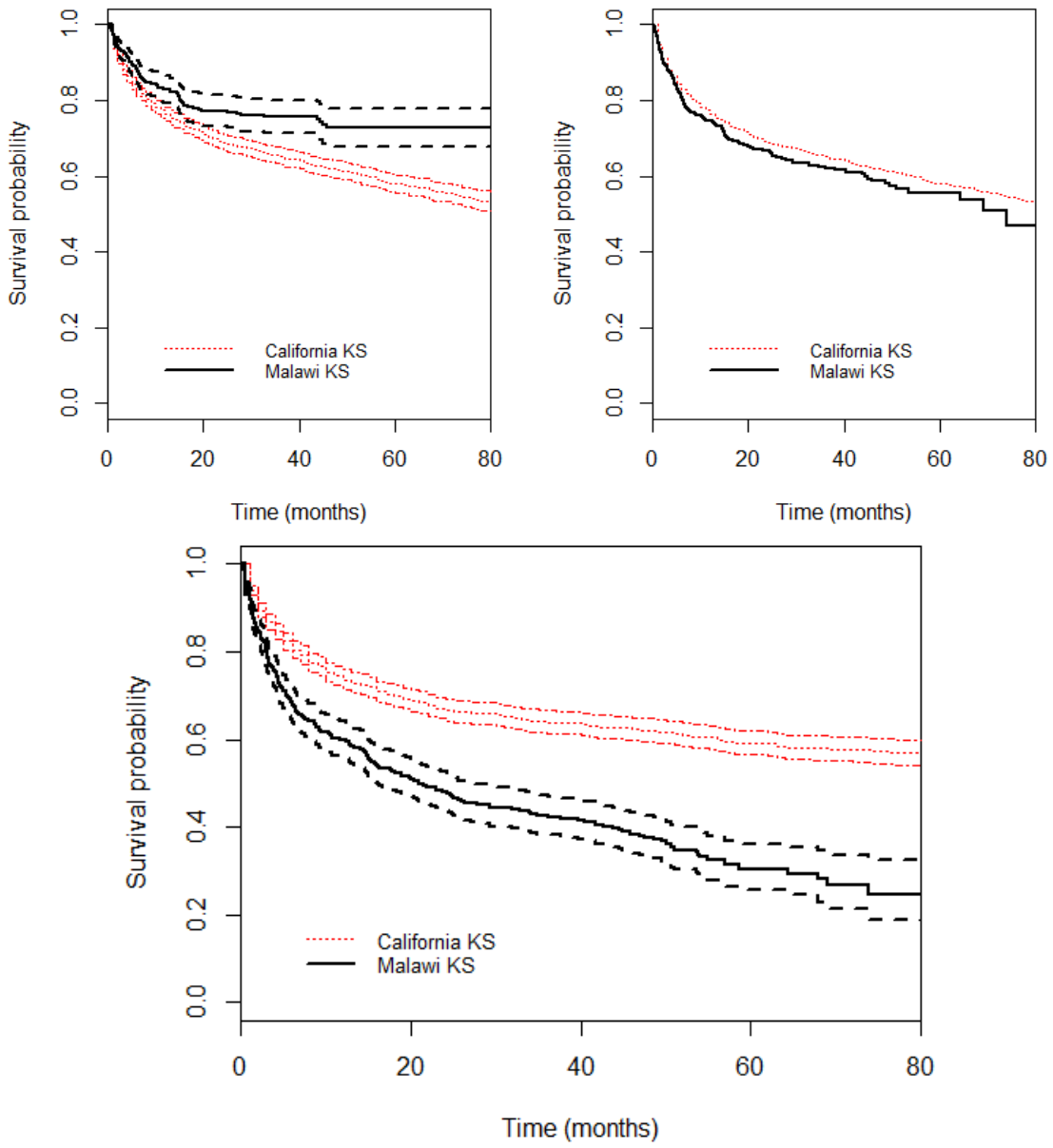

Figure 5.13: *Top Left panel*: Comparing survival plots of Malawi and US data based on the initial data. *Top Right panel*: The Malawi curve has been adjusted after accounting for 34% mortality in the loss to follow-up patients using California data as an upper bound for the survival curve. *Bottom plot*: Comparing survival plots of Malawi and US data assuming a worst case scenario (all LTFUs as deaths) in Malawi data.

Figure 5.13 (right panel) shows the "'best scenario"' K - M survival plot assuming 34% of deaths in LTFU patients. The estimated survival probability dropped to 0.47 (0.38 - 0.59) by 80 months of follow-up compared to 0.73 (0.68 - 0.78) in

the unadjusted estimates (left panel). With $c\_p = 34\%$, the median survival was estimated at 73.9 months (approximately 5 years).

5.7 Results from Semi-parametric and Parametric Models

In this section we present results from all competitive models. We also include results of correcting for LTFU using 34% as a minimum proportion of deaths among patients that were lost to follow-up. We end by presenting adjusted model results and the sensitivity analysis plots for the semi-parametric model.

5.7.1 Risk factors of loss to follow-up

After adjusting for TB status, treatment period, distance to the clinic and occupation, we found that male patients (HR: 1.39, 95%CI: 1.04 - 1.86) were significantly associated with being lost but older patients were less likely to be lost from the cohort (see Table 5.4). This analysis could have been done using a multivariate logistic model as well since in this case we are not concerned with censoring. However, we opted for a time-to-vent analysis approach so that we make the results comparable to the results from the risks of mortality from the survival models. Moreover, when we fitted the logistic model for the censored data, we noted that the direction of interpretation of estimates did not differ from the ones we obtained from the Cox models reported here.

5.7.2 Risk factors of mortality and survival

Table 5.5 provides a summary of results from different Cox models. The age was fitted as a continuous variable because the scatter plot smoothing of K-M estimates against patient age did not suggesting groups or clusters. The standard Cox model and extended Cox model with Gamma frailties gave similar estimates of the hazard ratios. All the models indicate that time of starting treatment and TB history are statistically significant factors for quantifying risk of death in these data. In particularly, patients starting HIV treatment between 2008 and 2011 had almost 50% lower risk than those starting before 2008 (HR 0.53, CI: 0.35 - 0.81).

In addition, high mortality risk was observed in patients who had TB or had TB episode in last two years prior to starting HIV, treatment. They were twice more likely to die than patients without TB (HR: 2.21, CI: 1.05 - 4.25). The 95% CI also suggest that the patients with TB could be four times higher in risk compared to their counterparts.

Baseline age, distance to the clinic, sex and occupation were not significant in the model. However, older patients tended to have increased risk, 1% higher than younger patients. Men also seemed to have a lower risk of dying than females (HR: 0.95). Furthermore, patients living near the HIV/KS Clinic tended to be 8% lower in risk of dying than those living more than 8 km away from the clinic. Lastly, the self - employed patients tended to have lower mortality risk than the employed, but students tended to have a higher risk (5.8% higher) than the employed.

The Weibull, Log-Normal, Log-logistic and generalised Gamma models were fitted

in AFT parametrization. This means that, instead of getting estimates of HR, or OR, we get time ratios (TR). Time ratios are also called accelerated factor. In this scale, we interpret the results in terms of survival. A value more than one indicates that survival rate is prolonged (accumulated), and vice-versa. The flexible PH estimates the cumulative HR, whereas flexible PO estimates cumulative OR of survival. Results from all AFT models lead to the same conclusion as Cox models. For instance, patients, starting treatment in 2008 - 2011 period had survival times twice longer than those that started in the 2004-2007 period (as shown in Table 5.6. Likewise, survival times for patients with TB in this cohort were shortened by a factor of 0.30, compared to those without TB.

Table 5.7 shows that the two models have similar estimates. The only difference is in the interpretation. For the flexible PH, the estimates are cumulative hazard ratios while for the flexible PO the estimates are interpreted as odds of survival. The results are quite similar to those reported in Table 5.5.

5.7.3 Describing patient heterogeneity

We also included two more complex models as an extension of the standard Cox PH model to model patient-level heterogeneity. The advantage of these two models is that they capture some information on unobserved individual-level heterogeneity that is otherwise not capture by the Cox model. Having this result informs the analyst the amount of variation in heterogeneity (*frailty*) among patients that is not explained by variables included in the model. Table 5.5 shows that the

extended Cox PH model with log-Gaussian frailties explains the data well. The inclusion of frailties to measure individual-level heterogeneity in the model significantly improved the model fit (partial likelihood ratio test: p-value < 0.001). The variance of the frailty value of 0.34 indicates the presence of heterogeneity not explained by the model with all covariates. We conducted a sensitivity analysis

Table 5.4: Hazard ratios for loss to follow-up using a Cox/log-Gaussian frailty model based on $c\_p = 34\%$

Variable	Adjusted HR (95%CI)	p-value
Age (years)		
15-29	Reference	
30-44	0.62(0.43 - 0.89)	**0.010**
45+	0.62(0.36 - 1.07)	NS
Sex		
Female	Reference	
Male	1.38(0.96 - 1.97)	NS
Initiation period		
2004-2007	Reference	
2008-2011	1.69(1.19 - 2.41)	**0.004**
TB status		
No TB	Reference	
With TB	1.05(0.51 - 2.16)	NS
Distance		
>8 km	Reference	
Within 8 km	0.79(0.56 - 1.10)	NS
Occupation		
Employed Reference		
Student	0.44(0.05 - 3.48)	NS
Self-Employed	1.09(0.57 - 2.07)	NS
Other	0.88(0.45 - 1.77)	NS

on the choice of knots k for both flexible models reported in Table 5.7. In a flexible PH model, we varied the knots from 3 to 5 using the `flexsurv` package in R. For instance, with $k = 5$ and $df = 15$, the HR for TB was 2.11 (1.05 - 4.24) which is similar to the result when $k = 3$ ($df = 13$) as reported in Table 5.7. All other covariate effects were similar to 2 decimal places in both PH and PO model. The

Table 5.5: Hazard ratios (95% CI) for death from Cox PH Models with all LTFUs treated as censored

Covariate	Cox	Cox/Gausian	Cox/Gamma
Age	1.01(0.976-1.03)	1.01(0.99-1.03)	1.01(0.976-1.03)
Sex			
Females	Reference		
Males	0.95(0.63-1.45)	0.96(0.62-1.47)	0.95(0.63-1.45)
ART Period			
2004-2007	Reference		
2008-2011	**0.53(0.35-0.81)**	**0.53(0.34-0.81)**	**0.53(0.35-0.81)**
TB Status			
No TB	Reference		
With TB	**2.12(1.06-4.26)**	**2.21(1.05-4.65)**	**2.12(1.06-4.26)**
Distance			
> 8 km	Reference		
≤8 km	0.92(0.62-1.37)	0.92(0.61-1.39)	0.92(0.62-1.37)
Occupation			
Employed	Reference		
Self-employed	0.88(0.43-1.80)	0.88(0.41-1.86)	0.88(0.43-1.80)
Student	1.55(0.32-7.60)	1.48(0.28-7.81)	1.55(0.32-7.60)
Other	1.01(0.47-2.14)	1.01(0.46-2.22)	1.01(0.47-2.14)
Frailty term(p-value)		0.34(0.37)	5e-05(0.75)

choice of the number of knots does not significantly change the size of estimates and direction of interpretation (Hinchliffe and Lambert, 2013).

5.7.4 Adjusting for Loss to follow-up and Results from Sensitivity Analysis

In this section, we look at how the mean HRs vary with changing datasets using the selected $c\_p = 34\%$. The value of 34% was identified as an approximate lower bound for the mortality among LTFUs but other studies on mortality in LTFUs reported mortality rates ranging from 20 - 83 % (see Table 3.2). We only report

Table 5.6: Time ratios (95% CI) for Parametric Models with all LTFUs treated as censored

Covariate	Weibull	LogN	LogLog	Gamma
Age	0.98(0.95-1.02)	0.99(0.95-1.03)	0.98	1.01(0.98-1.03)
Sex				
Females	Reference			
Males	1.11(0.54-2.26)	1.08(0.53-2.21)	1.07	1.11(0.66-1.88)
ART Period				
2004-2007	Reference			
2008-2011	**2.13(1.01-4.49)**	**2.19(1.07-4.48)**	2.31	1.58(0.91-2.76)
Distance				
> 8 kms	Reference			
≤8 kms	1.13(0.58-2.21)	1.12(0.56-2.22)	1.12	1.13(0.65-2.00)
TB History				
No TB	Reference			
With TB	**0.30**(0.09-1.01	**0.31**(0.08-1.18	3.91	0.89(0.28-2.81)
Occupation				
Employed	Reference			
Self-employed	1.02(0.35-4.13)	1.08(0.29-3.96)	1.19	0.50(0.21-1.67)
Student	0.33(0.21-5.04)	0.61(0.03-11.08	0.45	2.61(0.32-21.68)
Other	0.94(0.29-3.43)	0.90(0.23-3.57)	0.95	0.63(0.21-1.92)

results for the selected Cox/log-Gaussian model. A similar methodology can be used for parametric models.

In Table 5.8, we note that estimates of covariate effects are similar for model 2 and Model 2 in the majority of covariates. when comparing all the three models, the direction of interpretation was different in three covariates; male gender, student and "'other"' occupations.

Figure 5.14 shows mean HRs plotted against the proportion of deaths in LTFU for each covariate in the Cox/log-Gaussian model. The mean HRs are based on the 1000 times re-sampled data without replacement for each proportion of deaths (0-100%) in LTFU patients. Again, we note that ART period and TB status are significant for certain ranges of the number of deaths in LTFU patients. Also,

Table 5.7: Estimates from Flexible Parametric Proportion Hazards (Hazards Ratio) and Proportional Odds (Odds Ratio) Model with all LTFUs treated as censored

Covariate	Hazards Ratio(95% CI)	Odds Ratio(95%CI)
Age	1.01(0.98-1.03)	1.01(0.98-1.03)
Sex		
Females	Reference	
Males	0.96(0.60-1.53)	0.96(0.68-1.53)
ART Period		
2004-2007	Reference	
2008-2011	**0.53(0.35-0.80)**	**0.50(0.31-0.79)**
Distance		
> 8 kms	Reference	
≤8 kms	0.92(0.62-1.37)	0.92(0.59-1.44)
TB History		
No TB	Reference	
With TB	**2.12(1.06-4.25)**	**2.44(1.06-5.64)**
Occupation		
Employed	Reference	
Self-employed	0.88(0.43-1.81)	0.88(0.39-2.02)
Student	1.54(0.31-7.54)	1.35(0.23-8.06)
Other	1.01(0.47-2.14)	1.01(0.43-2.42)

the ranges of HRs and their 95% CIs are not very much different from the ones reported in the model results.

Table 5.8: Adjusted HR estimates from the Cox/Log-Gaussian Model after accounting for 34% of mortality among LTFU patients.

Covariate	Model 1	Model 2	Model 3
Age (years)	1.01(0.99, 1.03)	1.01(0.98, 1.03)	1.00(0.99, 1.02)
Sex			
Female Reference			
Male	1.02(0.69, 1.50)	0.97(0.63, 1.49)	0.99(0.70, 1.39)
ART Period			
2004-2007	Reference		
2008-2011	0.54(0.36, 0.82)	0.52(0.34, 0.81)	0.70(0.50, 0.99)
TB status			
No TB Reference	Reference		
With TB	1.97(0.95, 3.92)	2.19(1.04, 4.60)	1.45(0.73, 1.38)
Distance to ZCH			
>8 km	Reference		
Within 8 km	0.84(0.57, 1.25)	0.93(0.62, 1.42)	0.99(0.72, 1 .38)
Occupation			
Employed	Reference		
Student	1.08(0.22, 5.23)	1.45(0.28, 7.66)	0.92(0.18, 1.52)
Self-Employed	0.78(0.38, 1.61)	0.88(0.41, 1.87)	0.94(0.51, 1.72)
Other	0.96(0.44, 2.07)	1.02(0.46, 2.26)	1.03(0.55, 1.95)
Frailty variance		0.34(p<0.001)	0.37(p <0.001)

`Model 1 = Each covariate fitted separately in the model`

`Model 2 = Results based on the model with all covariates included and 0% deaths assumed among LTFUs`

`Model 3 = Results from a model with` $c\_p = 34\%$

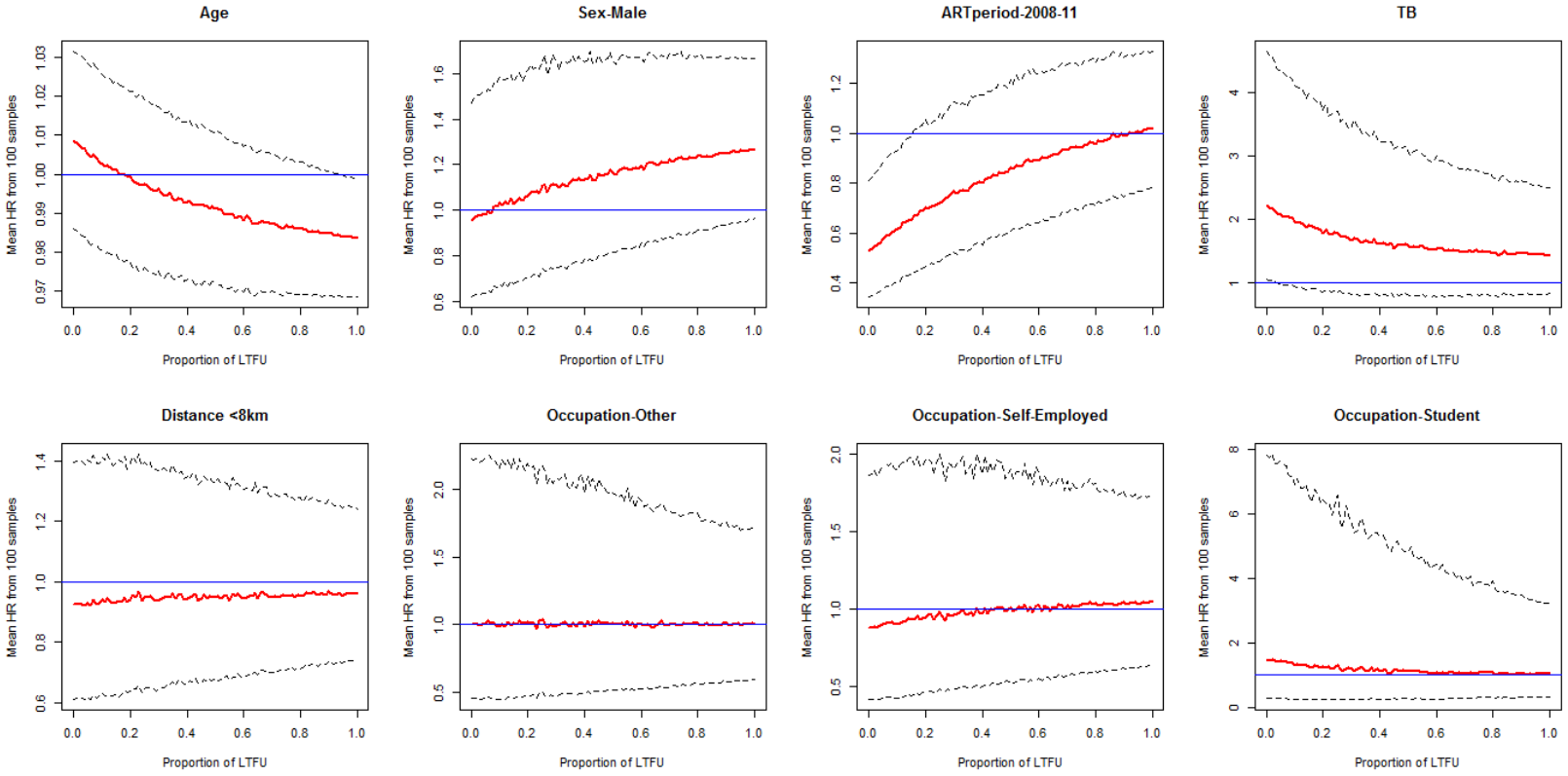

Figure 5.14: Sensitivity analysis plots for exponentiated mean β based on 1000 permutations of the data.

5.8 Model Diagnostics and Selection

It is good practice to select models that fit the data well amongst a set of competitive models. In this section, we present results from assessing PH assumptions, model selection and residual diagnostics. For comparison purposes, models were fitted on the data with no deaths assumed among LTFUs. However, the $c\_p = 34\%$ was applied to the selected Cox model (semi-parametric) and parametric model.

5.8.1 Checking the Proportional Hazards Assumption

The results from the Schoenfeld residual test for both individual covariates and the global test indicate non-violation of the PH assumption. Therefore, our data support the use of the proportional hazard assumption. Table 5.9 gives a summary of the PH test. If $\rho > 0$, then the linear trend is increasing, decreasing otherwise.

5.8.2 Comparing Cox Models

In Table 5.10, we have reported both the partial log-likelihood values and p-value from likelihood ratio test (LRT). The LRT shows a significant improvement in model fit from Cox PH model to Cox/log-Gaussian frailty model ($p < 0.001$), hence the Cox/log-Gaussian frailty model is the one that maximises the partial log-likelihood function (-PlogLik = 1155.43). This suggests the presence of unobserved individual heterogeneity among the patients.

Figure 5.15 shows three diagnostic plots based on Cox-Snell residuals against cumulative hazard estimates. As expected, the plot for Cox and Cox/Gamma are quite similar. However, residuals plot from a Cox/Log-Gaussian model indicates relatively higher residuals suggesting great variability in survival estimates. The great variability in the residuals is due to the presence of frailties which were found to be significant in Cox/log-Gaussian model.

Table 5.9: Test statistics for linear trend using the Schoenfeld residuals in a standard Cox PH model.

Covariate	rho	chisq	p
Age	0.14	2.34	0.13
Sex -Male	0.06	0.39	0.53
ART period-2008-2011	-0.09	0.77	0.38
TB -with TB	-0.01	0.01	0.92
Distance	-0.01	0.02	0.89
Occupat-Other	0.04	0.18	0.67
Occupat-Self-Employed	0.03	0.07	0.79
Occupat-Student	0.14	2.02	0.16
GLOBAL		5.29	0.73

Table 5.10: Comparing Cox models using the partial likelihood (LogLik) values.

Model	-2LogLik	partial LRT test
Standard Cox	1220.76	-
Cox/Gamma	1220.75	0.92
Cox/Log-Gaussian	**1155.43**	$\mathbf{1.30}\times 1^{-11}$

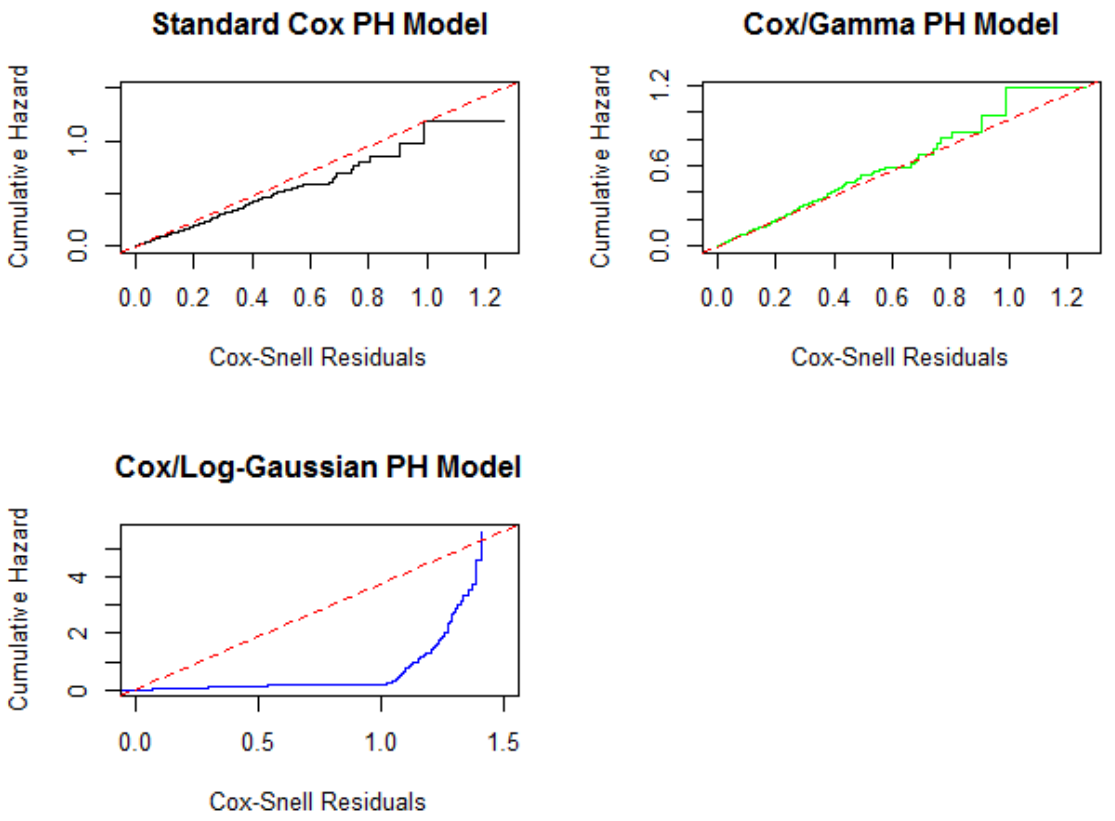

Figure 5.15: Diagnostic plots from Cox models using the Cox-Snell residuals.

