University of Nairobi

ISSN: 2410-1397

Master Project in Biometry

Joint Modelling of CD4 Count and Time To Wound Healing in HIV-Positive Men Following Circumcision

Research Report in Mathematics, Number 47, 2020

Okello Erick Otieno

December 2020

School of Mathematics

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Biometry

Master Project in Biometry **University of Nairobi**

December 2020

Joint Modelling of CD4 Count and Time To Wound Healing in HIV-Positive Men Following Circumcision

Research Report in Mathematics, Number 47, 2020

Okello Erick Otieno

School of Mathematics
College of Biological and Physical sciences
Chiromo, off Riverside Drive
30197-00100 Nairobi, Kenya

Master Thesis

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Biometry

Submitted to: The Graduate School, University of Nairobi, Kenya

Abstract

Background: Modelling longitudinal information and event time outcomes simultaneously helps in describing the progression of the disease over time. Past studies have mostly applied standard Cox proportional hazards model to establish the association between baseline CD4 count and time to wound healing following circumcision. However, Cox proportional hazards model does not take into account the special features of biomarkers besides not utilizing the entire longitudinal history of measurements. Consequently, results reported from Cox proportional hazards model could be biased or inefficient. To optimally investigate the association between CD4 count and time to wound healing, we used a joint modelling framework. In this framework, we utilized patients'entire longitudinal history of CD4 count, while also properly accounting for measurement error caused by biological variation and missing measurements.

Methods: In the first step, we fitted a linear mixed effects model to describe the evolution of square root CD4 count over time for each patient while adjusting for the priori selected baseline covarites. In the second step, we used the estimated evolution (square root CD4 count) in the Cox proportional hazards model to determine its relationship with time to wound healing. Some CD4 count values were missing for some patients at follow-up visits. This is a missing data problem synonymous with longitudinal studies and we assumed that the mechanism of missingness was missing at random (MAR), and thus, the results reported from the joint models, are still valid under MAR.

Results: 115 out 119 patients completed their follow-up visits and their wounds were certified fully healed. Median time to wound healing was 49 days (IQR:49-63). There was no association between the current true value of square root CD4 count and wound healing time (p-value=0.536). However, for patients with the same current true value of square root CD4 count at a given time point *t*, the log hazard ratio for a unit increase in the rate of change in square root CD4 count trajectory was 1.514 (95% CI: 1.121; 1.908).

Conclusion: Circumcising HIV-positive patients with any level of square root CD4 count is not harmful to their post-circumcision wound healing. However, patients with the same current true level of square root CD4 count could exhibit different slopes of the square root CD4 count trajectory at the same time point *t*, leading to different progression of wound healing between them.

Master Thesis in Mathematics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Mathematics

DISTRIBUTOR: School of Mathematics, University of Nairobi, Kenya

Declaration and Approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.

Signature Date

OKELLO ERICK OTIENO
Reg No. I56/11756/2018

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.

Njagi 04-11-2020

Signature Date

Dr. Edmund Njeru Njagi
Dept. of Non-Communicable Disease Epidemiology,
London School of Hygiene and Tropical Medicine,
London WC1E 7HT, United Kingdom.
E-mail: edmund.njeru.njagi@lshtm.ac.uk

November 11, 2020

Signature Date

Dr. Nelson Owuor
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: onyango@uonbi.ac.ke

Sarguta. 04-11-2020

Signature Date

Dr. Rachel Sarguta
School of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: rsarguta@uonbi.ac.ke

Dedication

To my dad and mum, this is for you both. God knows how much I love you.

Contents

Figures and Tables

Figures

Tables

Acknowledgments

First of all, I would like to thank God for giving me the resilience to complete this thesis in record time.

I am deeply grateful to my supervisor Dr. Edmund Njeru Njagi, of the London School of Hygiene and Tropical Medicine for his continuous support and guidance throughout my masters program. Dr. Njeru Njagi you have been a great mentor and I have really benefited a lot from our interactions. Even with the current difficulties the world is facing because of coronavirus, we have had successful virtual meetings which have been extremely insightful. I feel extremely honored being mentored by you.

Many thanks too to Dr. Nelson Owuor and Dr. Rachel Sarguta, for their valuable suggestions and inputs for this thesis. Your suggestions and inputs have been very helpful and I greatly acknowledge you.

My deepest gratitude to Nyanza Reproductive Health Society (NRHS) especially Dr. Fred Otieno and Prof. Robert C. Bailey for providing me with their research data for the purposes of this thesis. My sincere gratitude also goes to Phastar Limited for sponsoring my studies. Many thanks to The University of Nairobi especially School of Mathematics for being so supportive in the last 2 years.

To special colleagues at Phastar: Collins Mutua, you are just a wonderful boss. Kennedy Wanyonyi and Phelix Ojwang', thanks for the exam revision materials. James Mburu, Mary Wambua, Stella Mutua and Michael Ooko, thanks for your valuable thoughts and suggestions for this thesis.

To all my friends at UoN: Noel Kanini, Shelmith Macharia, Rachel Mburu, Maureen Mutisya, Stanley Nyoro, Emmanuel Ndubi, Abubakar Kalule, to name but a few; I greatly appreciate your friendship and support.

Lastly to my family members, thank you so much for your great love and support. To my best buddy Solange, I am grateful for always being there for me. Thank you all!

Okello Erick Otieno

Nairobi, 2020.

1 Introduction

1.1 Rationale

In most clinical studies, patients are followed until a predefined end point is observed. Research questions motivating joint