5.8.3 Comparing Parametric Models

Table 5.11 presents AIC and BIC values based on Equation 4.54 and Equation 4.55. The AIC selects the Flexible Proportional Hazard (FPH) model and the BIC selects the generalised Gamma model. Figure 5.17 and Figure 5.16 show cumulative hazard function and survival function estimated from the Cox/Log-Gaussian and the parametric models. Both the FPH and the FPO follow the Cox/Gaussian function more closely than other models. This is one of the reasons flexible models have found their place in survival analysis literature and are increasingly becoming popular because of their ability to flexibility model the hazard and survival function.

Table 5.11: Selecting a best fit model using AIC and BIC values

Statistic	Weibull	LogN	LogLog	Gamma	FPH	FPO
AIC	1886.47	1868.35	1879.91	1845.99	1839.57	1840.16
BIC	1925.41	1907.23	1918.84	1889.25	1891.49	1892.07

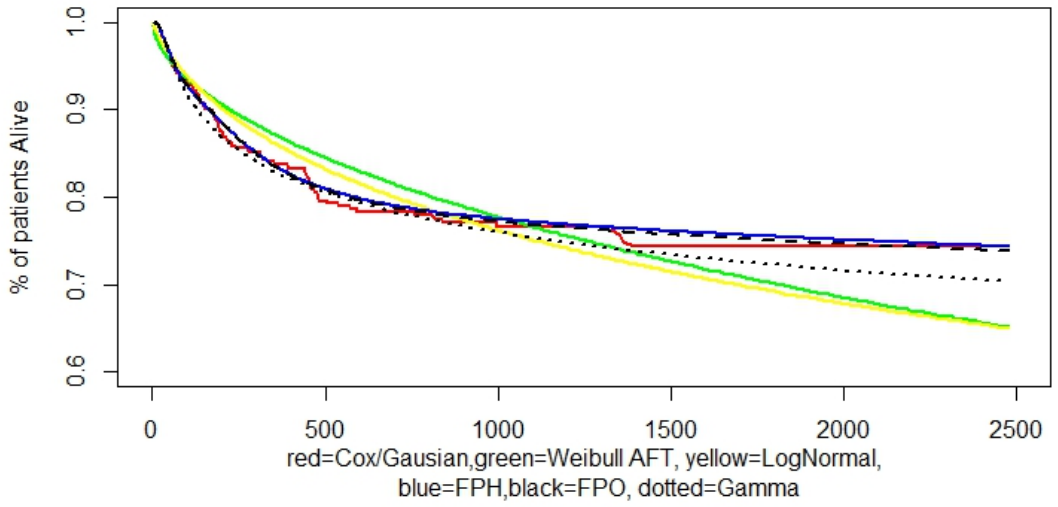

Figure 5.16: Survival functions from parametric model and a K-M plot from Cox/Log-Gaussian model.

5.9 Discussion and Conclusions

This is the first (at least to our knowledge) subgroup analysis of focusing on quantifying survival HIV patients diagnosed with KS at a time of starting HIV treatment. This is because other studies on KS patients such as Ziegler et al., (2003), Makombe, Harries and al, (2008), and Chu, Mahlangeni et al., (2010), only compared the risk in KS and no-KS patients. In addition, studies conducted by Wu et al., (2014) and Nelson et al., (2013) looked at incidences of KS in HIV population and also treatment options and outcomes available for HIV/KS patients. None of the studies extended the analysis to HIV/KS patients only. In general, the

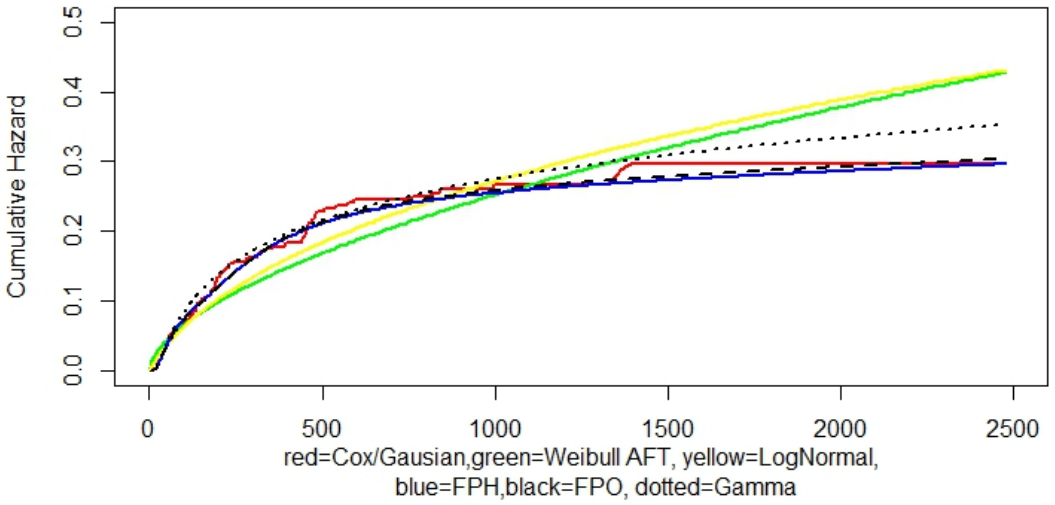

Figure 5.17: Cumulative Hazards functions from parametric models and Cox/Log-Gaussian model.

survival in this patient group was high, though lower than HIV patients without KS. There were 105 (18%) deaths, 219 (39%) patients were either lost to follow-up or transferred to another clinic and 235 (42%) were still alive by the end of the study. In a cross-sectional (retrospective) studies by Mwinjiwa et al., (2013) and Makombe, Harries and al, (2008) it was also reported that higher proportion of patients at the time of the end of the study were still in the cohort and close to 20% were either transferred or lost to follow -up. These statistics are also similar to the ones routinely reported in national HIV progress reports (Malawi Ministry of Health, 2011). These results continue to highlight the importance of establishing social services in HIV programmes to deal with this high defaulter rates.

Adjusting for other covariates in the model, the results showed that patients' age was not statistically significant in influencing the survival in this patients' group. However, the estimate of the hazard ratio suggests a slightly higher risk

of death in older patients. Other researchers such as Chu, Misinde et al., (2010) and Chu, Mahlangeni et al., (2010) fitted age as a categorical variable and none of the age groups were reported to be statistically significant, although the risk of death tended to increase across age groups. However, these results should keep reminding policymakers to be prepared for the ageing HIV population. Moreover, in the general population, older people are at more risk of dying since they tend to have a weaker immunity.

The period in which patients started treatment was found to be a significant factor influencing mortality. Adjusting for age, sex, TB history, occupation and distance, there was a significant reduction in the risk of dying for patients that started HAART between 2008 - 2011 period. (HR:0.53, CI:0.35 - 0.81). With the progress of provision of HAART, there has been increased awareness of starting HIV treatment in good time. The Ministry of Health (MOH) decreased and decentralised HIV services to rural health centres, making HIV programs easier. Furthermore, the MOH revised the HIV treatment guidelines for three times between 2004 - 2011 (Malawi Ministry of Health, 2008; Malawi Ministry of Health, 2011). These revisions are a sign of commitment by MOH to improve management of HIV patients including those with KS. No wonder, patients who started HIV treatment had a better prognosis, as they were likely to have started treatment early. Johansson, Robberstad and Norheim, (2010) reports early starting of HIV treatment associated with high survival.

Another covariate found to be significant is TB history, as shown in Table 5.7. Patients who had TB or had suffered from TB in the past two years had risk of

dying twice higher than the patients TB history. The results from the retrospective review by Makombe, Harries, Yu et al., (2007) showed high mortality in HIV-TB co-infected groups. With an addition of KS burden/patients with HIV,TB and KS are likely to be more at risk of death. Changes have been made in treatment guidelines to immediately initiate HAART for HIV patients with TB co-infection (Malawi Ministry of Health, 2011). HIV patients in these data had also access to a dedicated TB/HIV clinic within the Zomba hospital. These results support the need to continue increase TB prevention treatment and care in the general population. Normally, people with a compromised immune system have a higher risk of TB.

The results for the relationship of sex and survival showed that sex is not a significant factor of mortality. However, male patients appeared to have a greater risk of dying than females (though not significant). In the general HIV and KS population, the number of males with KS is always higher than males (Mosam, Carol and al, 2008; Mwinjiwa et al., 2013). This could be explained by the poor health-seeking behaviour amongst the general male population.

The baseline occupation status for HIV/KS patients was also not associated with mortality. However, the risk of dying was elevated in students, 48% higher than the employed, group, which included the teachers, security staff, and health care workers. The self-employed patients tended to have a reduced risk 12% lower than the employed. Our results mirror those in a case-control study by Ziegler et al., (2003) in Uganda in which they evaluated the association of having KS and occupation categorized as farmers and others . No significant differences were

noted in both case and controls. Despite the challenges, many organisations are introducing HIV workplace Policies to help their employees have early access to HAART and combat stigma and discrimination (Soko, Umar and Lakudzala, 2012; Bakuwa, 2010).

The calculation of the Euclidean distance to the Zomba hospital is one of the strengths of this study. However, it is unclear to what extent the Euclidean distance correlates with travel time and costs for travel to the hospital. Distance to the HIV/KS clinic was not associated with KS mortality in our study. We did not find other literature from the region on the relationship of distance to the clinic and KS mortality. The lack of association with survival may thus simply reflect a poor correlation between Euclidean distance and the difficulties patients experience travelling to the hospital (Rachlis et al., 2013).

The results from model selection showed that the Cox/Log-Gaussian frailty model and flexible Proportional Hazards models are adequate for explaining the risk in HIV/KS patients (refer to Table 5.5 and Table 5.7. Moreover, the Cox/Log-Gaussian model suggests the presence of unobserved heterogeneity in this patient group, looking at the value of variance for the frailty term in the model. As such, models fitted to these data ought to account for this heterogeneity. In particular, the flexible parametric proportional hazards model provide a very stable and flexible functional form for the cumulative hazards (Figure 5.17). As shown in Figure 5.17 and Table 5.7 the hazard functions and the HR estimates are similar for these two models.

From the sensitivity analysis, we can only estimate a lower bound of the proportion of deaths amongst LTFU. These results highlight the need for ascertaining the vital status (either through active tracing or other means) of all patients in survival analysis in order to obtain reliable estimates of survival and covariate effects. However, where ascertainment of vital status was not possible, our method provides a better alternative way of identifying mortality bounds amongst loss to follow-up. The majority of countries in the sub-Saharan country do not have a robust vital registration system to record vital statistics (such as deaths, births) for their citizens (Singogo et al., 2013). For countries with vital registration systems, the systems are weak and data on vital statistics are not easily accessible.

The issue of correcting estimates for loss to follow-up (LTFU) in HIV cohorts has been discussed and explored by colleagues from the International epidemiological Databases to Evaluate AIDS (IeDEA) network and elsewhere (Freeman, Semeere, Wenger et al., 2015; Kiragga et al., 2013; Henriques et al., 2012). However, the methods proposed for analysing these data are limited to HIV programmes with capacity for (some sort of) active tracing of LTFU patients which we did not have. However, these proposed methods assume high mortality as a reason for the loss to follow-up. As is in this study, where there is high survival rate in HIV cohort, it is relatively reasonable to assume low mortality as well in LTFU group as demonstrated by our comparison method and the sensitivity analyses. The results from our simulation methods were not very much different from the standard model in which vital status was not ascertained in LTFU patients. The only notable difference in the estimates was the smaller standard errors in the

model with $c\_p = 34\%$ deaths among LTFUs (not reported here).

This study has some limitations. One of the limitations is that we used routinely-collected information from clinic registers, master cards and an electronic monitoring system. This may have affected data quality to some degree, but the data generated from the national standardized monitoring and evaluation tools in the Malawi HIV programme have been found to be of good standard. This is because these data are audited every quarter by program experts from Ministry of Health headquarters. Any data quality issues are resolved on-site (Malawi Minstry of Health, 2014b). Secondly, we used data from 2004 to 2011 and the results may not reflect the current prognosis in the cohort.

In conclusion, it is important to identify potential factors that may influence patient's survival. In this study, we have seen that the subgroup analysis can be used to inform targeted patient management and treatment strategies. This can be achieved by formulating a more efficient triage system for care within a particular group of patients. For example, despite the fact that HIV/KS patients would likely be prioritized compared to non-KS HIV patients, there is still need to triage patients within the HIV/KS patients during ART initiation. This sub-triaging can also help monitor the high-risk patients over time. For instance, our results show HIV/KS patients with additional TB burden have poor prognosis compare to those without TB. In addition, patients with longer follow-up time, in particular, would need attention during care. There is also a need to strengthen tracing efforts for LTFU patients and bring them back to care. Future operational studies are needed to study geographical differences in KS epidemiology and treatment outcomes and

to describe the impacts of the ever-changing HIV treatment guidelines in line with one of the 2030 sustainable development goals of eliminating HIV/AIDS by 2030.

References II/5

Adetunji, J. and Bos, E. (2006) *Levels and Trends in Mortality in Sub-Saharan Africa: An Overview*, 2nd, The International Bank for Reconstruction and Development/The World Bank, Chapter 2.

Bakuwa, R. (2010) 'Adoption of formal HIV and AIDS workplace policies: an analysis of industry/sector variations', *Journal of SAHARA* 7, pp. 2–9.

Bruin, G.P.D. and Stefan, D.C. (2013) 'Children with Kaposi Sarcoma in Two Southern African Hospitals: Clinical Presentation, Management, and Outcome', *Journal of Tropical Medicine* 2013(2013).

Chu, K.M., Mahlangeni, G. et al. (2010) 'AIDS-associated Kaposi's sarcoma is linked to advanced disease and high mortality in a primary care HIV programme in South Africa.', *Journal of the International AIDS Society* 13, p. 23.

Chu, K., Misinde, D. et al. (2010) 'Risk factors for mortality in AIDS-associated Kaposi sarcoma in a primary care antiretroviral treatment program in Malawi.', *International health* 2 (2), pp. 99–102.

Collett, D. (2014) *Modelling Survival Data in Medical Research*, Third Edit, Boca Raton.

Freeman, E., Semeere, A., Wenger, M. et al. (2015) 'Pitfalls of practicing cancer epidemiology in resource-limited settings: the case of survival and loss to follow-up after a diagnosis of Kaposi's sarcoma in five countries across sub-Saharan Africa.', *BMC cancer* 16, p. 65.

Gbabe, O.F. et al. (2014) 'Treatment of severe or progressive Kaposi's sarcoma in HIV-infected adults.', *The Cochrane database of systematic reviews* Issue 9. Art. No.: CD003256.

Graham, S. et al. (2013) 'Loss to Follow-Up as a Competing Risk in an Observational Study of HIV-1 Incidence', *PLoS ONE* 8(3), e59480.

Henriques, J. et al. (2012) 'Comparison of Methods to Correct Survival Estimates and Survival Regression Analysis on a Large HIV African Cohort', *PLoS ONE* 7 (2), ed. by M.P. Fox, e31706.

Hinchliffe, S.R. and Lambert, P.C. (2013) 'Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions', *BMC Medical Research Methodology* 13(13).

Johansson, K.A., Robberstad, B. and Norheim, O.F. (2010) 'Further benefits by early start of HIV treatment in low income countries: Survival estimates of early versus deferred antiretroviral therapy', *AIDS Research and Therapy* 7, pp. 1–5.

Kiragga, A. et al. (2013) 'Comparison of Methods for Correction of Mortality Estimates for Loss to Follow-Up after ART Initiation: A Case of the Infectious Diseases Institute, Uganda', *PLoS ONE* 8 (12), ed. by L. Kaderali, e83524.

Makombe, S.D., Harries, A.D. and al, E. (2008) 'Outcomes of patients with Kaposi's sarcoma who start antiretroviral treatment under routine program conditions in Malawi', *Tropical Doctor* 38, pp. 5–7.

Makombe, S.D., Harries, A.D., Yu, J.K.-L. et al. (2007) 'Outcomes of tuberculosis patients who start antiretroviral therapy under routine program conditions in Malawi', *Int J Tuberc Lung Dis* 38, pp. 412–416.

Malawi Ministry of Health (2003) *Guidelines for the use of Antiretroviral Therapy in Malawi*, tech. rep., Malawi Ministry of Health, Lilongwe.

Malawi Ministry of Health (2008) *Guidelines for the use of Antiretroviral Therapy in Malawi*, tech. rep., Malawi Ministry of Health, Lilongwe http://apps.who.int/medicinedocs/documents/s18803en/s18803en.pdf; Accessed: 2013-12-04.

Malawi Ministry of Health (2011) *Guidelines for the use of Antiretroviral Therapy in Malawi*, tech. rep., Malawi Ministry of Health, Lilongwe http://www.who.int/hiv/pub/guidelines/malawi_art.pdf, Accessed: 2013-11-25.

Malawi Minstry of Health (2014b) *ART Progress Report-Quarter 3*, tech. rep., Malawi Ministry of Health, Lilongwe.

Malawi NSO (2008) *National Population Census 2008*, tech. rep., Malawi National Statistical Office(NSO), Zomba http://www.nsomalawi.mw/index.php?option=com_content&view=article&id=106&Itemid=6, Accessed: 2013-12-10.

Mosam, H.A.H.P., Carol, E. and al, E. (2008) 'Characteristics of HIV-1-associated Kaposi's sarcoma among women and men in South Africa', *International Journal of STD and AIDS* 1, pp. 400–405.

Msyamboza, K.P. et al. (2012) 'Burden of Cancer in Malawi; Common Types, incidence and Trends: National population-based cancer registry', *BMC Research Notes* 5, pp. 1–8.

Mwinjiwa, E. et al. (2013) 'Burden, characteristics, management and outcomes of HIV-infected patients with Kaposi's sarcoma in Zomba, Malawi', *Public Health Action* 3, pp. 180–185.

Nelson, B. et al. (2013) 'AIDS-associated Kaposi's sarcoma: Outcomes after initiation of antiretroviral therapy at a university-affiliated hospital in urban Zimbabwe', *International Journal of Infectious Diseases* 10, pp. 902–906.

Rachlis, B. et al. (2013) 'Using concept mapping to explore why patients become lost to follow up from an antiretroviral therapy program in the Zomba District of Malawi.', *BMC Health Serv Res* 13, p. 210.

Royston, P. and Parmar, M. (2002) 'Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects', *Statistics in Medicine* 21, pp. 2175–2197.

Singogo, E. et al. (2013) 'Village registers for vital registration in rural Malawi', *Tropical Medicine & International Health* 18, pp. 1021–24.

Soko, D., Umar, E. and Lakudzala, A. (2012) 'HIV and AIDS workplace interventions; Gaps between policy and practice at the College of Medicine', *Malawi Medical Journal* 24, pp. 52–55.

World Health Organization (2008) *Malawi Country Cooperation Strategy at glance*, tech. rep., World Health Organization, http://www.who.int/countryfocus/cooperation_strategy/ccsbrief_malawi_en.pdf?ua=1, Accessed: 2014-09-30.

Wu, X. et al. (2014) 'One hundred and five Kaposi's sarcoma patients: a clinical stud in Xinjiang, Northwest China', *Journal of European Academy of Dematology and Venereology* 28, pp. 1545–1552.

Yiannoutsos, C.T., An, M.-W. et al. (2008) 'Sampling-Based Approaches to Improve Estimation of Mortality among Patient Dropouts: Experience from a Large PEPFAR-Funded Program in Western Kenya', *PLoS ONE* 3 (12), ed. by B. Grinsztejn, e3843.

Yiannoutsos, C.T., Johnson, L.F. et al. (2012) 'Estimated mortality of adult HIV-infected patients starting treatment with combination antiretroviral therapy.', *Sexually transmitted infections* 88 Suppl 2, pp. i33–43.

Ziegler, J. et al. (2003) 'Risk factors for Kaposi's sarcoma: A case-control study of HIV-seronegative people in Uganda', *International Journal of Cancer* 103 (2), pp. 233–240.

Chapter 6

Estimating spatial variation in survival in HIV patients.

Summary

In this analysis, we examine geographical variations in survival in a sample of 1,931 HIV patients enrolled in ART program in Zomba district. Our motivation for this analysis comes from the fact that in resource-limited settings such as our study setting, there are enormous challenges in providing and accessing essential health services. With only a few specialised HIV clinics available, most patients are forced to travel long distances to access HIV treatment. Therefore, we are interested in studying how the risk of death varies in space. We use spatial survival models which are just extensions of the proportional hazards model to include the spatial frailties. Our model selection criteria indicate that a Weibull spatial survival model

with an exponential spatial covariance structure adequately describe the variation in survival in our data. More importantly, we produce risk maps highlight the areas with lower and high risk. We note that there is noticeable local geographical variation in terms of hazards of death, with the risk of death being elevated in the northern and southern parts of the district. Highlighting the areas that are likely to have poor survival outcomes and mapping the risk of mortality can help policymakers identifying priority areas during resource allocation process.

6.1 Introduction

Despite recent progress in reducing HIV/AIDS morbidity and mortality as a result of expansions of HIV-related activities in Africa, over 70% global AIDS-related deaths occur in Africa especially sub-Saharan Africa (World Health Organisation, 2016). Reports from Africa have also highlighted the in-country required variability in the HIV morbidity and mortality.

To study geographical variations of a particular disease, spatial methods are increasingly being applied. The spatial analysis methods have a long history in studying disease epidemiology and geographical variations. The spatial epidemiology has been extensively used in the modelling and prediction of the incidences of diseases such as HIV/AIDS, malaria, cancer, and Schistosomiasis, especially in sub-Saharan Africa region. For example, in Malawi, spatial analysis tools have been used in the modelling prevalences and incidences of pneumonia, malnutrition, malaria, and perinatal mortality (Kazembe and Kamndaya, 2016; Banda et al., 2016; Cuadros and Abu-Raddad, 2014; Kazembe, Muula et al., 2007; Kazembe, Kleinschmidt et al., 2006). All these studies highlighted the vital roles which the spatial epidemiology plays in increasing our understanding of the diseases as well as in the planning of resources to tackle these diseases.

Although considerable spatial epidemiology studies have been largely devoted to study HIV incidence and prevalence, rather less attention has been paid to study spatial variations in survival in HIV populations. Considering that HIV patients have compromised immunity, their survival is severely reduced especially if not on

ART treatment. Hence, studying geographical patterns in survival would help to plan and allocate resources to areas with likely to have poor treatment outcomes. Though HIV prevalence is high in urban areas, but the survival in this area is expected to be better compared to the rural areas because the urban areas usually have better access to health and social services. Therefore, we apply the spatial survival analysis methods to examine geographical variations in survival among ART patients in a rural district in Malawi, Africa.

6.2 Objectives

In this study, we address the third objective as listed in section 2.2; *Adjusting for individual-level covariate data, investigate the geographic variation in survival prognoses in the Zomba district.* In particular, are geographical determinants part of the unexplained observed variation in the hazards of death?

Our overall anticipation is that this study will add to our understanding of the epidemiology of Kaposi's sarcoma among HIV patients in Zomba, Malawi and whether the survival outcomes are characterized by the geographical distribution.

6.3 Data description

The data used in this study is from 1,931 HIV patients who started routine HIV treatment between 2004 and 2011 in Zomba district, Malawi. We randomly selected 1,500 patients (controls) from a cohort of 14,300 adult HIV patients from

Zomba HIV clinic. This cohort was selected using database IDs and efforts were made to have complete data for these patients. As such, a filed visit to Malawi was organised during summer in 2015 and recruited two research assistants to help in this exercise. These data were added to 559 HIV patients with KS (cases) at baseline. Although the total number of patients was 2059, 128 patients did not have complete data on sex, age and date of starting ART, all of which were required for in this analysis. Therefore, the final dataset that was used had 1931 patients. In both datasets, baseline data were used in the analysis.

During the period under consideration, all Kaposi's sarcoma (KS) patients in the district were centrally clinically managed at Zomba Central Hospital (ZCH), one of the two referral tertiary hospitals in the southern region of Malawi. Other HIV patients without KS were clinically managed either at this central hospital or at decentralized HIV clinics in the district. More complicated cases from decentralized clinics were referred to Zomba HIV clinic.

The following patient covariate information was available: age, date of registration, sex, tuberculosis history (TB within past two years), KS status and follow-up time. To get the geographical coordinates, we matched the names of the villages where patients came from to the names of landmark locations identified by the National Statistics office (NSO), Malawi. These locations included residences for traditional leaders, markets, hospitals, military camps, schools, churches, and mosques. We used the coordinates assigned to these landmark locations as a proxy for patients' point locations.

6.4 Methods

The model fitted in this chapter is of the form as in Equation (4.32) defined on page 61. We use the model that has the baseline hazard function from Weibull distribution. Other complex model exists but are more computation intensive and the Weibull distribution also offers a flexible distribution (increase or decrease) and is popularly used in medical studies.

Our aim is to model possible residual spatial variation in survival after accounting for patient-specific prognostic factors and unobserved individual frailties. For convenience sake, we repeat this model here;

$$h(t;\xi,x) = h_0(t)\exp\left\{X\beta + S(x)\right\}, \quad x \in \mathbb{R}^2 \tag{6.1}$$

where $\xi = (\beta, \alpha, \lambda, \tau^2, \sigma^2, \phi)$ is a vector of the model parameters, α and λ are shape and scale parameters for the baseline hazard function from the Weibull distribution, β and X as defined above, and $S(x)$ is a value of spatial continuous stationary Gaussian latent field measured at location of survival time i with σ^2 and ϕ being the marginal variance of the Gaussian field and spatial decay parameter the covariance functions. In this study, we consider the Exponential covariance function and Matern function as defined in Section 4.3.2.1 on page 62 and a baseline hazard from Weibull distribution. The choice of this baseline hazard function was largely based on the limited availability of other options in the R package `spatsurv` that we used for spatial survival data. At the time of this time, only Exponential and Weibull were available options for the baseline hazard but we opted for the Weibull

distribution because it is more flexible compared to the Exponential function.

Fixing the baseline hazard function (Weibull distribution), we fitted eight (8) variants of Equation (6.1) and selected the model that fits the data well. Here is a list of all the models considered;

Model 1: *Exponential model with spatial frailties only.* The parameters estimated in this model are; $\beta, \alpha, \lambda, \phi, \sigma^2$. Using an exponential covariance function, this model estimates spatial variance (σ^2) in survival estimates. It also estimates covariate effects (β), the spatial decay parameter and parameters from the baseline hazard function (α, λ). This model is same as Matern with $k = 0.5$.

Model 2: *Exponential model with both Individual-level and spatial frailties.* The parameters estimated in this model are; $\beta, \alpha, \lambda, \tau^2, \sigma^2, \phi$. In addition to parameter estimates in Model 1, this model provides one additional estimates: the variance τ^2 which describes the heterogeneity in patients.

Model 3: *Matern ($\kappa = 1$) model with spatial frailties only.* The parameters estimated in this model are; $\beta, \alpha, \lambda, \phi, \sigma^2$. This model is similar to Model 1, with the only difference being the covariance function.

Model 4: *Matern ($\kappa = 1$) model with Individual-level and spatial frailties.* The parameters estimated in this model are; $\beta, \alpha, \lambda, \tau^2, \sigma^2, \phi$. This model is similar to Model 2, with the only difference being the covariance function.

Model 5: *Matern ($\kappa = 1.5$) model with spatial frailties only.* The parameters estimated in this model are; $\beta, \alpha, \lambda, \phi, \sigma^2$. This model is similar to Model 3, with the only difference being an increase in smoothing parameter $k = 1.5$ specified in the covariance function.

Model 6: *Matern ($\kappa = 1.5$) model with Individual-level and spatial frailties.* The parameters estimated in this model are; $\beta, \alpha, \lambda, \tau^2, \sigma^2, \phi$. This model is similar to Model 4, with the only difference being an increase in smoothing parameter $k = 1.5$ specified in the covariance function.

Model 7: *Matern ($\kappa = 2$) model with spatial frailties only.* The parameters estimated in this model are; $\beta, \alpha, \lambda, \phi, \sigma^2$. This model is similar to Model 3 and Model 5, with the only difference being an increase in smoothing parameter $k = 2$ specified in the covariance function.

Model 8: *Matern ($\kappa = 2$) model with Individual-level and spatial frailties.* The parameters estimated in this model are; $\beta, \alpha, \lambda, \tau^2, \sigma^2, \phi$. This model is similar to Model 4 and Model 6, with the only difference being an increase in smoothing parameter $k = 2$ specified in the covariance function.

The resulting integrals from the likelihood functions of these models are very complex, necessitating the use of Bayesian approaches. To complete the Bayesian framework, we specify the independent Gaussian (IG) priors for the parameters to be estimated in the model. Since the parameters in the baseline function and those in the correlation function are always positive, the specified priors are given

on the natural logarithm scale. The priors for the parameters are specified as follows;

- $\beta_i \sim N(\mu = 0, sd = 10)$
- $\log(\alpha_i) \sim N(\mu = 0, sd = 10)$
- $\log(\lambda_i) \sim N(\mu = 0, sd = 10)$
- $\log(\sigma_i) \sim N(\mu = 0, sd = 0.5)$
- $\log(\tau_i) \sim N(\mu = 0, sd = 0.5)$
- $\log(\phi_i) \sim N(\mu = 8, sd = 0.25)$.

The prior for the spatial decay parameter was not vague in order to reduce the computation time. When we used vague priors for this parameter, it took long twice as much time as when it was not vague. Specifying a prior close to the true value of ϕ (used variogram) significantly decreased computation time by almost half when using high end computing (HEC) and by four fold when using a 64 bit/4GB RAM laptop. On HEC, each model took an average of 2 days to complete the run and about 6 days on a laptop (using non-vague parameters).

Ideally in a Bayesian analysis we would select model priors to be diffuse, and we have attempted to do this here, with the exception of the parameter ϕ, which is typically poorly identified by the data in spatial applications, our prior for $\log(\phi)$ lends support to spatial dependence between values of ϕ 1 and 7 kilometres. For diffuse priors; $\log(\sigma_i)$ and $\log(\tau_i)$ has a support between 0 and 7; for β_i, $\log(\alpha_i)$ and $\log(\lambda_i)$ the support is between 0 and 8e+15.

6.4.1 Model Diagnostics and Comparison

All the analyses were done in `R` software using `spatsurv package`(Taylor, Davies et al., 2015). The autocorrelation plots and log-posterior plots were used to check evidence of reasonable mixing of chains and convergence of chains respectively. These diagnostics plots were done using `mcmcplots package` in `R`.

The DIC is frequently reported in Bayesian models. However, the review of different criteria of comparing competitive Bayesian models by Gelman, Hwang and Vehtari, (2014) recommended the use of WAIC in selecting a best fit model. The WAIC is preferred to DIC because it is based on the posterior density unlike the DIC which is based on the posterior point estimate such as mean or median.

6.5 Descriptive Analysis

A total of 1931 HIV patients comprising of 560 patients diagnosed with Kaposi's sarcoma during treatment initiation and 1371 HIV patients without KS. There were a total of 238 deaths recorded, 105 among KS patients and 133 in HIV patients without KS.

Figure 6.1 shows parts of the district with low and higher number of patients who attended the HIV clinic. We created quadrant over the map and counted the number of patients falling inside each quadrant. This gives a rough estimate of how patients are distributed in the district. In our data, we observe that many patients attending the clinic tend to be those living in the centre of the town near

where the clinic is. This could be explained by the high population density around this area as the clinic is inside Zomba city.

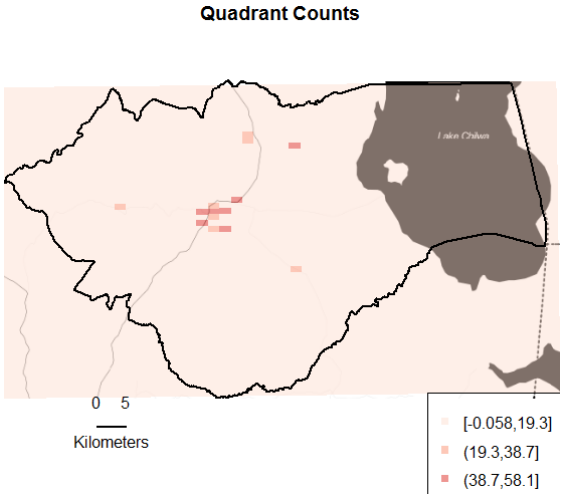

Figure 6.1: Quadrant counts showing areas with high number of HIV patients in Zomba. The counts were calculated by counting the number of patients in each quadrant created over the map.

However, the lake on the right side of the district has so many small islands and most can not be seen with the current map resolution. As such, subsequent (probability) maps in this chapter must be interpreted with caution especially for areas seen to be on the lake. Figure 6.2 shows a google maps snapshot of the islands on the northern part of the lake.

6.5.1 Mapping deaths, active follow-up and lost to follow-up

In total, there were 238 deaths and 605 patients in active follow-up at the end of the study. About 64% of patients with censored times were patients that were categorised as lost to follow-up. One possible reason for this high attrition rate could be due to poor documentation especially in cases where patients have been transferred to other HIV clinics.

Figure 6.2: Map of Lake Chilwa showing different small islands, with the majority of larger islands in the northern end. Also, some tiny islands also exist in the southern end of the lake

One of the interesting descriptive result would be to visualise how the patients that experienced the event (death) and LTFUs are distributed on the map. To do this, we used the smoothed Kernel intensity estimates to compare the patient counts for those who died versus those who were lost to follow-up, and the deaths versus those who were still in active clinic follow-up at the end of the study. We are interested in the exceedance probabilities $Prob\{\text{intensity}_{death} - \text{intensity}_a > 0\}$ where $a \in (active\ follow\text{-}up{:}alive,\ LTFU)$. We plotted these probabilities to identify areas with high probabilities of finding more deaths.

The plot on the left panel in Figure 6.3 (Death vs Alive) is an exceedance probability map showing how events (deaths) and active follow-up patients are distributed in Zomba. We plotted the difference in their the kernel estimate of point process intensities (deaths - alive), so that the colour corresponding to high probability values on the scale bar indicate the areas likely to have more counts of deaths (more yellow-red), and low probability values indicating areas likely have similar

intensity estimates for the two events being compared (more blue). From this map, it can be seen that more deaths were likely to be observed around Lake Chirwa (refer to Figure 2.1), the eastern direction of the map. Similarly on the right panel of Figure 6.3 (Death vs LTFU), the exceedance probability map is for the differences between the process intensity estimates for death events and loss to follow-up. Compared to observed LTFUs, we also note that in addition to the eastern region of the district, more deaths were also likely to be observed on the south- west and northern part of the district. These latter two areas coincide with the locations of two of the earliest HIV clinics in the Zomba district. On both maps, we see some areas in the lake with high probability of having more counts of deaths. This is because there some islands in the middle of the lake. Since we used smoothed mapping, even other points on the lake had an estimate but had zero probability of having death counts. This is one of the limitations for these maps because we did not have polygons for these islands so that they can be isolated.

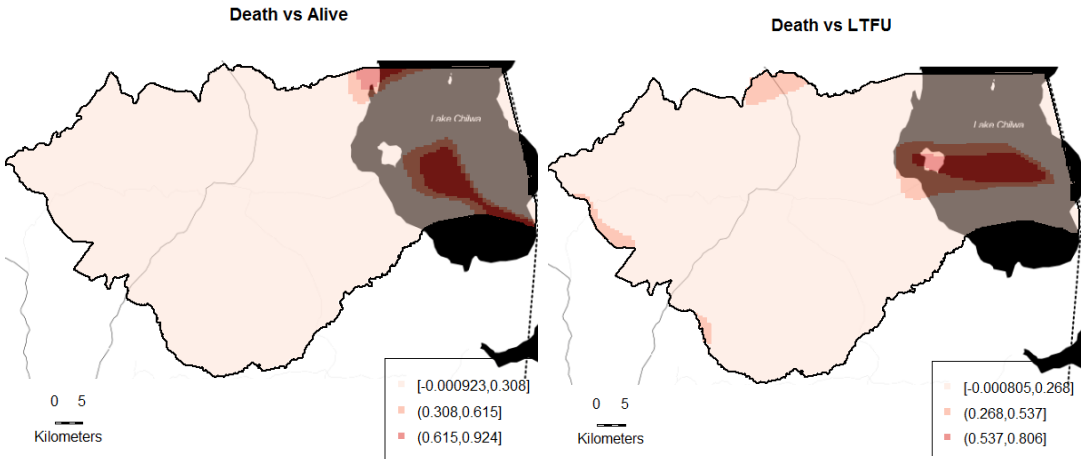

Figure 6.3: *Left panel*: The Zomba map showing areas with high probability of having more number of deaths compared to those in active clinic follow-up: using the difference in the kernel estimate of point process intensities for death events and alive patients. The colour corresponding to high probability values on the scale bar show areas with larger intensity estimates for the death. *Right panel*: The map on the right density differences of death events vs lost to follow-up (LTFU). Likewise, the colour corresponding to high probability values on the scale bar correspond to the areas with more deaths events compared to the LTFU.

The issue of LTFU is considered for reference when fitting the models in Section 6.6.3 and the maps for these are included in Figure 6.9. We used two proportions of deaths among LTFU: 34% and 60%.

6.5.2 Distribution of patients by ART period

In the study in Chapter 5, we found that the date of starting ART was significantly associated with the risk of death as well as being lost to follow-up, with patients who started in the later period having significantly lower hazards of deaths. We would like to explore if the distribution of patients between the two ART periods differ.

A total of 644 (33.4%) started HIV treatment (ART) between 2004 and 2007, and 1287 between 2008 and 2011. We used the kernel estimation to estimate point process intensities in order to compare the distribution of patients in these two ART periods. We also used exceedance probabilities as described above, but the difference in intensities was between estimates for patients who started ART in 2004-2007 period and those in 2008-2011 period. We were interested in areas of the district with this difference > 0, to indicate more patients who started in 2004-2007 observed.

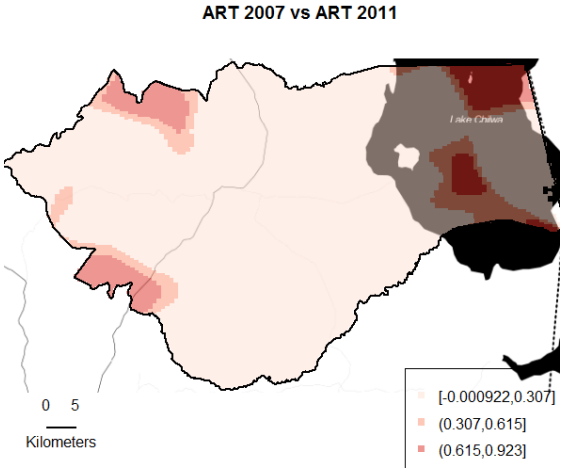

Figure 6.4: Map showing exceedance probabilities, with colours corresponding to higher values of probabilities showing areas that are likely to have more patients who started ART in the 2004-2007 period, otherwise no difference (more blue).

Figure 6.4 is an exceedance probability map and the colour corresponding to higher values of probabilities show areas likely to have more observed patients who started ART in the 2004-2007 period. The parts of the district with more blue colour (i.e. lower probabilities) indicates the point process intensities for the two periods are likely to be similar. From this map, it is suggested that the western, south-west, north-west and north-eastern parts (more red colour) of the map were likely

to have earlier ART initiations (2004-2007) compared to initiations in the later period (2008-2011). Again, some areas in islands on the lake did also have higher probability having earlier initiations.

6.5.3 Mapping TB Episodes and KS Diagnosis

In the analysis in the Chapter 5, we also found that TB status was an important predictor or mortality. Therefore, we wanted to identify areas that observed more TB episodes. In addition, we also plotted the KS cases on the map.

A total of 301 (15.6% of 1931 patients) were either diagnosed with TB or were on TB treatment at the recruitment time. The plot on the left panel of Figure 6.5 is an exceedance probability map obtained by calculating the probabilities that the difference between intensity estimates of having TB and no TB (TB vs no TB). Similarly, the colour corresponding to the greater probability values on the scale bar represent areas on the map likely to have more TB episodes observed compared to TB-free episodes. It can be seen that there were more TB episodes on the eastern and north-eastern parts of the district.

Among the 1931 patients, a total of 560 were diagnosed with Kaposi's sarcoma during routine initiation of ART in Zomba during the study period. We also used the differences in the intensity estimates to check areas with more KS cases compared to non-KS cases. Similar to other maps, the KS cases seemed to be likely to be observed in the western, north-western and eastern parts of the district as shown in the right panel of Figure 6.5.

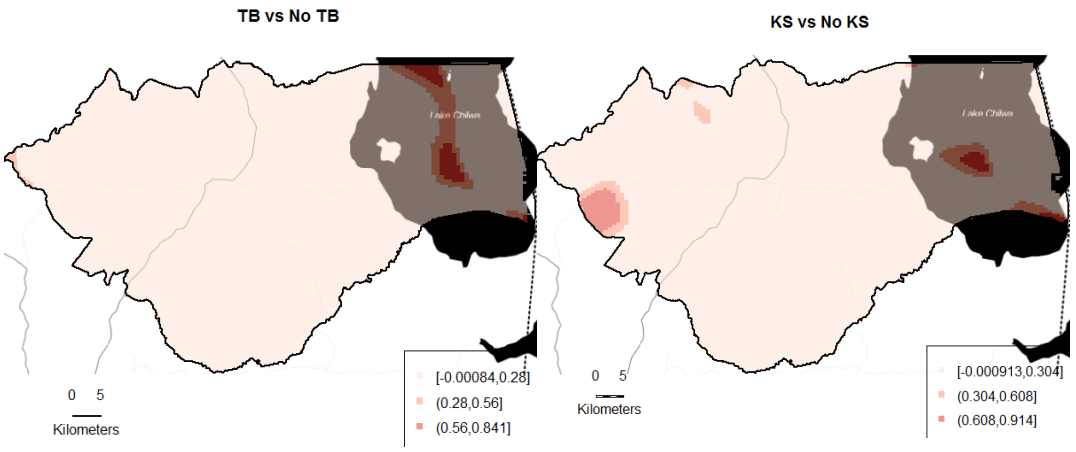

Figure 6.5: *Left panel*: The map of Zomba showing exceedance probabilities of the difference in intensity estimates between patients with TB and those without TB episodes (TB - no TB) being greater than zero. The greater probability values on the scale bar correspond to the areas with more TB episodes (more red colour). *Right panel*: An exceedance probability map showing areas where more KS cases were recorded: more red colour corresponding to areas likely to have more KS cases.

6.6 Spatial Survival Model Results

This section contains results from different competitive models considered in this analysis. We start by presenting model comparison results based on the WAIC and DIC values. Then we present the results from the MCMC diagnostics for the chosen model. Lastly, we end the current section by presenting the exceedance probability maps in order to provide an intuitive interpretation of the areas with high hazard rates by using example cut-off points.

6.6.1 Model Comparison

We compared eight (8) competitive spatial survival models, all with a Weibull baseline hazard function. We chose the best fitting model to the data using the

WAIC value as presented in the Table 6.1 below. The final model was chosen using WAIC values. We preferred WAIC to DIC because the WAIC uses the whole posterior density while the DIC uses a point estimate of the posterior distribution. The limitation of using the point estimates is that sometimes they do not provide a good summary of the distribution. In our case, the newly introduced WAIC addresses this limitation.

Table 6.1: Comparing Weibull spatial survival models with different covariance functions.

	Model	P_D	DIC	WAIC
Model 1	Exponential: Spatial*	-25.22	4500.82	**4564.38**
Model 2	Exponential: Individual-level + Spatial†	-8	4512.11	4580.78
Model 3	Matern ($\kappa = 1$): Spatial	-26.55	4498.59	4564.66
Model 4	Matern ($\kappa = 1$): Individual-level + Spatial	-10.94	4516.58	4583.24
Model 5	Matern ($\kappa = 1.5$): Spatial	-19.82	4511.71	4583.58
Model 6	Matern ($\kappa = 1.5$): Individual-level + spatial	-8.97	4525.15	4583.53
Model 7	Matern ($\kappa = 2$): Spatial	-25.89	4518.98	4584.41
Model 8	Matern ($\kappa = 2$): Individual-level + spatial	-8.03	4521.17	4582.63

*`Spatial= Spatial frailties`

†`Individual-level + Spatial = Individual-level + Spatial frailties`

After comparing the WAIC values, a Weibull spatial model with an exponential correlation function was found to fit these data well. However, a Weibull spatial model with a Matern $\kappa = 1$ had a similar WAIC value to the exponential model (4564.66 vs 4564.38). To our understanding, this difference in the WAIC values

is very minimal such that these two models should give comparable parameter estimates. Lastly, we note that in all the models, the inclusion of the Individual-level frailties was not beneficial to the overall model fitness using both DIC and WAIC values.

6.6.2 MCMC Diagnostics

We give details of the MCMC diagnostics for the Weibull Spatial frailty model with the exponential covariance function. We run 2,000,000 iterations with burn-in of 100,000 and thinning every 1000^{th} sample. The computation time varied on personal computers and high end computing (HEC). On a desktop it took an average of five days to complete a model while on HEC it took an average of two days for the same models.

We conducted diagnostic checks for reasonable mixing and convergence of the chains. In Figure 6.6, a lag-1 autocorrelation plot (first plot from left) of posterior MCMC samples shows independence of in the samples drawn (autocorrelation clustered around zero). The log-posterior density plot in Figure 6.6 (second plot from left) indicates the chain has found a mode and remained around it for the duration of the run. This is evidence that the chain has left the transitory stage.

The last two plots in Figure 6.6 are the density plots for the prior and posterior distributions for the parameters of the correlation function. When we fit a Bayesian model, we input a prior density for our parameters of interest and the data modify the prior through the likelihood to arrive at the posterior. We can therefore

compare plots of the prior and posterior to get an idea of the information content in the data. When the data are well able to identify a parameter in our model, we expect the prior and the posterior to look very different, however if the prior and posterior look similar, then the data do not provide much information on the parameter. These plots indicate that there is some information added from the data to the prior for the spatial variance parameter (σ) but not in the case of the distance parameter, ϕ, a common phenomenon in spatial analyses. For the spatial variance parameter σ, there is a small shift to the right in the posterior density.

All trace plots for fixed and random effects parameter indicate reasonable mixing in the chains as shown in Figure 6.7 and also shown in Figure 6.8. The MCMC chains for the parameters of the baseline hazard function (Weibull, α, λ) and spatial correlation function (ϕ, σ^2) are shown in Figure 6.7 while chains for fixed effects are shown in Figure 6.8.

The diagnostic plots for the other models considered in this analysis looked similar.

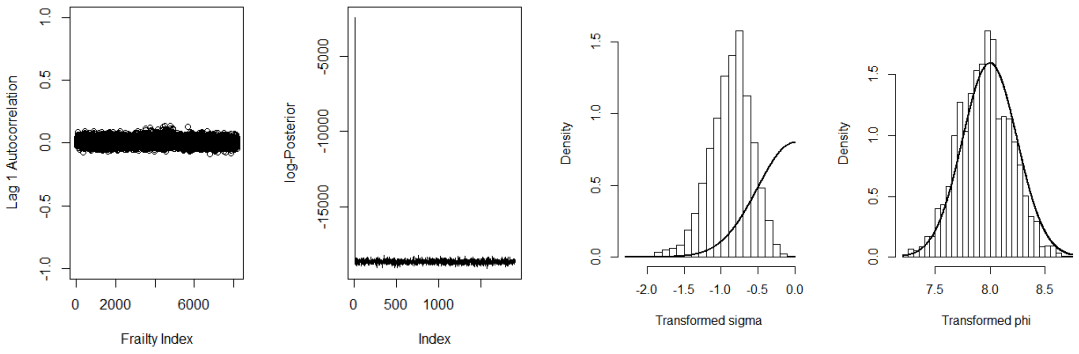

Figure 6.6: The first two plots on the left are plots for convergence,indicating close-to-zero correlation between samples. The last two plots are overlay plots of posterior density over prior density show how information from data improve the prior knowledge

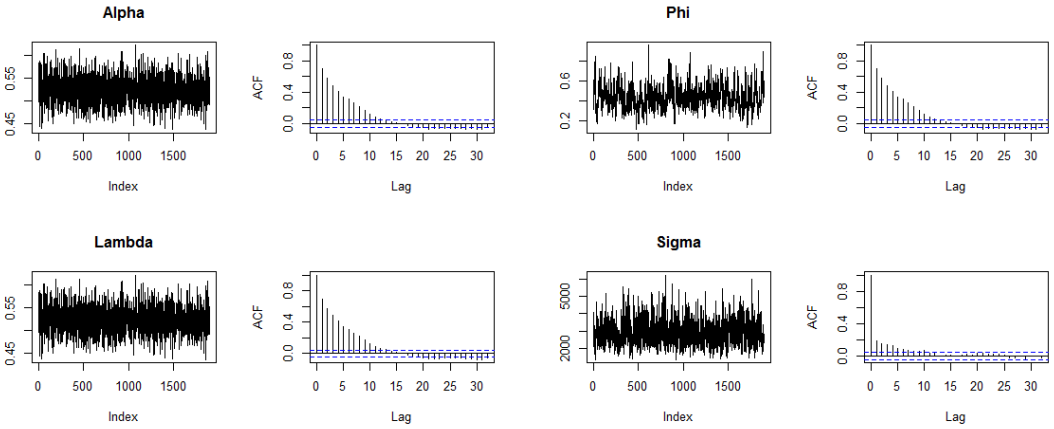

Figure 6.7: Trace and autocorrelation plots for the chains of parameters for the baseline hazard and for the spatial correlation function.

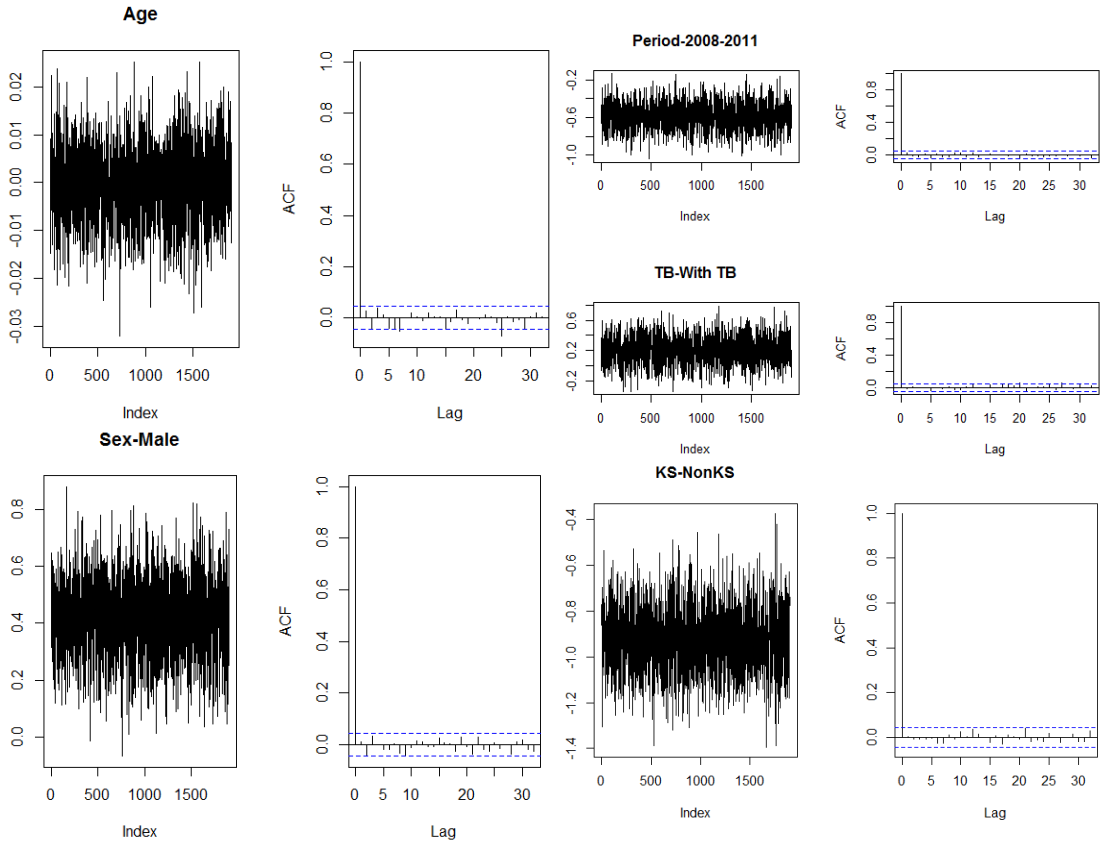

Figure 6.8: Trace and autocorrelation plots of the chain for the fixed effects βs

6.6.3 Estimates of model parameters

Table 6.2 shows summary posterior estimates of the fixed and random effects from the Weibull spatial survival model with an exponential correlation function. From this table, only sex, ART period and (not) diagnosed of KS were significantly associated with risk of death. Male HIV patients were 50% more at greater risk of dying than female patients (HR: 1.509, 95% HPD: 1.149 - 1.961). Low mortality risk was observed in patients who started HIV treatment from 2008 onwards (HR: 0.558, 95% HPD: 0.421 - 0.740). Low mortality was also observed in HIV patients who were not diagnosed with KS, with approximately 60% reduction in risk (HR: 0.398, 95%HPD: 0.306 - 0.527). No statistically significant associations were observed among patients with different ages and whether patients had TB or not.

The estimate of the spatial variance indicate evidence of unobserved spatial heterogeneity in the survival in the district (see Figure 6.9).

Table 6.2: Posterior summaries of fixed and random Effects from Weibull spatial model.

Covariate		50%	2.5%	97.5%
Age	β_1	1.000	0.984	1.015
Sex	Ref:Females			
Males	β_2	1.509	**1.149**	**1.961**
ART Period	Ref: 2004 - 2007			
2008-2011	β_3	**0.558**	**0.421**	**0.740**
TB Diagnosis	Ref: No TB			
With TB	β_4	1.218	0.851	1.737
KS status	Ref: With KS			
No KS	β_5	**0.398**	**0.306**	**0.527**
Spatial variance	σ^2	0.428	0.226	0.689
Correlatio scale	ϕ	2858.184	1784.839	4517.591
Weibull shape	α	0.52310	0.46623	0.5836

6.6.4 Exceedance Probability Maps

Probability maps are maps that are generated by plotting the probability of a certain phenomenon happening. In our case, we are interested in generating maps that can be used to show areas with high probability of having low survival estimates (or high hazards for deaths). For this purpose, exceedance probabilities are commonly used. An exceedance probability is defined as probability that the hazard rates (as in our case) exceed a certain threshold value. The choice of a threshold value is largely based on the expert advice or clinical relevance of the cut-off value. Arguably, the use of the probability maps in spatial epidemiology provides an intuitive interpretation of how the risk is spatially distributed. Secondly, when calculating the exceedance probabilities, the mean prediction and standard errors are taken into account, thereby simplifying the burden of comparing two different maps: one for mean prediction and another for standard error.

The quantity $\exp(S)$ which is the hazard rates (HR) is a post-estimate of Equation (6.1). However, we plotted the exceedance probabilities $\mathbb{P}[\exp(S) > c]$ for a particular spatial location, with c taking the following values; 1.3 and 1.5. In our case, the choice of reference value was arbitrary but clinically interpretable. We also tried other cut-off values such as 2 (twice the hazard rates) and 2.5 but the maps for these values were very faint.

To account for the issue of lost to follow-up (LTFU) in cohort studies, fitted three different variants of the final spatial model: i) the model assuming 0% mortality rate among LTFU patients, ii) a model assuming 34% mortality among LTFU

patients, and iii) a model assuming 60% mortality among LTFU patients. Due to the complexity of spatial models and the associated high computation cost, we did not perform sensitivity analysis for the assumed choices of 34% and 60% mortality rates. As such, we did not report covariate effects from the spatial model for these models but instead produced exceedance probability maps based on these model results.

We plotted three different maps, each having a different representation of proportion of mortality among the LTFU patients. In Figure 6.9, the maps in the first row were produced with all LTFU patients censored (0% deaths among LTFU patients). The middle row of Figure 6.9 were produced after assuming 60% mortality among LTFU patients. The last row was produced using the minimum value 34% mortality among LTFU, the value which was identified in the comparison method for identifying an acceptable lower bound of death rate among LTFU patients as reported in Chapter 5 under Section 5.6. Unlike in Chapter 5, we did not conduct sensitivity analysis of spatial model because of the high computation cost involved in fitting these models.

In Figure 6.9, the first two rows of the plots show similar areas where probability is likely to exceed 1.3 and 1.5 (with a lake on the eastern side). These plots indicate the northern and southern parts of the district having a higher probability of the hazard rates exceeding 1.3 and 1.5. However, the choropleth maps on the bottom row highlight different areas to the other plots on the first and second rows. The plots on the bottom row indicate that the areas around the city centre of the district are the ones that likely to have the highest probabilities of exceeding the

reference HR values. A possible explanation for this difference could be due to sampling bias in deaths among LTFU since this area is densely populated.

Figure 6.9: Maps showing probability of hazard ratios of death using two cut-off values; 1.3 and 1.5. *Top row:* $P_0[\exp(Y) > 1.3]$ (left) and $P_0[\exp(Y) > 1.5]$ (right) indicate the areas on the map with the probability of HR exceeding the two cut-off values when mortality amongst the LTFU patients is ignored. *Middle row:* $P_{34}[\exp(Y) > 1.3]$ (left) and $P_{34}[\exp(Y) > 1.5]$ (right) indicate the areas on the map with the probability of HR exceeding the two cut-off values when the proportion of deaths amongst the LTFU patients is assumed to be 34%. *Bottom row:* $P_{60}[\exp(Y) > 1.3]$ (left) and $P_{60}[\exp(Y) > 1.5]$ (right) indicate the areas on the map with the probability of HR exceeding the two cut-off values when the proportion of deaths amongst the LTFU patients is assumed to be 60%.

6.7 Discussion

Overall, we noted that the majority of patients on the HIV treatment were living around Zomba Central Hospital (see Figure 6.1). When comparing the distribution of patients by ART outcomes (deaths, active follow-up and lost to follow-up (LTFU)), we noted that more deaths than the number of patients in active clinic follow-up were observed around the eastern part of the district. Comparing the number of observed deaths to the number of LTFU patients, we noted again that there were more deaths observed in the northern outskirts and eastern parts of the district. In the northern part of the district, there is an HIV clinic that started providing ART in 2004 and this could explain the observed mortality due to ageing cohort. In eastern part, there is a lake and many fishing and other economic activities happen in the area and that could increase the risk of HIV/AIDS and mortality in this area.

When mapping TB episodes and KS using intensity maps, we noted that the majority of TB episodes observed in the district were likely to be observed in the western, eastern and north-eastern part of the district. We did not have an immediate explanation for this observation but around these areas of the map, there are two important features: the lake where many economic activities happen; and on the western part of the district, there is an HIV clinic which is one of the earliest clinics in the district.

In calculating the exceedance probabilities in Section 6.5.1, Section 6.5.2, and Section 6.5.3, we made an assumption that the standard errors for "‘cases"’ and

"'controls"' are independent. This is an assumption of convenience and will not be always true in practise, and so these exceedance probabilities are likely to be conservative.

When comparing results in Chapter 5, the effect of TB status and age at recruitment on mortality changed in a spatial survival model. The ART initiation age tended to be associated with mortality while TB was not significant. This change could be due to the addition of the spatial dimension to the model.

Model comparison showed that models with spatial frailties were out-performing model that included both patient-level frailties and spatial frailties. The use of grids was to reduce the computation time of fitting the models and does not affect model choice. Two competitive models were found: an Exponential model and Matern 1 (kappa=1) model both with spatial frailties only. The final model results and hazards prediction was based on Exponential model. Similar results can be obtained using a Matern 1 model. There was good mixing of chains and a reasonable convergence was reached.

Adjusting for age, TB status and spatial location, the following variables were significant in the model; sex, ART period and KS status. Male patients were at a greater risk of dying. The hazards of death were so higher in males than in females. These result in consistent with common knowledge about low health seeking behaviour in men compared to females. A multinational study in resource limited setting also found that female had low risk of death compared to males across all age groups (O'Brien et al., 2016).

Patients who started ART from 2008 to 2011 had significant reduction in hazards of death. This result can be explained by better service provision and better scale-up of HIV treatment programs over time. With the new efforts to start all HIV patients on treatment immediately, it will be interesting to see how results will compare spatially.

The exceedance probability maps show that HIV patients living in the northern and southern ends of the district tend to have lower survival. This result is also consistent with (sensitivity analysis) results obtained after adjusting for 60% deaths among loss to follow-up patients. However, after accounting for deaths amongst LTFU patients, the exceedance probability increased considerably from a maximum of 0.6 to 0.8 and 0.4 to 0.6 for HR >1.3 and HR > 1.5 respectively (20% increase in both cases). Patients living closer to the town have better prognosis than those leaving far away. We are not able to explain the poor survival observed in areas far north and south of the district because there are two relatively big HIV clinics around those areas. However, possible explanations could be that early ART initiations were from these regions and also the type of quality of care and treatment services provided in these regions. These geographical variations should not be ignored but should inform routine HIV programme and form a basis for further operational research to better understand these differences.

One of the limitations of this study is that there was high loss to follow-up which can impair estimation of survival in this cohort. However, our sensitivity analysis results support the general finding of existence geographical variations in survival in Zomba district. Also, we used data that is routinely collected as part of program

activities and quality of data has been a great issue in such study settings but the data were generated from the nationally standardized monitoring and evaluation tools used in the Malawi HIV programme and have been found to be of good standard.

In conclusion, this study highlights the geographical variations in survival in KS epidemiology and HIV treatment outcomes often not reported in most sub-Saharan Africa HIV programmes. We found significant geographical variation in survival among different sexes, ART period and KS status. The relative hazards of death increase in males, almost halved in patients who started ART during 2008 -2011 period, and a three-fold decrease in relative hazards in patients not diagnosed with KS. Overall variation in survival in the district should not be ignored in routine HIV programming. Similar analyses at regional and nation-level to study KS epidemiology and other types of cancer in HIV programmes could inform current efforts to improve HIV programming in resource-limited settings.

References II/6

Banda, M. et al. (2016) 'Spatial modelling of perinatal mortality in Mchinji, Malawi.', *Spatial and spatio-temporal epidemiology* 16, pp. 50–58.

Cuadros, D.F. and Abu-Raddad, L.J. (2014) 'Spatial variability in HIV prevalence declines in several countries in sub-Saharan Africa'.

Gelman, A., Hwang, J. and Vehtari, A. (2014) 'Understanding predictive information criteria for Bayesian models', *Statistical Computing* 24, pp. 997–1016.

Kazembe, L.N. and Kamndaya, M.S. (2016) 'Hierarchical spatial modelling of pneumonia prevalence when response outcome has misclassification error: Applications to household data from Malawi.', *Spat Spatiotemporal Epidemiol* 16, pp. 35–42.

Kazembe, L.N., Muula, A.S. et al. (2007) 'Modelling the effect of malaria endemicity on spatial variations in childhood fever, diarrhoea and pneumonia in Malawi.', *Int J Health Geogr* 25, p. 33.

Kazembe, L.N., Kleinschmidt, I. et al. (2006) 'Spatial analysis and mapping of malaria risk in Malawi using point-referenced prevalence of infection data.', *International journal of health geographics* 5, p. 41.

O'Brien, D. et al. (2016) 'Risk factors for mortality during antiretroviral therapy in older populations in resource-limited settings', *J Int AIDS Soc* 19, p. 20665.

Taylor, B.M., Davies, T.M. et al. (2015) 'lgcp:An R Package for Inference with Spatial and Spatio-Temporal Log-Gaussian Cox Processes', *Journal of Statistical software* 52, pp. 1–39.

World Health Organisation (2016) *Global Health Observatory Data: HIV/AIDS*, Available at http://www.who.int/gho/hiv/en/.

Chapter 7

Multistate Markov modelling in TB Epidemiology

Abstract

Multistate models in survival analysis are used for analysing data in which patients experience different intermediate states (events) before experiencing the event of interest. In this final case study, we consider the time to different states in which TB patients in HIV cohorts go through (TB free, TB infection, cured, relapse and death). A time-homogeneous Markov model is applied to a sample of 1483 patients randomly selected from an HIV cohort of approximately 20,000 patients. The total follow-up period was 120 months (10 years) and approximately 12.6% of total patients had at least one TB episode during the follow-up period. Mortality in these patients was around 9%, with most of patients that died being patients

that were diagnosed with TB after they had already started ART. The main transitions observed in this cohort were; remaining TB free, no-TB directly to death, TB infection to cured, no-TB then TB then cured, no-TB then TB then died, in that order. Patients that had finished TB treatment and were not cleared of TB at the end of TB treatment were more likely to restart TB treatment again (not cured). Mortality was three times higher in patients who got TB after starting ART than in patients who started ART while on TB treatment. Among patients with no TB at baseline, male patients were at significantly greater risk of dying than female patients. Also, patients that started ART from 2008 had significantly lower risk of TB infection if they did not have TB at the time of starting ART. Our study presents different transition quantities between the states of interest that can be easily be understood by clinicians and public health specialists. More importantly, the results provide clear pathways taken by HIV patients as they move from one TB state to another, an aspect mostly often not reported in TB analyses. More importantly, the clinical implications of these findings suggest that patients that had a non-conclusive TB treatment result could be restarted on TB treatment rather than wait for further confirmation results. Also, from public health interest, these findings suggest the protective effect of ART on TB episodes with patients that start TB after scale-up of ART having less likelihood of experiencing a TB episode.

7.1 Introduction

In Chapter 5 and Chapter 6, we modelled the time to an event happening, namely death. In addition, in Chapter 5 we considered; i) death as an outcome, ii) death as a competing risk to occurrence of lost to follow-up event i.e. dying before being lost to follow-up and lost to follow-up before death being recorded or observed. In both cases, one of the assumptions was that all patients had the same origin status (also called initial state). In Chapter 6, only the death outcome was considered but an extension was made to study how the hazard of death varied geographically. However in some diseases, patients go through multiple states before experiencing a terminal state (such as death). To get a better understanding of the disease progression, it is therefore important to model this process.

In this chapter, we consider an example of epidemiological study in which patients visit four (4) states; TB free, TB infected/Receiving TB treatment, TB treatment completed and death. These states are based on nationally standardised definition of TB treatment outcomes in Malawi. In particular, unlike having a single unique transition record per patient ($alive \rightarrow dead$ or to another competing event), we consider all transitions between states. The methods used to analyse these kinds of data are generally mathematical modelling of infectious diseases and 2) the multistate models, also known as event history modelling.

In general mathematical modelling of infectious diseases uses a set of differential equations. For these equations the following must be established or specified: the number of individuals who are susceptible to the disease; the number of individuals

exposed to the disease; the number of individuals infected by the disease and ; the number of individuals that recovered or died from the disease. All the parameters that govern the infection process ought to be specified too. Such parameters include; the infection rate, recovery rate, and the basic reproduction number of the disease. The modelling process becomes complex when modelling co-infections such as HIV and TB co-infection. Available literature on modelling the dynamics of HIV/TB co-infections have been done at cellular level, hence requiring immunological measurements such as T cells, CD4 cell count, CD8+ cell counts and viral load (Dodd et al., 2014; Shah and Gupta, 2014; Wang, Yang and Zhang, 2013; Castillo-Chavez and Song, 2004; Blower et al., 1995). These measurements indicate the level of compromise of an individual's immunity with TB and HIV co-infection. Since we did not have all these measurements, we did not consider mathematical modelling, and this is one of the limitations of these data.

In our our literature review, we did not find any studies on TB treatment outcome that used multistate models. This lack of literature in public domain is echoed in an overview by Beyersmann, Wolkewitz et al., (2011). However, multistate models have been used to study disease progression and efficacy of TB diagnostics and treatment regimen with studies from Malawi, South Africa, Zambia and Kenya (Keiser et al., 2011; Bwayo et al., 1995; Heymann, 1993). In other studies, Markov multistate models were used in economic evaluations of preventive therapy regimes of TB aimed at preventing the spread of tuberculosis (Mandalakas et al., 2013; Bell, Rose and Sacks, 1999; Bachmann, 2006).

7.2 Objective

We use these multistate models to study the epidemiology of TB in HIV cohorts. Specifically, the purpose of this chapter is to address the last objective: *Using multistate survival models, investigate the different infection pathways for individuals with both TB and HIV* as listed under section 2.2 on page 7. In particular, the goals of this analysis are to; i) model the probability of transitioning between each pair of states and ii) assess the effects of covariate on these transition probabilities. The overall motivation of this analysis is that a better understanding of the complex epidemiology of HIV and TB co-infection could be vital in providing insights into care and treatment practices in HIV programmes. A detailed description of methods used in multistate modelling have been described under subsection 4.3.3 from page 64.

7.3 Recap: HIV and TB co-infection

Tuberculosis is one of the most common opportunistic infections in persons living with HIV. The disease takes advantage of the compromised immunity the HIV patients have and being diagnosed with TB is a sign of progression to AIDS (CDC, 2016). TB examination is either through sputum smear examination (very common) or chest X-ray examination. Once diagnosed with TB, in addition to HIV treatment drugs (ART) patients are started on first line TB drugs for a minimum of six months. The most common drugs are isoniazid (INH) and rifampicin (RMP). Sometimes HIV patients develop a more complex form of TB known as

Multi-drug-resistant tuberculosis (MDR-TB). MDR-TB patients are treated for a longer period of time- usually a minimum of eight months (Malawi Ministry of Health, 2011). The possible outcomes of the TB treatments are cured, relapse, complete (TB outcome not yet confirmed), death or loss to follow-up. In low-income countries with high burden of HIV disease, research in TB epidemiology is becoming increasingly useful in informing the integration of HIV and TB services especially in sub-Saharan Africa. Some of the studies reporting efforts on the HIV and TB integration services include studies by Sculier, Getahun and Lienhardt, (2011), Zachariah et al., (2011) and Howard and El-Sadr, (2010) and Churchyard et al., (2007).

7.4 Data Description

The data used in this analysis were collected from 1, 500 HIV patients randomly selected (without replacement) from an HIV cohort with approximately 20, 000 patients on HIV treatment in Zomba district in Malawi. The selection of the random sample was restricted to only patients who started the lifelong HIV treatment antiretroviral therapy (ART) between 2004 and September 2011. The size of the sample was constrained by resources and time. After obtaining permission from the hospital officials, we organised and collected patients records for patients that started ART during the study period. With the help of five research assistants, we retrospectively reviewed paper records containing clinic visits for all 1, 500 selected patients and entered the data in an electronic database. Here, TB

is defined in general as having a TB diagnosis regardless whether it is a simple TB or MDR-TB. This is because the MDR-TB testing was not available for the majority of the study period. In addition, all the data from sputum smear and chest X-ray examinations are included. After data cleaning, only 1483 patients were eventually used in the final analysis, because the other 17 patients did not have complete information on important variables (sex, date of birth and date of starting HIV treatment). In total, there were more than 35, 000 clinic visits made by the 1483 sampled patients in the study period.

7.5 Multistate modelling

We identified four possible TB states (status) through which HIV patients go through. These states are;

- *TB free.* This is a state when patients do not have TB or have been cured after TB diagnosis and have successfully completed TB treatment as per national treatment guidelines. Patients in this state could be misdiagnosed as TB free and this in turn would overestimate the number of patients of cured of TB.

- *Receiving TB treatment.* This is a state for all patients diagnosed with TB and are receiving TB treatment in addition to their routine ART treatment. There are two scenarios for an HIV patient to be on TB treatment; Firstly, all HIV patients screened for TB at baseline and subsequent follow-up clinic visits. If diagnosed and confirmed with TB, they are started on TB treatment

in additional of their usual ART drugs. Secondly, the other entry point is the TB clinic. All patients in TB clinic are routinely offered HIV testing and counselling services as part their continuum care package. Therefore, at the time of starting HIV ART (HIV treatment) some patients might already have started TB treatment.

- *TB treatment completed.* The patients in this is state are patients that have completed TB treatment but their TB status or outcome is not yet confirmed. For some patients, this was taking up to three months to have TB status results. The possible final outcomes from this state are; 1) patients confirmed cured of TB (go to *state 1*) , 2) patients not cured are restarted on TB treatment (go to *state 2*), and 3) other patients die while awaiting ascertainment of TB status results (go to *state 4*).

- *Death.* The state for patients that died while on TB treatment or after successful TB treatment or died without any TB episode. The limitation of this categorisation is that such deaths are not attributed to TB/HIV as a cause, its treated as all-cause mortality even though TB/HIV could be the highly probable cause of death.

All possible transitions between these states are summarised in the flow diagram in Figure 7.1.

The recruitment (baseline) time was defined as the time since starting the lifelong antiretroviral therapy. We used a Counting Process (CP) format for our data. This means that the time to a particular state is calculated from the recruitment time

as opposed to resetting the clock to zero every time the patient leaves a particular state (see an example given in Table 4.3). The reason for adopting this type of time scale is because the order in which the states (events) occur is important in our data. For example, while you can die without having any TB episode ($1 \rightarrow 4$), it is impossible to be cured unless you first had TB and were on TB treatment. So TB infection starts first before being cured of TB.

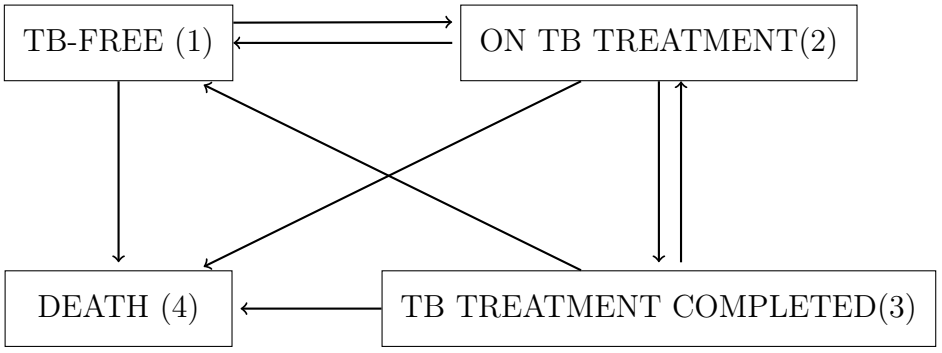

Figure 7.1: A model for TB events following initiation of HIV treatment(ART)

7.6 Model formulation and Parameter Estimation

Different approaches can be used to analyse multistate data in the counting process approach. In the book by Beyersmann, Allignol and Schumacher, (2012), the authors use a stratified Cox model to model transition hazards and probabilities. However in the present analysis, a Markov model was used to model different transition quantities of interest because the order in which events occurred is important in our data. This order of occurrence of events is not captured by a stratified Cox model. For instance, for patients to be in *state 3*, the must have been on TB treatment first and also one can be cured of TB if they were first

diagnosed of TB and put on treatment. The order of these possible transitions is very important especially in estimating the probability of transition to a next state i but not j.

In this study, we used a time-homogeneous Markov model. This assumption is important as it simplifies the interpretation of different transition quantities. However, this is a strong assumption especially in TB epidemiology where the hazards of TB infection could change over time and exposure due to change in public health systems. Nevertheless, the motivation and availability of software for analysing such data played a big role in choosing this model. This limitation is also reflected through out in the discussion of results from these models.

To complete the formulation of a Markov model, a transition matrix specifying all possible transitions must be defined. The diagonal elements are set to zero as they denoted staying in the same state. In this study, the transition matrix Q was defined as follow;

$$Q = \begin{Bmatrix} 0 & q_{12} & 0 & q_{14} \\ q_{21} & 0 & q_{23} & q_{24} \\ q_{31} & q_{32} & 0 & q_{34} \\ 0 & 0 & 0 & 0 \end{Bmatrix}$$

One of the goals of this analysis is to assess the effects of covariates on transition quantities. To do this, we introduce a model for the hazard of transition between each pair of states. We included covariates in our model in order to estimate the effect on transitions between states. Equation 7.1 to Equation 7.7 give the series

of all models for the transition hazards fitted in this study.

$$q_{12}(t,x) = q_0^{12}(t)\exp(X\beta_{12}) \tag{7.1}$$

$$q_{14}(t,x) = q_0^{14}(t)\exp(X\beta_{14}) \tag{7.2}$$

$$q_{21}(t,x) = q_0^{21}(t)\exp(X\beta_{21}) \tag{7.3}$$

$$q_{23}(t,x) = q_0^{23}(t)\exp(X\beta_{23}) \tag{7.4}$$

$$q_{24}(t,x) = q_0^{24}(t)\exp(X\beta_{24}) \tag{7.5}$$

$$q_{32}(t,x) = q_0^{32}(t)\exp(X\beta_{32}) \tag{7.6}$$

$$q_{34}(t,x) = q_0^{34}(t)\exp(X\beta_{34}) \tag{7.7}$$

where $q_0(t)$ is the baseline transition hazard at time t and x is an individual covariate value.

The estimation of the parameters $\beta_{12}, \beta_{14}, \beta_{21}, \beta_{23}, \beta_{24}, \beta_{32}$ and β_{34} is through maximum likelihood estimation using the likelihood function defined in Equation 4.49. Detailed description of these methods is given under subsection 4.3.3.

The `msm` was the man package used in model fitting. The other two packages developed by Beyersmann, Allignol and Schumacher, (2012) were also used in this analysis as follows; `mvna`-multivariate Nelson-Aalen estimator of the cumulative transition hazards for plotting cumulative transition hazards, and `etm`-Empirical Transition Matrix, also called Aalen-Johansen estimator for plotting transition probabilities.

7.7 Descriptive Analysis

The total follow-up period was 11.23 years with a total of 35, 263 patient visits. Among the 1, 483 HIV patients, a total of 130 (8.7%) patients died and a total of 187 (12.6%) patients had at least one TB episode (generally called HIV/TB co-infection). A total of 1194 patients did not have any TB episode during the follow-up period (Table 7.2).

Among 130 patients that died, 102 (78.5%) did not experience any TB episode during the follow-up period and a total of 28 (21.5%) of 130 patients that died had at least one TB episode. This represents approximately 14.9% deaths among the HIV/TB patients.

A total of 102 of 187 (54.5%) TB patients had TB at baseline, of which 8 (7.8%) died. The remaining 85 of the 187 TB patients were diagnosed with TB after starting ART, of which 20 (23.5%) died. Table 7.1 is a summary of all transitions that happened in this sample.

The fewer numbers in *state 3* is good from programme point of view since this means more patients get post-treatment results in good time, no need to be in *state 3* awaiting their final TB result. However, in modelling this may lead to great uncertainty due to large standard errors, hence wide confidence intervals.

From Table 7.1, there are no transitions from death (4) since it is an absorbing state. Since patients cannot transition from *state 4* (death), this is the only column that has unique number of patients and the column total is therefore equal to the

number of patients that died.

Table 7.1: Summary of state transitions. Transition numbers are as defined in Figure 7.1

		To			
		1	2	3	4
From	1	1325	87	0	104
	2	139	0	33	22
	3	7	5	0	4

There were a total of 20 unique transitions that patients passed through as summarised in Table 7.2. For instance, among the 130 patients that died, 102 moved directly from *state* 1 to *state* 4 (died without any TB episode during follow-up); 18 patients made the transition $1 \rightarrow 2 \rightarrow 4$ (no TB at baseline and got TB and died while on TB treatment); 4 patients made the transition $2 \rightarrow 4$ (had TB at baseline and died while on TB treatment); 2 patients made the transition $2 \rightarrow 3 \rightarrow 4$ (had TB at baseline and died while awaiting ascertainment of TB status after TB treatment), 2 patients made the transition $1 \rightarrow 2 \rightarrow 3 \rightarrow 4$ (no TB at baseline then got TB and died while awaiting ascertainment of TB status after TB treatment); and 2 patients made transitions $2 \rightarrow 1 \rightarrow 4$ (had TB at baseline and got cured of TB but later died).

A large number of patients had either one or two transitions only (Table 7.2). Seven (7) patients experienced a relapse of TB, with 4 out 7 of them being cured; the status of the other half was not ascertained by the time the study ended.

Table 7.2: A list of transitions and number of patients in each combination of transitions (1= TB free, 2= TB diagnosis and on TB treatment, 3=TB treatment Completed/Awaiting confirmation of TB Status, 4=Dead)

Serial #	Transitions	Number of patients
1	1→ 1	1194
2	1 → 4	102
3	2 → 1	78
4	1 → 2 → 1	49
5	1 → 2 → 4	18
6	2 → 3	9
7	1 → 2 → 3	6
8	2 → 3 → 1	5
9	1 → 2 → 1	4
10	2 → 4	4
11	2 → 3 → 4	2
12	1 → 2 → 3 → 2 → 1	2
13	1 → 2 → 3 → 4	2
14	2 → 1 → 4	2
15	1 → 2 → 1 → 2 → 1	1
16	1 → 2 → 1 → 2 → 3	1
17	1 → 2 → 3 → 2 → 3	1
18	1 → 2 → 3	1
19	2 → 3 → 2 → 1	1
20	2 → 3 → 2 → 3 → 1	1

For patients with no TB at baseline, the median time to TB diagnosis was 10.5 months (95% CI: 6.7, 14.3). There was no noticeable difference in mean ages between patients that had TB episodes and those without TB episode (Figure 7.2). However, there was noticeable differences in mean age between female and male patients, with male patients being older than females on average (Figure 7.3).

For patients that did not have TB at baseline, the probability of surviving a TB diagnosis in the first five years ($\mathbb{P}\{\text{surviving TB in t}\leq 5 \text{ yrs}|\text{no TB at baseline}\}$) was estimated at 93% [95% CI: 91 - 95]. The time variable in this case is defined as the time to TB diagnosis.

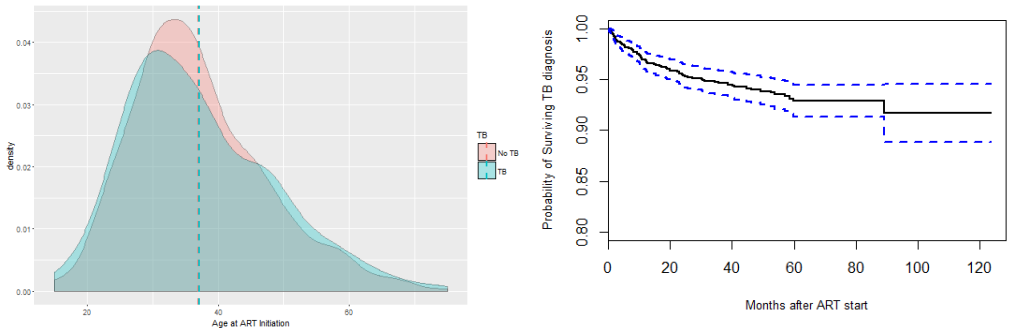

Figure 7.2: *Left*: Density plot for patients with and without TB episode during follow-up, with dotted lines representing the mean age in each group. *Right*: Probability of surviving TB diagnosis for patients who started HIV treatment without TB diagnosis at baseline.

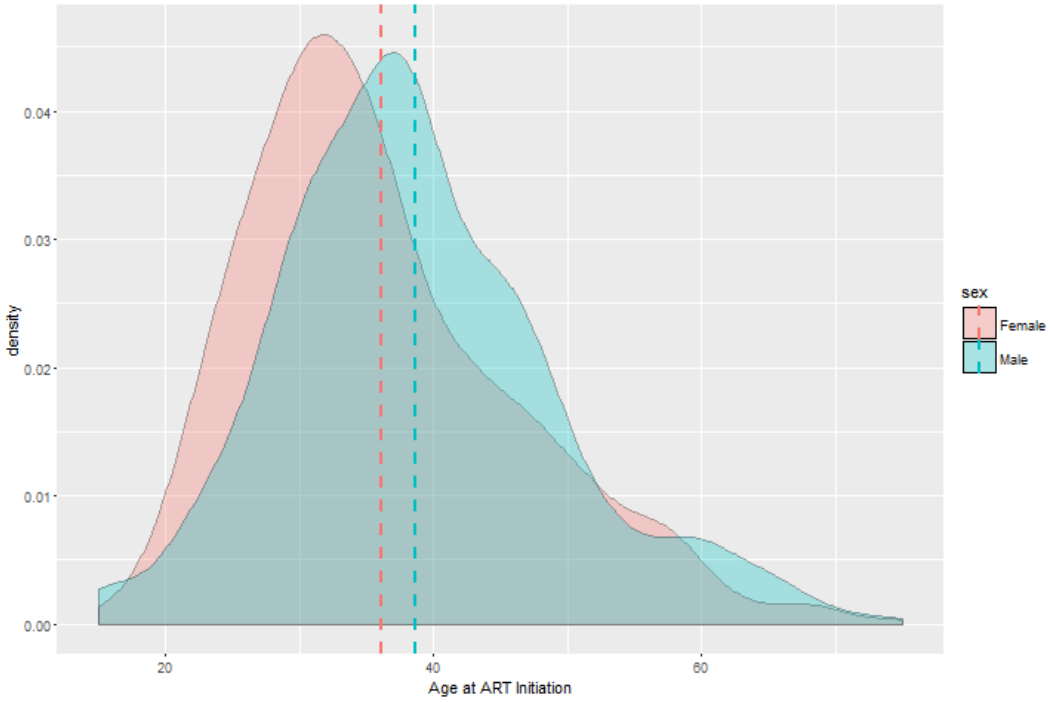

Figure 7.3: *Left*: Density plot for patients' age and sex (male/female). The dotted lines are the mean ages at initiation of ART for each sex category.

7.8 Model results and Prediction

We report the transition intensities, transition probabilities and transition hazards (see pages 73 and 74 for definitions). These estimates were obtained using methods implement in `msm` R package. Comparable results for transition probabilities and cumulative transition hazards[1] were obtained using `etm` and `mvna` packages respectively. All transition plots were done using `etm` and `mvna`. We discuss in turn each of these quantities in the next three subsections.

7.8.1 Transition Intensities

The transition intensities describe the rates of transitions from one state i at time t_1 into another state j (possibly $j = i$) at time t_2, $t_2 \geq t_1$. The row transition intensities in a transition matrix sum up to zero since the intensities of remaining in the same state are defined as a negative summation of transition intensities from this particular state to other states, $q_{ii} = -\sum_{j \neq i} q_{ij}$. Table 7.3 provides a summary of the crude (no covariates included) and baseline covariate-adjusted transition intensities with their corresponding 95% confidence intervals based on the normal distribution. These baseline adjusted transition rates are obtained with covariates set at their means. For intensity matrix that depends on a covariate, see Table 7.6 on page 204. The transition rates are calculated per 120 months (10 years), the length of the follow-up.

[1] Cumulative transition hazards: the cumulative hazards of transitioning from one state to another adjusting for patients' covariates.

Table 7.3: 10-Year Transition rates TR (95% CI) using a time-homogeneous Markov Model(Adjusted for: *initiation age, sex and ART period*)

Transition	Crude TR	95% CI	Adjusted TR	95% CI
$1 \rightarrow 1$	-0.00350	(-0.00405,-0.00303)	-0.00318	(-0.00373,-0.00272)
$1 \rightarrow 2$	0.001893	(0.00153,0.00234)	0.00172	(0.00137,0.00217)
$1 \rightarrow 4$	0.001611	(0.00132,0.00197)	0.00146	(0.00117,0.00182)
$2 \rightarrow 1$	0.192888	(0.16309,0.22813)	0.19622	(0.16515,0.23313)
$2 \rightarrow 2$	-0.26647	(-0.30743,-0.23097)	-0.25759	(-0.29996,-0.22120)
$2 \rightarrow 3$	0.04547	(0.03194,0.06474)	0.03909	(0.02533,0.06034)
$2 \rightarrow 4$	0.02811	(0.01772,0.04460)	0.02228	(0.01198,0.04142)
$3 \rightarrow 2$	0.08772	(0.03679,0.20914)	0.07883	(0.02203,0.28213)
$3 \rightarrow 3$	-0.15329	(-0.29306,-0.08018)	-0.14840	(-0.37588,-0.05859)
$3 \rightarrow 4$	0.06557	(0.02446,0.17578)	0.06957	(0.01813,0.26695)

The transition rate describes the rate of moving from one state to another. For instance, there were 2 per 1000 HIV patients per 10 years who were infected of TB and were put on TB treatment i.e they move from *state 1* to *state 2*.

As seen in Table 7.3, there are no big differences in estimates obtained from un-adjusted and adjusted model (adjusted for covariates).

The interest is on how transition from one state to another compare as well as how different combination of transitions compare. From Table 7.3, it can be seen that patients with no TB were 1.18 (95%CI: 0.85 - 1.64) times more likely to be diagnosed with TB (q_{12}) than die (q_{14}).

If patients were on TB treatment, the rates of moving to "cured state" were higher than dying while on TB treatment or having the ascertainment of their TB status prolonged after completing TB treatment. Patients were 5 (95% CI: 3 - 8) times likely to be declared cured of TB (q_{21}) than to remain without ascertainment of their TB status (q_{23}). Also if on TB treatment, patients were 9 (95% CI: 5 - 17) times more likely to complete TB treatment and be cured of TB (q_{21}) than die while on TB treatment (q_{24}). Once on TB treatment, patients were 69% more likely to complete the treatment (q_{23}) than die while on treatment (q_{24}).

Patients who completed TB treatment and were awaiting ascertainment of their TB status were 1.13 (95% CI: 0.17 - 7.17) more likely to restart TB treatment (not cured, q_{32}) than die (q_{34}).

7.8.2 Transition Probability

The transition probability is the probability of occupying a given state j at time t_2 conditional on occupying state i at time t_1, $t_2 \geq t_1$. When the new state occupied is an absorbing state, then the transition probability is equivalent to an event probability. This is also analogous to cumulative incidence functions discussed in the competing risk modelling (see subsubsection 4.3.3.1 for details on the cumulative incidence function).

Since probabilities are defined for periods, the use of landmark time points to aid interpretation of results is common practice in medical statistics. A landmark time point is a time s that is chosen to mark a milestone of interest and can be used to

check the progress of an outcome at different times. The landmark time approach was first proposed by Anderson, Cain and Gelber, (1983). When calculating the probability at a particular landmark time point, individuals are removed from the data if they experience the transition of interest before this time (Anderson, Cain and Gelber, 1983; Beyersmann, Allignol and Schumacher, 2012, p. 187). Two comparisons between transitioning to TB diagnosis and to death are considered as follows; $\mathrm{P}_{12}(s,t)$ vs $\mathrm{P}_{14}(s,t)$ and $\mathrm{P}_{14}(s,t)$ vs $\mathrm{P}_{24}(s,t)$ in Figure 7.4 and Figure 7.5 respectively. The estimated transition probabilities are based on Equation 4.47 on page 73. Let Y be a stochastic process for the TB states indexed by time on ART with T= c$(t_{min}, t_1, \ldots, t_{max})$, for $t_i < t_{i+1}$. These transition probabilities are defined as;

$$
\begin{aligned}
P_{12}(s,t) &= P(Y_t = 2 | Y_s = 1) \\
P_{14}(s,t) &= P(Y_t = 4 | Y_s = 1) \\
P_{24}(s,t) &= P(Y_t = 4 | Y_s = 2)
\end{aligned}
$$

In this analysis, three landmark time points were chosen at 6, 12, 18 and 24 months after starting HIV treatment which are regarded as crucial time points for HIV patients. The points were chosen based on the ART survival milestones reported by the national HIV programme in Malawi (Malawi Minstry of Health, 2011).

In Figure 7.4, it can be noted that for each fixed s, we have $\widehat{\mathrm{P}}_{12}(s,t) \leq \widehat{\mathrm{P}}_{14}(s,t)$. This difference is more pronounced with increasing period on ART for any fixed landmark time point s. In general, $\widehat{\mathrm{P}}_{12}(s,t)$ is decreasing over time while $\widehat{\mathrm{P}}_{14}(s,t)$

is increasing. Since TB diagnosis and on treatment (*state* 2) is a transient state, some patients leave the state explaining the fluctuating behaviour of the probabilities over time t for a fixed landmark time point s. Unlike patients on TB treatment (*state* 2), patients who die (*state* 4) do not leave the state (absorbing) hence the probabilities are bound to increase or be constant as time t increases for a fixed landmark time point s. The width of the confidence bounds (dotted lines) for $\widehat{\mathrm{P}}_{14}(s,t)$ is increasing indicate growing uncertainty in the estimated probabilities. If patients were in TB free (*state 1*) after a long time on ART, the likelihood of being diagnosed with TB decreases substantially although there are two spikes between t=80 and t=100 with very wide confidence bounds. A possible explanation for this observation could be due to the small number of TB cases during this period causing wide and unstable confidence bounds.

Overall in Figure 7.5, the probability of dying while TB free (includes those cured of TB) is consistently lower than dying while on TB treatment i.e. for each fixed s, $\widehat{\mathrm{P}}_{14}(s,t) \leq \widehat{\mathrm{P}}_{24}(s,t)$. This difference is much smaller if patients did not have TB in the first six months on ART but widens with increasing landmark time point. Also note the increasing uncertainty of $\widehat{\mathrm{P}}_{24}(s,t)$ with increasing values of s. However, overall these results indicate if patients with longer follow-up time are diagnosed with TB it is an indicator of worsening prognosis(as evidenced in increasing probability of dying). Patients with longer follow-up time given that they were TB free at s=18 and s=24 had a lower risk of death compared to the risk in early follow-up time.

Table 7.4 shows the estimated 10-year transition probabilities for remaining in the

same state and also transitioning to other states. The transition probabilities from the same state sum up to 1. The crude probabilities are obtained from a model without covariates and transition probabilities adjusted for age, sex and period of starting ART are presented under the "Adjusted" column. The estimates from both models lead to the same conclusions although 7 out of 12 times the adjusted probabilities were slightly lower than the crude estimates.

From Table 7.4, we note a very high probability of remaining in *state 1* using both crude and adjusted results (Probability:0.9967, 95%CI: 0.9820-0.9971). Some of the reasons for this high probability could be due to: fewer TB infections amongst patients who started HIV treatment while TB free at baseline; low number of TB relapses in patients that have successfully completed TB treatment and are cured; and also it could be due to the long period used (10 years) could be long. However, comparing transitions to other states, we note patients in *state 1* (TB free or cured of TB) had higher probability of being TB infected and put on TB treatment than dying (Table 7.4).

Once on TB treatment (state 2), patients were more likely to complete the treatment and be cured than dying while on TB treatment (0.169; 0.146-0.197). We also note that patients on TB treatment had the lower probability of transitioning to a death state was 3%, while transitioning to a state where they had to wait for their final TB treatment result was 4%.

For patients awaiting ascertainment of TB status after completing TB treatment (*state 3*), the probabilities of restarting TB treatment (not cured, going back to

state 2) and probability of dying were both higher than being declared TB free (cured, going to *state 1*). There was a 6% chance of dying if patients were in *state 3*. This result suggest restarting these patients on TB treatment would be a safer option from a public health perspective.

Table 7.4: Estimates of 10-Year Transition Probabilities* (95% CI) using a time-homogeneous Markov Model

Transition	Crude	Adjusted
$1 \rightarrow 1$	0.99667(0.99620, 0.99710)	0.99698(0.99642, 0.99740)
$1 \rightarrow 2$	0.00166(0.00133, 0.00203)	0.00152(0.00121, 0.00194)
$1 \rightarrow 3$	0.00004(0.00003, 0.00006)	0.00003(0.00002, 0.00005)
$1 \rightarrow 4$	0.00163(0.00133, 0.00201)	0.00147(0.00120, 0.00187)
$2 \rightarrow 1$	0.16914(0.14550, 0.19650)	0.17279(0.14754, 0.20140)
$2 \rightarrow 2$	0.76782(0.73700, 0.79210)	0.77429(0.74294, 0.80210)
$2 \rightarrow 3$	0.03691(0.02571, 0.05230)	0.03195(0.02060, 0.04774)
$2 \rightarrow 4$	0.02613(0.01752,0.04129)	0.02097(0.01255, 0.03705)
$3 \rightarrow 1$	0.00736(0.00313, 0.01609)	0.00676(0.00187, 0.02002)
$3 \rightarrow 2$	0.07120(0.02984, 0.15420)	0.06442(0.01769, 0.18320)
$3 \rightarrow 3$	0.85953(0.72980, 0.92110)	0.86336(0.66850, 0.93750)
$3 \rightarrow 4$	0.06191(0.02590, 0.16430)	0.06546(0.01894, 0.21780)

*$\mathbb{P}(0, t)$ for each pair of states, with t being the longest follow-up time.

Table 7.5 provides probabilities of the state being the immediate state to be visited, $-q_{ij}/q_{ii}$ state j to be visited immediately from state i. These probabilities are only from all possible transitions from state i. For example, only two transitions are possible from TB free status (*state 1*); transition to *state 2* (TB infection/TB treatment); transition to *state 4* (dying while TB free). So we that TB infection was likely to be the first state compared to death, with a ten-point difference in probability (54% vs 46%). However, patients on TB treatment had a very higher chance of being cured first before going to other states (72% vs 17% & 11%). Also,

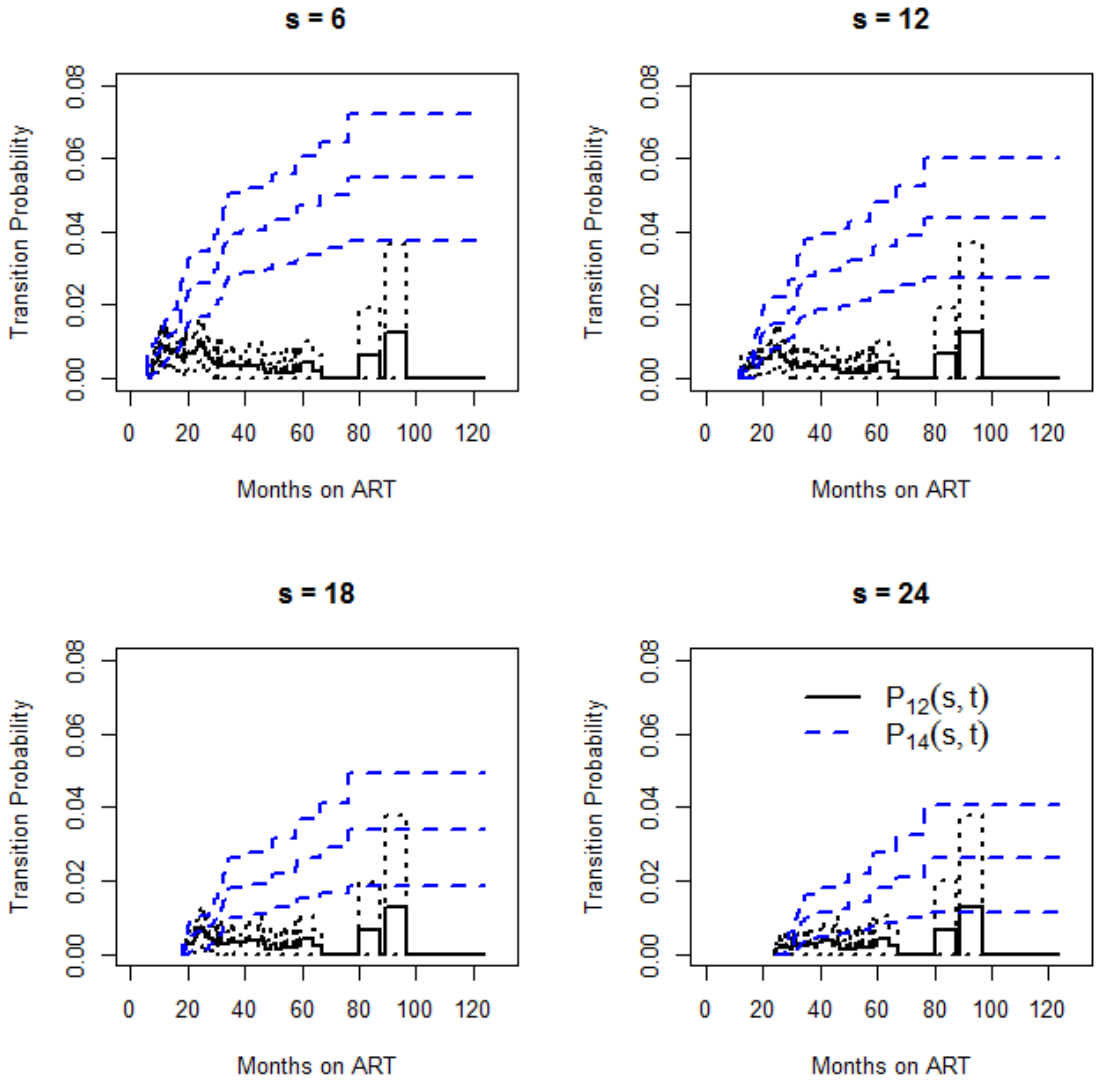

Figure 7.4: Probability of TB diagnosis ($P_{12}(s,t)$) compared to transitioning to a death state from a TB free state ($P_{14}(s,t)$). Probabilities are plotted at different landmark time points s since starting HIV treatment; 6 months (*top left*), 12 months (*top right*), 18 months (*bottom left*), 24 months (*bottom right*)

we note that patients in *state 3* were likely to transition to *state 2* first (restarting TB treatment) compared to moving to *state 4* (death).

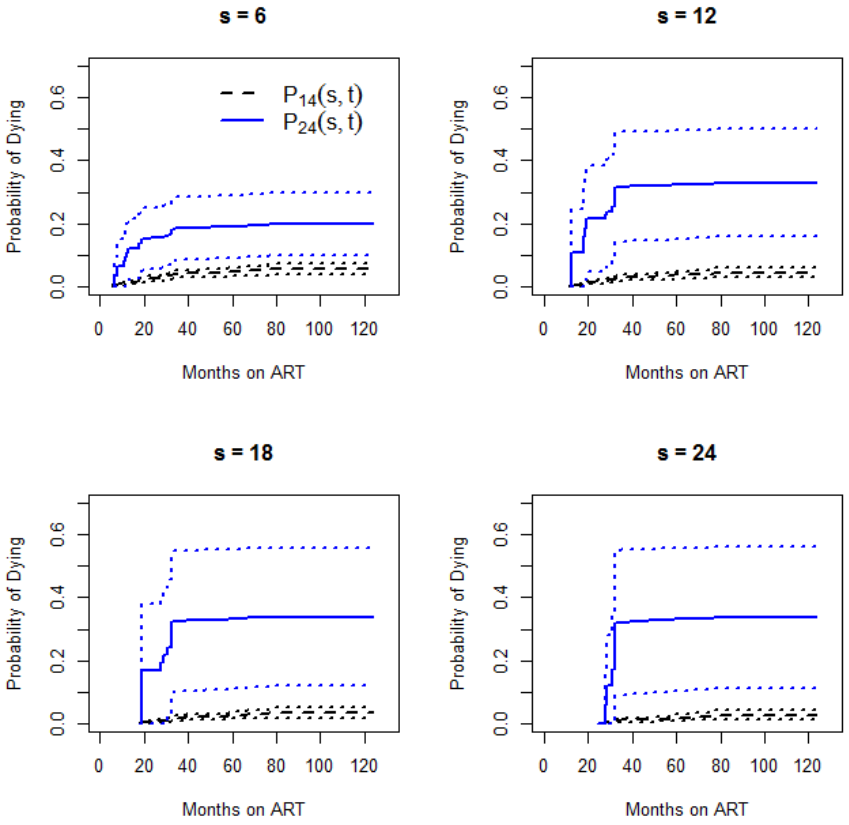

Figure 7.5: Comparing transition probabilities of death without TB ($P_{14}(s,t)$) versus death while on TB treatment ($P_{24}(s,t)$). Probabilities are plotted at different landmark time points s since starting HIV treatment; 6 months (*top left*), 12 months (*top right*), 18 months (*bottom left*), 24 months (*bottom right*)

Table 7.5: 10-Year Probability of being the next immediate outcome to be experienced

	No TB	TB	TB Complete	Death
No TB	0	0.54(0.47,0.61)	0	0.46(0.39,0.53)
TB	0.72(0.65,0.78)	0	0.17(0.12,0.23)	0.11(0.07,0.16)
TB Complete	0	0.57(0.28,0.83)	0	0.43(0.17,0.72)

7.8.3 Transition Hazards Ratios

One of the important goals in survival analysis is to assess the effects of covariates on the time to an event occurring. Like in the standard proportional hazards models, the transition hazards defined in Equation 7.1 to Equation 7.7 are assumed to be proportional transition hazards models. This assumption eases the interpretation of the estimated transition hazards. The likelihood function for these models is defined as in Equation 4.49 with the transition intensity function defined in terms of the corresponding transition probability.

Table 7.6 shows the estimated transition hazards ratios (THR) for age, sex and the era in which the patients started ART (ART period). The hazards of remaining in the same state are used as reference values (THR = 1). There were significant differences in the transition hazards of TB infection given they were TB free ($1 \rightarrow 2$), with patients who started ART after 2007 having a significantly reduced risk of TB infection (THR: 0.524, 95% CI : 0.339, 0.810). Compared to female patients, male patients had significantly higher hazards of death given they were TB free $1 \rightarrow 4$ (THR: 2.274, 95% CI: 1.508, 3.428). For patients on TB treatment (in *state 2*), there were no significant differences noted although a unit increased in age and being a female patient had lowering effects on the hazards of being cured of TB ($2 \rightarrow 1$).

Table 7.6: Estimates of 10-Year Transition Hazards Ratios (95% CI) using a time-homogeneous Markov Model(Reference values: *Female* for Sex and *2004-2007* for ART period)

Transition	Initiation Age	Sex: Male	ART Period:2008-2011
$1 \rightarrow 2$	1.014(0.992, 1.036)	1.303(0.825, 2.057)	**0.524(0.339, 0.810)**
$1 \rightarrow 4$	0.981(0.960, 1.003)	**2.274(1.508, 3.428)**	0.692(0.459, 1.046)
$2 \rightarrow 1$	0.998(0.982, 1.013)	0.858(0.599, 1.230)	1.248(0.853, 1.826)
$2 \rightarrow 3$	1.009(0.975, 1.045)	1.677(0.774, 3.632)	0.866(0.379, 1.978)
$2 \rightarrow 4$	1.037(0.994, 1.082)	1.600(0.580, 4.409)	0.461(0.171, 1.240)
$3 \rightarrow 2$	1.016(0.949, 1.087)	0.967(0.150, 6.237)	1.126(0.109, 11.644)
$3 \rightarrow 4$	0.975(0.888, 1.070)	0.297(0.038, 2.375)	0.163(0.016, 1.650)

In Figure 7.6, a comparison of cumulative hazards of death and TB diagnosis is made for patients that are in the TB-free state. It can be noted that cumulative transition hazards for death are higher (steep slope) than TB diagnosis in early days on ART treatment but the opposite is true for later periods.

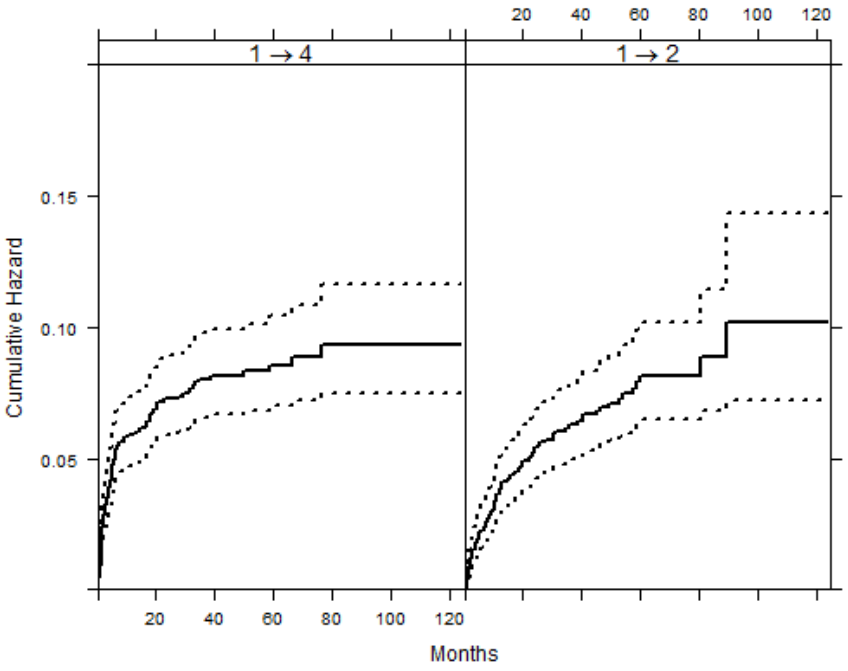

Figure 7.6: Cumulative transition hazards of death without TB diagnosis ($1 \rightarrow 4$) compared to the hazards of TB diagnosis ($1 \rightarrow 2$)

7.9 Model Assessment

It is important to assess if the fitted model suits the data well. In multistate models, two approaches are used: plotting the number of observed and expected counts in each state and Pearson-type goodness of fit (which is analogous to Pearson χ^2 test).

Suppose all patients start in state j and have a common initial time t. Then the expected number of patients in state k at time t is $n(t)P(t)_{j,k}$. Using this definition, we compare and plot at one-year intervals the expected and observed number of patients using the R-function `prevalence.msm` (Jackson, 2011). Here, prevalence is defined as percentages of individuals in each state at a set of times.

From Figure 7.7, it can be noted that the prevalence for the observed and expected are similar for "'TB Complete"' and "' On TB treatment"' states. However under this model, the observed and expected are not predicted well for the "'TB free"' and "'Dead"' states. Less individual were observed than expected in the "'TB Free"' state and more individuals were observed than expected in the "'Death"' state. This means that the model does not describe well these irregularly-observed transition processes. From about 50 months (4 years), the model tend to overestimate the predicted number of patients who die. Also at around the same time, the model tend to underestimate the number of patients who are TB free.

The possible reasons for this poor fit could be due to the following reasons. Firstly, this could mean failure of the Markov assumption, meaning the transition intensities may well depend on the time spent in previous states, the process history. To address this, (Jackson, 2011) suggest the use of a Semi-Markov process model as a possible solution but it is often difficult to account for the process history as the process is only observed through series of snapshots. The other reason for poor fit could be that the non-homogeneous time model is ideal for these data. To go around this problem, one way is to consider modelling the transition intensities as a piecewise-constant function of age.

The observed and expected number of patients in Pearson-type test are calculated by (Jackson, 2011; Titman, 2009; Aguirre-Hernández and Farewell, 2002):

$$O_{hl_hrscg} = \sum I\left\{S(t_{i,j+1}) = s, S(t_{ij}) = r\right\}$$

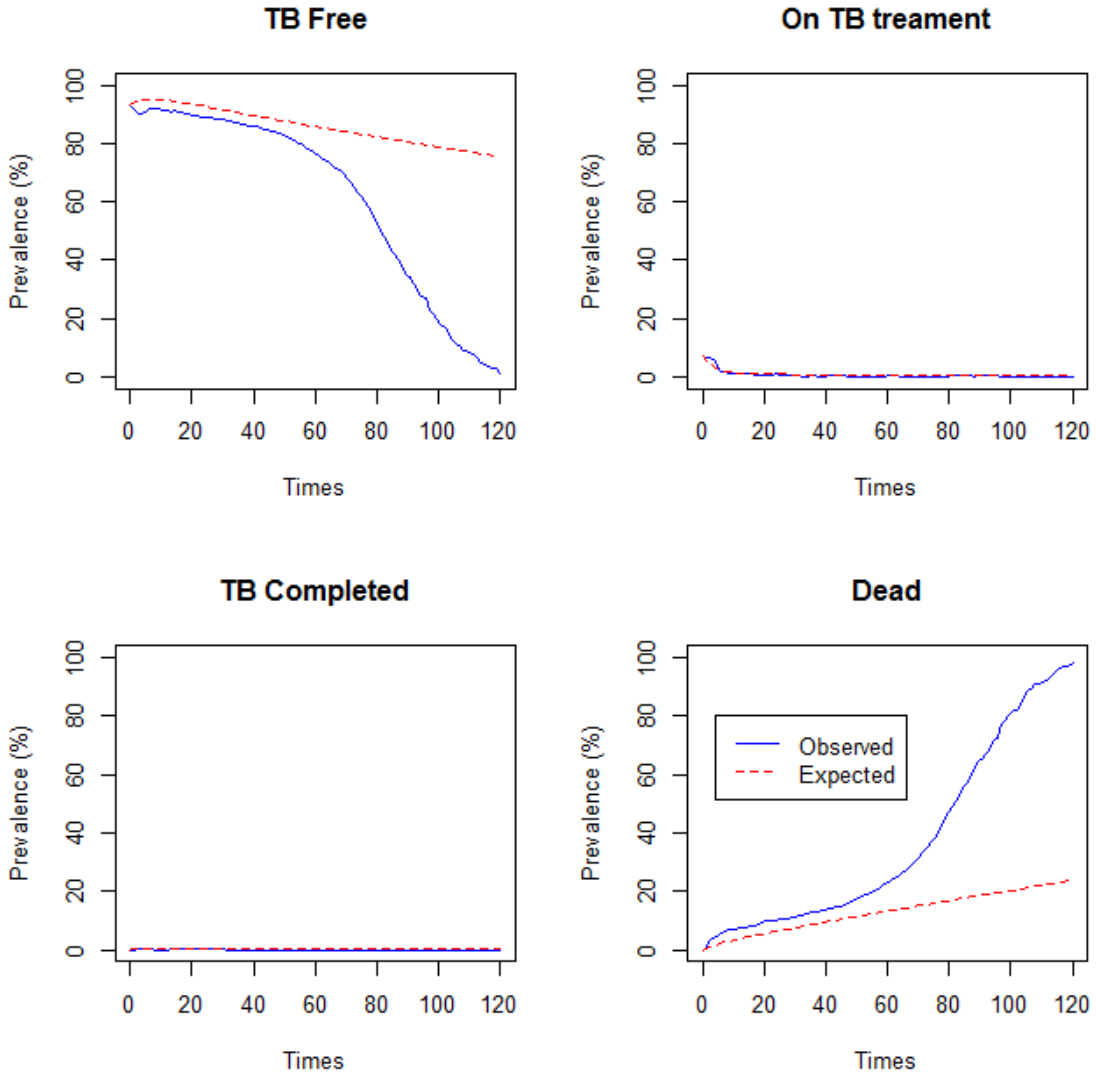

Figure 7.7: Comparing the observed and expected number of individuals in each state

$$e_{hl_hrscg} = \sum P\left\{S(t_{i,j+1}) = s | S(t_{ij}) = r\right\}$$

where r is the starting state, s is the finishing state, h is time between the start of the process and the first observed pair, l_h is the time interval between observations, g is for diagnosing lack of fit and c is the impact of covariates summarised by q_{rr}. Here $I(E)$ is the indicator function for an event E, the summation is over all individuals and over the set of transitions in the groups defined by h, l_h, c, g. The

test statistic T is then;

$$T = \sum_{hl_h rscg} \frac{O_{hl_h rscg} - e_{hl_h rscg}}{e_{hl_h rscg}}$$

The results in Table 7.7 from the Pearson test (using `pearson.msm` in R, see Jackson, (2011)) suggest there is a problem in the computation of the test. We get;

Table 7.7: Model diagnostics using Pearson-type test

stat	df.lower	p.lower	df.upper	p.upper
468	NA	NA	147	0

This means that the observed sample in some transition paths is not possible under the null hypothesis. This is noted in Table 7.2 that there are very few individuals in some transition paths (note: On TB treatment and TB complete are transient states). In this case, this test may not useful, as a result we use the plots above as a diagnostic tool.

7.10 Discussion and Conclusion

In medical and epidemiological studies, the odds ratios, relative risks and hazard ratios are the quantities often reported and well understood by wider readership in health field. However, in this study different quantities are modelled and their relationship explained. The quantities reported in this analysis are the transition intensities, transition probabilities and the transition hazard ratios. The overall

aim was to provide an alternative interpretation of model results in order to aid a better understanding of the epidemiology of diseases. In particular, in this study, these methods were used to explain and better understand the TB epidemiology in HIV cohorts.

We considered a multistate model as opposed to modelling as transition separately because this accounts for the fact that patients can also visit other competitive states. The results from the model better explains the transitions and their dependencies as patients move from one state to another. Moreover, the probabilities obtained from the models describe the random movements of patients providing possible competitive pathways patients are likely to take.

There were three assumptions made for this analysis. Firstly, it was assumed that the transitions we observed were unrelated to ART regimen and the latter was unobserved. Secondly, it was assumed that TB treatment type was not responsible for the transitions observed between the states. Lastly, our analysis focusses on all-cause mortality because data on cause-specific were not captured in this HIV cohort.

A larger proportion of TB patients were already on TB treatment at the time of starting HIV treatment. This could be that these patients were already infected with HIV and were just tested late for HIV since TB is an opportunistic illness that takes advantage of worsening body immunity. The main clinical implication of this result is to consider close monitoring of patients with HIV/TB co-infections.

The main transitions observed in this cohort were; remaining TB free, no TB to

death, TB infection to cure, no TB then TB then cure, no TB then TB then died in that order (Table 7.2). Mortality was three times higher in patients who got TB after starting ART than in patients who started ART while on TB treatment. Overall, mortality was low among TB patients (15%) compared to the rates reported from other studies in the tropics (Loveday et al., 2015; Velásquez et al., 2015; Diendéré et al., 2015; Vijay et al., 2011; Collins et al., 2010; Makombe, Harries, Yu et al., 2007). This result is consistent with current efforts that are being to made in intensifying TB screening in HIV cohorts. Form public health perspective, HIV programs should consider integrating more TB service in HIV cohorts rather than HIV patients accessing TB services in standalone TB clinics.

Both unadjusted and adjusted models lead to same conclusions though inclusion of patient covariates in adjusted models lowered the parameter estimates in the majority of the cases (transition intensities, probabilities and transition hazards). Patients in TB free status had a relatively higher probability of being diagnosed with TB than dying TB free. However, the cumulative transition hazards were higher for "TB free $\rightarrow$ death" transitions compared to the "TB free $\rightarrow$ TB infection" transitions during the early days on HIV treatment. Thereafter, the cumulative hazards are similar and tend to level-off with time. Clinically, this implies that the early period on ART is crucial and better practices for management HIV patients during this period have been advocated elsewhere (Rabie et al., 2015; Mosam, Uldrick et al., 2011; Vijay et al., 2011; Collins et al., 2010).

Once patients were on TB treatment, they were more likely to complete the treatment than to die while on TB treatment. There was an eight fold chance of

completing the TB treatment and be cured compared to patients dying while on the TB treatment. This also observed in high treatment TB treatment success rate of 76.5%. From clinical and public health perspective, this finding would be pleasing to clinicians and public health specialists because the use of ART is supposed to boost the immunity of HIV patients hence should have a spilling-over protective effect on opportunistic illnesses such tuberculosis. Several studies have reported the protective and beneficial effective of ART on opportunistic illness (Belayneh, Giday and Lemma, 2015; Vijay et al., 2011; Collins et al., 2010; Johansson, Robberstad and Norheim, 2010).

The majority of TB patients who completed treatment had their TB status ascertained soon after finishing TB treatment. However there were nine TB patients who, despite finishing their TB treatment, their ascertainment results about TB status after took longer than usual. For these patients, they were more likely to restart TB treatment compared to declared cured or dying while waiting confirmation results. The stock out of TB diagnostic test kits could be the main cause of the delay of after-treatment confirmation tests. While there are continued efforts in improving logistics in HIV and TB programming, countries in sub-Saharan countries face so many logistics due to demands from disease burdens such as malaria and cancer. These challenges if not addressed could erode the gains already made in controlling TB in the general population.

There were only two transitions in which significant differences in patient covariates were observed; transitions from TB free state to TB infection state and TB free state to death (Table 7.6). Patients who were TB free and had started HIV

treatment between 2008 and 2011 had significantly lower hazards of transitioning to TB infection status compared to the hazards of directly transitioning to the death state. Also TB free male patients were twice at risk of dying compared to TB free women. In all other transitions, no significant differences were noted neither by age, sex nor period of starting treatment. Gender differences in TB outcomes as well ART outcomes have been reported in other studies in Africa (Takarinda et al., 2015; Jarrin, Gestus and al, 2008).

The model assessment indicated that the homogeneous continuous time multistate Markov model did not consistently fit the data well. The poor prediction seen in t *state 1* (TB free) and *state 4* (death) could suggest a failure of the Markov assumption or indeed the time homogeneous assumption used. To address this prediction, one could consider modelling the transition intensities as a piecewise-constant function of age. Also, one could consider relaxing the Markov assumption and model the whole process history. Also, the Pearson-type test failed to be computed because of the small number of observations in some transitions especially transitions into "'TB treatment complete"' state. The limitation of this test was also highlighted by Aguirre-Hernández and Farewell, (2002).

The issue of lost follow-up (LTFU) has been discussed in all previous chapters. Unlike the first and second chapters, in this chapter the main focus was on the diagnosis of TB. The data on TB patients used in this analysis was collected and periodically updated at a district-level and was mostly complete with all outcomes ascertained in the majority of patients. Nationally, the TB treatment default rate (missing treatment) has fallen from 5% in 2006 to 2% in 2009 to 2010 period

(Malawi Ministry of Health, 2011). In this sampled data, a low proportion of loss to follow-up of approximately 10% was observed. We anticipate that even after accounting for deaths in the 10% LTFU group the reported results will not be drastically different from the ones reported in this study. Nevertheless, loss to follow-up in TB patient cohorts is still a big concern in many sub-Saharan countries. More importantly TB patients who default on their treatment pose a great danger of infecting other people in their communities. Poor adherence to TB treatment can lead to developing resistance to standard first line drugs. Many TB programs in Africa are doing commendable efforts to address this issue, with a more recent concerns for TB patients detained in prisons (Mburu et al., 2016; Puchalski Ritchie et al., 2015; Loveday et al., 2015; Feasey et al., 2013).

The limitations of this study are mainly failure to account for the impact of diagnostic tool (sputum smear versus X-ray) and type of TB (PTB, EPTB, MDR-TB). Including these variables would help to measure they impact transition quantities reported here. Also being a retrospective study, we were limited in terms of socio-demographics and clinical variables to include in the models. Being an airborne disease, TB prevalence varies by social factors such as being in crowded prisons and tobacco smoking. Inclusion of such social could provide an extra layer of understanding TB in HIV populations. Another limitation of this analysis is that the infection rate reported in this analysis could be lower than the rate in the general population. This is because we only used TB cases in HIV clinic and might have missed some TB cases in the general HIV population. Complete capturing of TB cases in many settings like Malawi with poor surveillance systems

is challenging. Lastly, the issue of loss to follow-up could also have an impact on the true outcomes of patients who did not die. Although the majority (>90%) of patients were alive and were in active follow-up, the other 10% of patients' statues were based on their last known clinic visit (a lag of 3 months). This could potentially underestimate the number of deaths not Tb infections due to its public health importance. Despite these limitations, the results of this analysis are consistent with results reported in other studies in the region. In additional, by using multistate models we have demonstrated how presenting different transitions in TB can improve our understanding of TB epidemiology in HIV cohorts. In our literature review, we noted that multistate models are widely applied in oncology and pharmacology (Svensson and Simonsson, 2016; Andersen and Keiding, 2002; Anderson, Cain and Gelber, 1983).

Further areas of research extending this work can be considered in the following areas. Firstly, it will be interesting to investigate if the introduction of the universal provision of ART to all HIV patients regardless of their CD4 levels will have an impact on incidences of TB episodes in HIV populations. As part of assessing the impact of this universal ART treatment policy, it will be of interest to see how transitions in TB states will be affected in the long run. Secondly, a possible methodology work is to extend the present work to include spatial effects (spatial multistate models). This could help to understand regional effects on the transition probabilities: geography is increasingly becoming an important element in the planning process of health programmes. With increasing development of computational tools and software, these more complicated models can be fitted

although this is not straight-forward.

In conclusion, the results in this analysis provide clear pathways taken by patients as they move from one state to another. It is noted that the following pathways were common; remaining TB free, no TB to death, TB infection to cure, no TB then TB then cure, no TB then TB then died. Cumulatively, we noted that there was a high probability that TB-free patients would die than TB infection as their immediate state. Also significant gender differences were noted in TB-free to death transitions and being diagnosed with TB significantly varied with the period of starting a lifelong HIV treatment. Although mortality among TB patients was low compared to other cohort studies in the region, continued efforts for early TB diagnosis and treatment initiation are crucial for public health reasons. In HIV cohorts, patients with long follow-up time should be given the necessary attention as they tend have a higher risk of TB infection. The methods presented in this chapter are a complementary way of interpreting results from comparable models. Apart from reporting the traditional hazards and time ratios in survival analysis, transition probabilities and intensities are also reported here. As argued, this could potentially increase in the way in which the public health readership community can understand the epidemiology of a particular disease. Continued operations research and best practices research in TB epidemiology is vital to inform the overall public health interventions for controlling TB, especially in this rapidly changing era of the lifelong antiretroviral therapy.

References II/7

Aguirre-Hernández, R. and Farewell, V.T. (2002) 'A Pearson-type goodness-of-fit test for stationary and time-continuous Markov regression models', *Statistics in Medicine* 21 (13), pp. 1899–1911.

Andersen, P.K. and Keiding, N. (2002) 'Multi-state models for event history analysis.', *Statistical methods in medical research* 11 (2), pp. 91–115.

Anderson, J.R., Cain, K.C. and Gelber, R.D. (1983) 'Analysis of survival by tumor response.', *Journal of clinical oncology* 1 (11), pp. 710–9.

Bachmann, M.O. (2006) 'Effectiveness and cost effectiveness of early and late prevention of HIV/AIDS progression with antiretrovirals or antibiotics in Southern African adults', *AIDS Care* 18 (2), pp. 109–120.

Belayneh, M., Giday, K. and Lemma, H. (2015) 'Treatment outcome of human immunodeficiency virus and tuberculosis co-infected patients in public hospitals of eastern and southern zone of Tigray region, Ethiopia.', *The Brazilian journal of infectious diseases* 19 (1), pp. 47–51.

Bell, J.C., Rose, D.N. and Sacks, H.S. (1999) 'Tuberculosis preventive therapy for HIV-infected people in sub-Saharan Africa is cost-effective', *AIDS* 13 (12), pp. 1549–1556.

Beyersmann, J., Allignol, A. and Schumacher, M. (2012) *Competing Risks and Multistate Models with R*, New York, NY: Springer New York.

Beyersmann, J., Wolkewitz, M. et al. (2011) 'Application of multistate models in hospital epidemiology: advances and challenges.', *Biometrical journal. Biometrische Zeitschrift* 53 (2), pp. 332–350.

Blower, S.M. et al. (1995) 'The intrinsic transmission dynamics of tuberculosis epidemics', *Nature Medicine* 1(8), pp. 815–821.

Bwayo, J.J. et al. (1995) 'Comparison of the declines in CD4 counts in HIV-1-seropositive female sex workers and women from the general population in Nairobi, Kenya', *Journal of Acquired Immune Deficiency Syndromes & Human Retrovirology:* 10 (4).

Castillo-Chavez, C. and Song, B. (2004) 'Dynamical models of tuberculosis and their applications', *Mathematical biosciences and engineering* 1(2), pp. 361–404.

CDC (2016) *Living with HIV: Opportunistic Infections*, http://www.cdc.gov/hiv/basics/livingwithhiv/opportunisticinfections.html.

Churchyard, G.J. et al. (2007) 'Tuberculosis Preventive Therapy in the Era of HIV Infection: Overview and Research Priorities', *The Journal of Infectious Diseases* 196 (s1), S52–S62.

Collins, J. et al. (2010) 'Effect of antiretroviral therapy on survival of HIV-infected tuberculosis patients in Peru', *Rev Panam Infectol* 12 (2), pp. 37–43.

Diendéré, E.A. et al. (2015) 'Clinical outcomes and mortality associated factors in patients infected with HIV receiving a presumptive anti-tuberculosis treatment in a tertiary level hospital in Burkina Faso.', *AIDS care* 27 (10), pp. 1250–4.

Dodd, P.J. et al. (2014) 'Burden of childhood tuberculosis in 22 high-burden countries: a mathematical modelling study', *The Lancet Global Health* 2(8), e453–e459.

Feasey, N.A. et al. (2013) 'How can the community contribute in the fight against HIV/AIDS and tuberculosis? An example from a rural district in Malawi.', eng, *The International Journal of Tuberculosis and Lung Disease* 8 (8), pp. 174–177.

Heymann, S. (1993) 'Modelling the efficacy of prophylactic and curative therapies for preventing the spread of tuberculosis in Africa', *Transactions of the Royal Society of Tropical Medicine and Hygiene* 87 (4), pp. 406–411.

Howard, A.A. and El-Sadr, W.M. (2010) 'Integration of Tuberculosis and HIV Services in Sub-Saharan Africa: Lessons Learned', *Clinical Infectious Diseases* 50 (s3), S238–S244.

Jackson, C.H. (2011) 'Multi-State Models for Panel Data: The msm Package for R', *Journal of Statistical Software* 38 (8), pp. 1–28.

Jarrin, I., Gestus, R. and al, E. (2008) 'Gender Differences in HIV progression to AIDS and Death in indistrialized Countries: Slower Disease Progression following HIV Seronconversion in Women', *American Journal of Epidemiology* 168, pp. 532–540.

Johansson, K.A., Robberstad, B. and Norheim, O.F. (2010) 'Further benefits by early start of HIV treatment in low income countries: Survival estimates of early versus deferred antiretroviral therapy', *AIDS Research and Therapy* 7, pp. 1–5.

Keiser, O. et al. (2011) 'Outcomes of antiretroviral treatment in programmes with and without routine viral load monitoring in Southern Africa.', *AIDS (London, England)* 25 (14), pp. 1761–9.

Loveday, M. et al. (2015) 'Community-based care vs. centralised hospitalisation for MDR-TB patients, KwaZulu-Natal, South Africa.', *The International Journal of Tuberculosis and Lung Disease* 19 (2), pp. 163–71.

Makombe, S.D., Harries, A.D., Yu, J.K.-L. et al. (2007) 'Outcomes of tuberculosis patients who start antiretroviral therapy under routine program conditions in Malawi', *Int J Tuberc Lung Dis* 38, pp. 412–416.

Malawi Ministry of Health (2011) *Guidelines for the use of Antiretroviral Therapy in Malawi*, tech. rep., Malawi Ministry of Health, Lilongwe http://www.who.int/hiv/pub/guidelines/malawi_art.pdf, Accessed: 2013-11-25.

Malawi Minstry of Health (2011) *ART Progress Report-Quarter 3*, Malawi Ministry of Health, Lilongwe.

Mandalakas, A.M. et al. (2013) 'Modelling the cost-effectiveness of strategies to prevent tuberculosis in child contacts in a high-burden setting.', *Thorax* 68 (3), pp. 247–55.

Mburu, G. et al. (2016) 'Detention of People Lost to Follow-Up on TB Treatment in Kenya: The Need for Human Rights-Based Alternatives', *Health and Human Rights Journal* 18 (1).

Mosam, A., Uldrick, T.S. et al. (2011) 'An evaluation of the early effects of a combination antiretroviral therapy programme on the management of AIDS-associated Kaposi's sarcoma in KwaZulu-Natal, South Africa.', *International journal of STD & AIDS* 22 (11), pp. 671–673.

Puchalski Ritchie, L.M. et al. (2015) 'A knowledge translation intervention to improve tuberculosis care and outcomes in Malawi: a pragmatic cluster randomized controlled trial.', *Implementation science : IS* 10, p. 38.

Rabie, H. et al. (2015) 'Tuberculosis: opportunities and challenges for the 90-90-90 targets in HIV-infected children.', *Journal of the International AIDS Society* 18 (Suppl 6), p. 20236.

Sculier, D., Getahun, H. and Lienhardt, C. (2011) 'Improving the prevention, diagnosis and treatment of TB among people living with HIV: the role of operational research.', *Journal of the International AIDS Society* 14 Suppl 1 (Suppl 1), S5.

Shah, N.H. and Gupta, J. (2014) 'Modelling of HIV-TB Co-infection Transmission Dynamics', *American Journal of Epidemiology and Infectious Disease* 2, pp. 1–7.

Svensson, R. and Simonsson, U. (2016) 'Application of the Multistate Tuberculosis Pharmacometric Model in Patients With Rifampicin-Treated Pulmonary Tuberculosis', *CPT: Pharmacometrics & Systems Pharmacology* 5 (5), pp. 264–273.

Takarinda, K.C. et al. (2015) 'Gender-related differences in outcomes and attrition on antiretroviral treatment among an HIV-infected patient cohort in Zimbabwe: 2007-2010.', *International Journal of Infectious Diseases* 30, pp. 98–105.

Titman, A.C. (2009) 'Computation of the asymptotic null distribution of goodness-of-fit tests for multi-state models', *Lifetime Data Analysis* 15 (4), pp. 519–533.

Velásquez, G.E. et al. (2015) 'Impact of HIV on mortality among patients treated for tuberculosis in Lima, Peru: a prospective cohort study', *BMC Infectious Diseases* 16 (1), p. 45.

Vijay, S. et al. (2011) 'Treatment Outcome and Mortality at One and Half Year Follow-Up of HIV Infected TB Patients Under TB Control Programme in a District of South India', *PLoS ONE* 6 (7), ed. by M. Pai, e21008.

Wang, X., Yang, J. and Zhang, F. (2013) 'Dynamic of a TB-HIV Coinfection Epidemic Model with Latent Age', *Journal of Applied Mathematics* 2013(2013).

Zachariah, R. et al. (2011) 'The HIV/AIDS epidemic in sub-Saharan Africa: thinking ahead on programmatic tasks and related operational research.', *Journal of the International AIDS Society* 14 Suppl 1 (Suppl 1), S7.

Part III

General Discussion and Further Research

The work in this thesis has addressed some of the most important research questions in HIV epidemiology: survival, geographical variation in risk and TB co-infection. As such, the purpose of this Part III is to summarise the discussion and conclusion points from presented in the three (3) results chapters: Chapter 5, Chapter 6 and Chapter 7.

In Chapter 5, we have reported risk factors associated with mortality in ART cohorts. In addition, we have provided a novel way of identifying the lower bound for the proportion of loss to follow-up (LTFU) patients that are really deaths. In summary, we found that the time of starting ART was associated with reduced hazards of death, while patients diagnosed with TB were twice at greater risk of death compared to those without TB. These results were also echoed in some results from sensitivity analysis of our approach of addressing LTFU. Our approach of addressing LTFU in HIV cohorts is one of the major contributions of this thesis as a whole. However, we hasten that to caution the reader that our approach is not the ideal method for correcting parameter estimates in the model using data with high LTFU. The methods for that purpose are available: see Kiragga et al., (2013) and Henriques et al., (2012) and also a dedicated HIV research group in South Africa, see `http://www.iedea-sa.org/` for details. The key public health message in this chapter concerns advocating for continued funding for HIV programs to better trace individuals who are lost from HIV/AIDS care.

In Chapter 6, we extended the analysis in Chapter 5 by addressing the aspect of geographical epidemiology. With the use of exceedance probability maps, we noted that the northern and southern parts of Zomba district had unusually high

probability of having greater hazards for death. The key contribution of this study to the general HIV epidemiology is the intuitive way of interpretive the risk in space. Identifying areas with poor survival outcomes becomes easy and this could facilitate knowledge sharing among policymakers and health practitioners. Such information could be used in the planning and distribution of resources targeting areas with high disease burden. By highlighting geographical variation in risk, this could also inform the global HIV/AIDS campaigns such as the 90-90-90 HIV treatment goals aimed at eradicating AIDS by 2030 (UNAIDS, 2014).

In Chapter 7, we have addressed another important opportunistic disease in HIV and AIDS cohorts. Due to compromised immunity, HIV patients are often vulnerable to attacks from several opportunistic illnesses including from TB infection (CDC, 2016; Chu, Mahlangeni et al., 2010). In this analysis, we noted that (all-cause) mortality was three times higher in HIV patients that contracted TB after they had started ART compared to those that were already on TB treatment. This result is indicative of worsening prognosis in patients that were diagnosed with TB after recruitment. Few relapses (7) were observed among the TB patients that successfully completed TB treatment at some time point. This low number of relapses is encouraging finding highlighting the effect of ART (by boosting immunity) and similar findings have been reported from studies in sub-Saharan Africa (Belayneh, Giday and Lemma, 2015; Johansson, Robberstad and Norheim, 2010). Also, we noted that once patients were on TB treatment, they were more likely to complete the treatment than to die while on TB treatment. This is good because to gives hope to the fight against TB. Lastly, we observed gender differences in mortality

risk, with men having high risk of death: should continue to worry clinicians and public health specialists. This undermines the continued efforts to eradicate TB and gender mainstreaming in TB treatment programmes could provide vital. We anticipate that the findings in this study will help to inform the current global campaigns to end TB by 2030: WHO End TB Strategy, with a target of 35% reduction in TB cases by 2020; Towards Zero campaigns by 2030- TB, AIDS, Malaria, Poverty etc (WHO, 2016)

A number of future research studies could be considered for both epidemiology and methodology purposes. For epidemiology research, future operational studies are needed to study geographical differences in KS epidemiology and treatment outcomes. Similar analyses at regional and nation-level to study KS epidemiology and other types of cancer in HIV programmes could inform current efforts to improve HIV programming in resource-limited settings. Also, there is need to find better metrics for distance from Zomba Clinic to the patient's location. Operational and best practices research should be encouraged in HIV and TB epidemiology in order to inform the overall public health interventions for controlling TB, especially in this rapidly changing era of HIV treatment. For methodology, a possible methodology work is to extend the present multistate models to include spatial effects (spatial multistate models) and their computation to supplement current work by Nathoo and Dean, (2008) and Brezger, Kneib and Lang, (2005). This could help to understand regional effects on the transition probabilities as geography is increasingly becoming an important element in epidemiology as a whole.

Complete Bibliography

Aalen, O., Borgan, O. and Gjessing, H. (2008) *Survival and Event History Analysis: A Process Point of View*, 1st, New York.

Adebayo, S.B. and Fahrmeir, L. (2005) 'Analysing child mortality in Nigeria with geoadditive discrete-time survival models.', *Statistics in medicine* 24 (5), pp. 709–28.

Adetunji, J. and Bos, E. (2006) *Levels and Trends in Mortality in Sub-Saharan Africa: An Overview*, 2nd, The International Bank for Reconstruction and Development/The World Bank, Chapter 2.

Aguirre-Hernández, R. and Farewell, V.T. (2002) 'A Pearson-type goodness-of-fit test for stationary and time-continuous Markov regression models', *Statistics in Medicine* 21 (13), pp. 1899–1911.

Alvaro-Meca, A. et al. (2012) 'Impact of comorbidities and surgery on health related transitions in pancreatic cancer admissions: a multi state model.', *Cancer epidemiology* 36 (2), e142–6.

Andersen, Per Kragh and Keiding, N. (2006) *Survival and Event History Analysis*, England: John Wiley & Sons Ltd.

Andersen, P.K. and Keiding, N. (2002) 'Multi-state models for event history analysis.', *Statistical methods in medical research* 11 (2), pp. 91–115.

Anderson, J.R., Cain, K.C. and Gelber, R.D. (1983) 'Analysis of survival by tumor response.', *Journal of clinical oncology* 1 (11), pp. 710–9.

Ando, T. (2010) *Bayesian Model Selection and Statistical Modeling*, Florida: CRC Press. Chapter 7.

Bachmann, M.O. (2006) 'Effectiveness and cost effectiveness of early and late prevention of HIV/AIDS progression with antiretrovirals or antibiotics in Southern African adults', *AIDS Care* 18 (2), pp. 109–120.

Bakuwa, R. (2010) 'Adoption of formal HIV and AIDS workplace policies: an analysis of industry/sector variations', *Journal of SAHARA* 7, pp. 2–9.

Banda, M. et al. (2016) 'Spatial modelling of perinatal mortality in Mchinji, Malawi.', *Spatial and spatio-temporal epidemiology* 16, pp. 50–58.

Banerjee, S. and Carlin, B.P. (2004) 'Parametric spatial cure rate models for interval-censored time-to-relapse data', *Biometrics* 60 (1), pp. 268–275.

Banerjee, S. and Dey, D.K. (2005) 'Semiparametric proportional odds models for spatially correlated survival data', *Lifetime Data Analysis* 11, pp. 175–191.

Banerjee, S. and Carlin, B.P. (2003) 'Semiparametric spatio-temporal frailty modeling', *Environmetrics* 14, pp. 523–535.

Banerjee, S., Carlin, B.P. and Alan E, G. (2004) *Hierarchical Modelling and Analysis for Spatial Data*, 2nd ed., Florida: CHAPMAN & HALL/RCR.

Banerjee, S., Wall, M.W. and Carlin, B.P. (2003) 'Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota', *Biostatistics* 4, pp. 123–142.

Belayneh, M., Giday, K. and Lemma, H. (2015) 'Treatment outcome of human immunodeficiency virus and tuberculosis co-infected patients in public hospitals of eastern and southern zone of Tigray region, Ethiopia.', *The Brazilian journal of infectious diseases* 19 (1), pp. 47–51.

Bell, J.C., Rose, D.N. and Sacks, H.S. (1999) 'Tuberculosis preventive therapy for HIV-infected people in sub-Saharan Africa is cost-effective', *AIDS* 13 (12), pp. 1549–1556.

Beyersmann, J., Allignol, A. and Schumacher, M. (2012) *Competing Risks and Multistate Models with R*, New York, NY: Springer New York.

Beyersmann, J., Wolkewitz, M. et al. (2011) 'Application of multistate models in hospital epidemiology: advances and challenges.', *Biometrical journal. Biometrische Zeitschrift* 53 (2), pp. 332–350.

Blower, S.M. et al. (1995) 'The intrinsic transmission dynamics of tuberculosis epidemics', *Nature Medicine* 1(8), pp. 815–821.

Breslow, N.E. (1975) 'Analysis of Survival Data under the Proportional Hazards Model', *International Statistics Review* 45, pp. 45–58.

Brezger, A., Kneib, T. and Lang, S. (2005) 'BayesX: Analyzing Bayesian Structured Additive Regression Models', *Journal of Statistical Software* 14 (11), pp. 1–22.

Bruin, G.P.D. and Stefan, D.C. (2013) 'Children with Kaposi Sarcoma in Two Southern African Hospitals: Clinical Presentation, Management, and Outcome', *Journal of Tropical Medicine* 2013(2013).

Bwayo, J.J. et al. (1995) 'Comparison of the declines in CD4 counts in HIV-1-seropositive female sex workers and women from the general population in Nairobi, Kenya', *Journal of Acquired Immune Deficiency Syndromes & Human Retrovirology:* 10 (4).

Castillo-Chavez, C. and Song, B. (2004) 'Dynamical models of tuberculosis and their applications', *Mathematical biosciences and engineering* 1(2), pp. 361–404.

CDC (2016) *Living with HIV: Opportunistic Infections*, http://www.cdc.gov/hiv/basics/livingwithhiv/opportunisticinfections.html.

Chi, B.H. et al. (2011) 'Universal definition of loss to follow-up in HIV treatment programs: a statistical analysis of 111 facilities in Africa, Asia, and Latin America.', *PLoS medicine* 8 (10), e1001111.

Chokunonga, E. et al. (1999) 'AIDS and Cancer in Africa: the evolving epidemic in Zimbabwe', *AIDS* 13, pp. 2583–2588.

Chu, K.M., Mahlangeni, G. et al. (2010) 'AIDS-associated Kaposi's sarcoma is linked to advanced disease and high mortality in a primary care HIV programme in South Africa.', *Journal of the International AIDS Society* 13, p. 23.

Chu, K., Misinde, D. et al. (2010) 'Risk factors for mortality in AIDS-associated Kaposi sarcoma in a primary care antiretroviral treatment program in Malawi.', *International health* 2 (2), pp. 99–102.

Churchyard, G.J. et al. (2007) 'Tuberculosis Preventive Therapy in the Era of HIV Infection: Overview and Research Priorities', *The Journal of Infectious Diseases* 196 (s1), S52–S62.

Collett, D. (2014) *Modelling Survival Data in Medical Research*, Third Edit, Boca Raton.

Collins, J. et al. (2010) 'Effect of antiretroviral therapy on survival of HIV-infected tuberculosis patients in Peru', *Rev Panam Infectol* 12 (2), pp. 37–43.

Cox, D.R.(R. and Miller, H.D.(D. (1965) *The theory of stochastic processes.* Methuen, p. 398.

Cox, R.D. (1972) 'Regression Models and Life-Tables', *Journal of the Royal Statistical Society, Series B* 34, pp. 187–220.

Cuadros, D.F. and Abu-Raddad, L.J. (2014) 'Spatial variability in HIV prevalence declines in several countries in sub-Saharan Africa'.

David, T.D. (2014) *HIV/AIDS, and Kaposi's Sarcoma*, http://www.webmd.com/hiv-aids/guide/aids-hiv-opportunistic-infections-kaposis-sarcoma_.

Dedicoat, M., Vaithilingum, M. and Newton, R. (2003) 'Treatment of Kaposi's sarcoma in HIV-1 infected individuals with emphasis on resource poor settings', *Cochrane Database Systematic Reviews*.

Diendéré, E.A. et al. (2015) 'Clinical outcomes and mortality associated factors in patients infected with HIV receiving a presumptive anti-tuberculosis treatment in a tertiary level hospital in Burkina Faso.', *AIDS care* 27 (10), pp. 1250–4.

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998) 'Model-Based Geostatistics', *RSS Applied Statistics* 47, pp. 299–350.

Diggle, P.J. and Ribeiro, P.J. (2007) *Modelbased Geostatistics*, New York: Springer.

Diva, U., Banerjee, S. and Dey, D.K. (2007) 'Modelling spatially correlated survival data for individuals with multiple cancers', *Statistical Modelling* 7, pp. 191–213.

Dodd, P.J. et al. (2014) 'Burden of childhood tuberculosis in 22 high-burden countries: a mathematical modelling study', *The Lancet Global Health* 2(8), e453–e459.

Efron, B. (1977) 'The Efficiency of Cox's Likelihood Function for Censored Data', *Journal of Maerican Statistical Association* 72, pp. 557–565.

Egger, M. et al. (2011) 'Correcting Mortality for Loss to Follow-Up: A Nomogram Applied to Antiretroviral Treatment Programmes in Sub-Saharan Africa', *PLoS Medicine* 8 (1), ed. by G.P. Bisson, e1000390.

Faraggi, D. and Simon, R. (1998) 'Bayesian variable selection method for censored survival data', *Biometrics* 54, pp. 1475–85.

Feasey, N.A. et al. (2013) 'How can the community contribute in the fight against HIV/AIDS and tuberculosis? An example from a rural district in Malawi.', eng, *The International Journal of Tuberculosis and Lung Disease* 8 (8), pp. 174–177.

Fenner, L. et al. (2013) 'Tuberculosis and the risk of opportunistic infections and cancers in HIV-infected patients starting ART in Southern Africa.', *Tropical medicine & international health : TM & IH* 18 (2), pp. 194–8.

Fine, J.P. and Gray, R.J. (1999) 'A Proportional Hazards Model for the Subdistribution of a Competing Risk', *Journal of the American Statistical Association* 94 (446), p. 496.

Freeman, E., Semeere, A. and Wenger, M.e.a. (2016) 'Pitfalls of practicing cancer epidemiology in resource-limited settings: the case of survival and loss to follow-up after a diagnosis of Kaposi's sarcoma in five countries across sub-Saharan Africa.', *BMC Cancer* 16, p. 65.

Freeman, E., Semeere, A., Wenger, M. et al. (2015) 'Pitfalls of practicing cancer epidemiology in resource-limited settings: the case of survival and loss to follow-up after a diagnosis of Kaposi's sarcoma in five countries across sub-Saharan Africa.', *BMC cancer* 16, p. 65.

Gbabe, O.F. et al. (2014) 'Treatment of severe or progressive Kaposi's sarcoma in HIV-infected adults.', *The Cochrane database of systematic reviews* Issue 9. Art. No.: CD003256.

Gelman, A., Carlin, J.B. et al. (2010) *Bayesian Data Analysis*, Second, Florida: CRC Press.

Gelman, A., Hwang, J. and Vehtari, A. (2014) 'Understanding predictive information criteria for Bayesian models', *Statistical Computing* 24, pp. 997–1016.

Graham, S. et al. (2013) 'Loss to Follow-Up as a Competing Risk in an Observational Study of HIV-1 Incidence', *PLoS ONE* 8(3), e59480.

Henderson, R., Shimakura, S. and Gorst, D. (2002) 'Modeling Spatial Variation in Leukemia Survival Data', *American Statistical Association* 97, pp. 965–972.

Henriques, J. et al. (2012) 'Comparison of Methods to Correct Survival Estimates and Survival Regression Analysis on a Large HIV African Cohort', *PLoS ONE* 7 (2), ed. by M.P. Fox, e31706.

Heymann, S. (1993) 'Modelling the efficacy of prophylactic and curative therapies for preventing the spread of tuberculosis in Africa', *Transactions of the Royal Society of Tropical Medicine and Hygiene* 87 (4), pp. 406–411.

Hinchliffe, S.R. and Lambert, P.C. (2013) 'Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions', *BMC Medical Research Methodology* 13(13).

Hosmer, D. and Lemeshow, S. (1999) *Applied Survival Analysis*, 2nd, New York: John Wiley & Sons.

Hougaard, P. (2000) *Analysis of Multivariate Survival Data*, New York.

Howard, A.A. and El-Sadr, W.M. (2010) 'Integration of Tuberculosis and HIV Services in Sub-Saharan Africa: Lessons Learned', *Clinical Infectious Diseases* 50 (s3), S238–S244.

Hurvich, C. and Tsai, C. (1986) 'Regression and time series model selection in small samples', *Biometrika* 76, pp. 297–307.

Ibrahim, J.G., Chen, M.-H. and Sinha, D. (2010) *Bayesian Survival Analysis*, New York: Springer.

Jackson, C.H. (2011) 'Multi-State Models for Panel Data: The msm Package for R', *Journal of Statistical Software* 38 (8), pp. 1–28.

Jarrin, I., Gestus, R. and al, E. (2008) 'Gender Differences in HIV progression to AIDS and Death in indistrialized Countries: Slower Disease Progression following HIV Seronconversion in Women', *American Journal of Epidemiology* 168, pp. 532–540.

Johansson, K.A., Robberstad, B. and Norheim, O.F. (2010) 'Further benefits by early start of HIV treatment in low income countries: Survival estimates of early versus deferred antiretroviral therapy', *AIDS Research and Therapy* 7, pp. 1–5.

Kalbfleisch, J.D. and Prentice, R.L. (2002) *The Statistical Analysis of Failure Time Data*, 2nd, Wiley Series in Probability and Statistics, Hoboken, NJ, USA: John Wiley & Sons, Inc.

Kaplan, E.L. and Meier, P. (1958) 'Nonparametric Estimation from Incomplete Observations', *Journal of American Statistical Association* 53, pp. 457–481.

Kazembe, L.N. and Kamndaya, M.S. (2016) 'Hierarchical spatial modelling of pneumonia prevalence when response outcome has misclassification error: Applications to household data from Malawi.', *Spat Spatiotemporal Epidemiol* 16, pp. 35–42.

Kazembe, L.N., Muula, A.S. et al. (2007) 'Modelling the effect of malaria endemicity on spatial variations in childhood fever, diarrhoea and pneumonia in Malawi.', *Int J Health Geogr* 25, p. 33.

Kazembe, L.N., Kleinschmidt, I. et al. (2006) 'Spatial analysis and mapping of malaria risk in Malawi using point-referenced prevalence of infection data.', *International journal of health geographics* 5, p. 41.

Keiser, O. et al. (2011) 'Outcomes of antiretroviral treatment in programmes with and without routine viral load monitoring in Southern Africa.', *AIDS (London, England)* 25 (14), pp. 1761–9.

Kendiga, C.E. et al. (2013) 'Cancer treatment in Malawi: A Disease of Palliation', *World Jouranl of Oncology* 4, pp. 142–146.

Kiragga, A. et al. (2013) 'Comparison of Methods for Correction of Mortality Estimates for Loss to Follow-Up after ART Initiation: A Case of the Infectious Diseases Institute, Uganda', *PLoS ONE* 8 (12), ed. by L. Kaderali, e83524.

Klein, J.P. et al. (2013) *Handbook of survival analysis*, p. 640.

Kleinbaum, D.G. and Klein, M. (2012) *Survival Analysis-A Self-Learning Text*, 2nd, USA: Springer.

Li, Y. and Ryan, L. (2002) 'Modeling Spatial Survival Data using Semiparametric Frailty Models', *International Biometric Society* 58, pp. 287–297.

Liang, H. and Zou, G. (2008) 'Improved AIC Selection Strategy for Survival Analysis', *Comput Stat Data Anal* 52(5), pp. 2538–48.

Loveday, M. et al. (2015) 'Community-based care vs. centralised hospitalisation for MDR-TB patients, KwaZulu-Natal, South Africa.', *The International Journal of Tuberculosis and Lung Disease* 19 (2), pp. 163–71.

Makombe, S.D., Harries, A.D. and al, E. (2008) 'Outcomes of patients with Kaposi's sarcoma who start antiretroviral treatment under routine program conditions in Malawi', *Tropical Doctor* 38, pp. 5–7.

Makombe, S.D., Harries, A.D., Yu, J.K.-L. et al. (2007) 'Outcomes of tuberculosis patients who start antiretroviral therapy under routine program conditions in Malawi', *Int J Tuberc Lung Dis* 38, pp. 412–416.

Malawi Ministry of Health (2003) *Guidelines for the use of Antiretroviral Therapy in Malawi*, tech. rep., Malawi Ministry of Health, Lilongwe.

Malawi Ministry of Health (2008) *Guidelines for the use of Antiretroviral Therapy in Malawi*, tech. rep., Malawi Ministry of Health, Lilongwe http://apps.who.int/medicinedocs/documents/s18803en/s18803en.pdf; Accessed: 2013-12-04.

Malawi Ministry of Health (2011) *Guidelines for the use of Antiretroviral Therapy in Malawi*, tech. rep., Malawi Ministry of Health, Lilongwe http://www.who.int/hiv/pub/guidelines/malawi_art.pdf, Accessed: 2013-11-25.

Malawi Minstry of Health (2011) *ART Progress Report-Quarter 3*, Malawi Ministry of Health, Lilongwe.

Malawi Minstry of Health (2014a) *ART Progress Report-Quarter 3*, Malawi Ministry of Health, Lilongwe.

Malawi Minstry of Health (2014b) *ART Progress Report-Quarter 3*, tech. rep., Malawi Ministry of Health, Lilongwe.

Malawi Minstry of Health (2016) *Malawi Integrated HIV Program Report, January-March 2016*, Malawi Ministry of Health, Lilongwe.

Malawi NSO (2008) *National Population Census 2008*, tech. rep., Malawi National Statistical Office(NSO), Zomba
http://www.nsomalawi.mw/index.php?option=com_content&view=article&id=106&Itemid=6, Accessed: 2013-12-10.

Mandalakas, A.M. et al. (2013) 'Modelling the cost-effectiveness of strategies to prevent tuberculosis in child contacts in a high-burden setting.', *Thorax* 68 (3), pp. 247–55.

Matern, B. (1986) *Spatial Variation*, Berlin: Springer-Verlag.

Mburu, G. et al. (2016) 'Detention of People Lost to Follow-Up on TB Treatment in Kenya: The Need for Human Rights-Based Alternatives', *Health and Human Rights Journal* 18 (1).

Minhas, V. and Wood, C. (2014) 'Epidemiology and transmission of Kaposi's sarcoma-associated herpesvirus', *Viruses* 6, pp. 4178–4194.

Mosam, A., Uldrick, T.S. et al. (2011) 'An evaluation of the early effects of a combination antiretroviral therapy programme on the management of AIDS-associated Kaposi's sarcoma in KwaZulu-Natal, South Africa.', *International journal of STD & AIDS* 22 (11), pp. 671–673.

Mosam, H.A.H.P., Carol, E. and al, E. (2008) 'Characteristics of HIV-1-associated Kaposi's sarcoma among women and men in South Africa', *International Journal of STD and AIDS* 1, pp. 400–405.

Msyamboza, K.P. et al. (2012) 'Burden of Cancer in Malawi; Common Types, incidence and Trends: National population-based cancer registry', *BMC Research Notes* 5, pp. 1–8.

Mwinjiwa, E. et al. (2013) 'Burden, characteristics, management and outcomes of HIV-infected patients with Kaposi's sarcoma in Zomba, Malawi', *Public Health Action* 3, pp. 180–185.

Nardi, A. and Schemper, M. (2003) 'Comparing Cox and parametric models in clinical studies', *Statistics in medicine* 22, pp. 3597–3610.

Nathoo, F.S. and Dean, C.B. (2008) 'Spatial Multistate Transitional Models for Longitudinal Event Data', *Source: Biometrics* 64 (1), pp. 271–279.

Nelson, B. et al. (2013) 'AIDS-associated Kaposi's sarcoma: Outcomes after initiation of antiretroviral therapy at a university-affiliated hospital in urban Zimbabwe', *International Journal of Infectious Diseases* 10, pp. 902–906.

Ngarina, M. et al. (2013) 'Reasons for poor adherence to antiretroviral therapy postnatally in HIV-1 infected women treated for their own health: experiences from Mitra Plus study in Tanzania', *BMC Public Health* 450 (13).

O'Brien, D. et al. (2016) 'Risk factors for mortality during antiretroviral therapy in older populations in resource-limited settings', *J Int AIDS Soc* 19, p. 20665.

Patel, K., Kay, R. and Rowel, L. (2006) 'Comparing proportional hazards and accelerated failure time models: an application in influenza', *Pharm Stat* 5(3), pp. 213–24.

Puchalski Ritchie, L.M. et al. (2015) 'A knowledge translation intervention to improve tuberculosis care and outcomes in Malawi: a pragmatic cluster randomized controlled trial.', *Implementation science : IS* 10, p. 38.

Rabie, H. et al. (2015) 'Tuberculosis: opportunities and challenges for the 90-90-90 targets in HIV-infected children.', *Journal of the International AIDS Society* 18 (Suppl 6), p. 20236.

Rachlis, B. et al. (2013) 'Using concept mapping to explore why patients become lost to follow up from an antiretroviral therapy program in the Zomba District of Malawi.', *BMC Health Serv Res* 13, p. 210.

Rohner, E. et al. (2014) 'Incidence rate of Kaposi sarcoma in HIV-infected patients on antiretroviral therapy in Southern Africa: a prospective multicohort study.', *Journal of acquired immune deficiency syndromes (1999)* 67 (5), pp. 547–54.

Royston, P. and Parmar, M. (2002) 'Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects', *Statistics in Medicine* 21, pp. 2175–2197.

Sankaranarayanan, R. et al. (2010) 'Cancer survival in Africa, Asia, and Central America: a population-based study', *Lancet Oncology* 11, pp. 165–173.

Sculier, D., Getahun, H. and Lienhardt, C. (2011) 'Improving the prevention, diagnosis and treatment of TB among people living with HIV: the role of operational research.', *Journal of the International AIDS Society* 14 Suppl 1 (Suppl 1), S5.

Shah, N.H. and Gupta, J. (2014) 'Modelling of HIV-TB Co-infection Transmission Dynamics', *American Journal of Epidemiology and Infectious Disease* 2, pp. 1–7.

Singogo, E. et al. (2013) 'Village registers for vital registration in rural Malawi', *Tropical Medicine & International Health* 18, pp. 1021–24.

Sitas, F., Bezwoda, W.R. et al. (1997) 'Association between human immunodeficiency virus type 1 infection and cancer in the black population of Johannesburg and Soweto, South Africa', *British Journal of Cancer* 75, pp. 1704–1707.

Sitas, F., Parkin, M. et al. (2006) *Disease and Mortality in Sub-Saharan Africa*, 2nd, The International Bank for Reconstruction and Development/The World Bank, Chapter 20.

Soko, D., Umar, E. and Lakudzala, A. (2012) 'HIV and AIDS workplace interventions; Gaps between policy and practice at the College of Medicine', *Malawi Medical Journal* 24, pp. 52–55.

Svensson, R. and Simonsson, U. (2016) 'Application of the Multistate Tuberculosis Pharmacometric Model in Patients With Rifampicin-Treated Pulmonary Tuberculosis', *CPT: Pharmacometrics & Systems Pharmacology* 5 (5), pp. 264–273.

Takarinda, K.C. et al. (2015) 'Gender-related differences in outcomes and attrition on antiretroviral treatment among an HIV-infected patient cohort in Zimbabwe: 2007-2010.', *International Journal of Infectious Diseases* 30, pp. 98–105.

Taylor, B.M., Davies, T.M. et al. (2015) 'lgcp:An R Package for Inference with Spatial and Spatio-Temporal Log-Gaussian Cox Processes', *Journal of Statistical software* 52, pp. 1–39.

Taylor, B.M. and Rowlingson, B.S. (2014) 'spatsurv: an R package for Bayesian inference with spatial survival models', *Available at https://arxiv.org/abs/1501.01665*.

Therneau, T.M. and Grambsch, P.M. (2000) *Modeling Survival Data: Extending the Cox Model*, New York: Springer.

Titman, A.C. (2009) 'Computation of the asymptotic null distribution of goodness-of-fit tests for multi-state models', *Lifetime Data Analysis* 15 (4), pp. 519–533.

UNAIDS (2014) *90-90-90- An ambitious treatment target to help end the AIDS epidemic*, http://www.unaids.org/en/resources/documents/2014/90-90-90.

UNAIDS (2015) *Malawi HIV and AIDS Estimates*, tech. rep., Switzerland: The Joint United Nations Programme on HIV/AIDS http://www.unaids.org/en/regionscountries/countries/malawi, Accessed: 2016-04-15.

UNAIDS (2016) *FACT SHEET 2016*, tech. rep., Switzerland: The Joint United Nations Programme on HIV/AIDS http://www.unaids.org/en/resources/fact-sheet, Accessed; 2016-04-15.

Velásquez, G.E. et al. (2015) 'Impact of HIV on mortality among patients treated for tuberculosis in Lima, Peru: a prospective cohort study', *BMC Infectious Diseases* 16 (1), p. 45.

Vijay, S. et al. (2011) 'Treatment Outcome and Mortality at One and Half Year Follow-Up of HIV Infected TB Patients Under TB Control Programme in a District of South India', *PLoS ONE* 6 (7), ed. by M. Pai, e21008.

Volinsky, C. and Raftery, A. (2000) 'Bayesian information criterion for censored survival models', *Biometrics* 56, pp. 256–62.

Wabing, H.R. et al. (2000) 'Trends in cancer incidence in Kyandondo County, Uganda, 1969-1977', *British Journal of Cancer* 82, pp. 1585–1592.

Wang, X., Yang, J. and Zhang, F. (2013) 'Dynamic of a TB-HIV Coinfection Epidemic Model with Latent Age', *Journal of Applied Mathematics* 2013(2013).

Watanabe, S. (2010) 'Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory.', *Journal of Machine Learning Research* 11, pp. 3571–94.

WHO (2014a) *Malawi Tuberculosis Financing Profile* , http://www.who.int/tb/country/data/profiles/en/.

WHO (2014b) *Prevalence of HIV among adults aged 15-49 (%)* , http://www.who.int/gho/hiv/epidemic_status/prevalence/en/.

WHO (2016) *WHO End TB Strategy: Global strategy and targets for tuberculosis prevention, care and control after 2015*, http://www.who.int/tb/post2015_strategy/en/.

Wienke, A. (2011) *Frailty models in survival analysis*, CRC Press.

World Health Organisation (2016) *Global Health Observatory Data: HIV/AIDS*, Available at http://www.who.int/gho/hiv/en/.

World Health Organization (2008) *Malawi Country Cooperation Strategy at glance*, tech. rep., World Health Organization, http://www.who.int/countryfocus/cooperation_strategy/ccsbrief_malawi_en.pdf?ua=1, Accessed: 2014-09-30.

World Health Organization (2011a) *Cancer Fact Sheet no 297*, tech. rep., World Health Organization, Geneva, http://www.who.int/mediacentre/factsheets/fs297/en/.

World Health Organization (2011b) *HIV/AIDS Fact Sheet*, tech. rep., World Health Organization, Geneva, http://www.who.int/hiv/topics/tb/hiv_tb_factsheet_june_2011.pdf, Accessed; 2014-01-15.

World Health Organization (2016) *Tuberculosis: Fact Sheet Fact sheet N°*104, tech. rep., World Health Organization, Geneva, http://www.who.int/mediacentre/factsheets/fs104/en/, Accessed: 2016-05-10.

Wu, X. et al. (2014) 'One hundred and five Kaposi's sarcoma patients: a clinical stud in Xinjiang, Northwest China', *Journal of European Academy of Dematology and Venereology* 28, pp. 1545–1552.

Yiannoutsos, C.T., An, M.-W. et al. (2008) 'Sampling-Based Approaches to Improve Estimation of Mortality among Patient Dropouts: Experience from a Large PEPFAR-Funded Program in Western Kenya', *PLoS ONE* 3 (12), ed. by B. Grinsztejn, e3843.

Yiannoutsos, C.T., Johnson, L.F. et al. (2012) 'Estimated mortality of adult HIV-infected patients starting treatment with combination antiretroviral therapy.', *Sexually transmitted infections* 88 Suppl 2, pp. i33–43.

Zachariah, R. et al. (2011) 'The HIV/AIDS epidemic in sub-Saharan Africa: thinking ahead on programmatic tasks and related operational research.', *Journal of the International AIDS Society* 14 Suppl 1 (Suppl 1), S7.

Zhang, J. and Lawson, A.B. (2011) 'Bayesian Parametric Accelerated Failure Time Spatial Model and its Application to Prostate Cancer', *Journal of Applied Statistics* 38(2), pp. 591–603.

Zhang, M.-J. and Fine, J. (2008) 'Summarizing differences in cumulative incidence functions', *Statistics in Medicine* 27(24), pp. 4939–4949.

Ziegler, J. et al. (2003) 'Risk factors for Kaposi's sarcoma: A case-control study of HIV-seronegative people in Uganda', *International Journal of Cancer* 103 (2), pp. 233–240.

Appendix: Addressing the STROBE Statement in Chapter 4, 5 and 6

	Item No	Recommendation	Sections where addressed
Title and abstract	1	(a) Indicate the study's design with a commonly used term in the title or the abstract	Page 90, 147,179
		(b) Provide in the abstract an informative and balanced summary of what was done and what was found	Page 90, 147,179
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	Page 92, 149,181
Objectives	3	State specific objectives, including any prespecified hypotheses	Page 94, 150,183
Methods			
Study design	4	Present key elements of study design early in the paper	Page 95, 150,184
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	Page 95, 150,184
Participants	6	(a) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up	Page 99, 150,184
		(b) For matched studies, give matching criteria and number of exposed and unexposed	
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	Page 99, 152,185
Data sources measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	Page 10, 98 for all

Bias	9	Describe any efforts to address potential sources of bias	Page 116, 169,208
Study size	10	Explain how the study size was arrived at	Page 95, 150,184
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	Page 10, 98 for all
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding	Page 99, 152,185
		(b) Describe any methods used to examine subgroups and interactions	Page 99, 152,185
		(c) Explain how missing data were addressed	Page 99, 152,185
		(d) If applicable, explain how loss to follow-up was addressed	Page 116, 169,208
		(e) Describe any sensitivity analyses	Page 116, 169,208
Results			
Participants	13	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	Page 101, 156,190
		(b) Give reasons for non-participation at each stage	- -
		(c) Consider use of a flow diagram	Page 101, 156,190
Descriptive data	14	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	Page 101, 156,190
		(b) Indicate number of participants with missing data for each variable of interest	Page -
		(c) Summarise follow-up time (eg, average and total amount)	Page 101, 156,190
Outcome data	15	Report numbers of outcome events or summary measures over time	Page 101, 156,190
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	Page 121, 163,194
		(b) Report category boundaries when continuous variables were categorized	Page 121, 163,194
		(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	-

Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	Page 133, 173,208
Discussion			
Key results	18	Summarise key results with reference to study objectives	Page 133, 173,208
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	Page 133, 173,208
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	Page 133, 173,208
Generalisability	21	Discuss the generalisability (external validity) of the study results	Page 133, 173,208
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	Page vi

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.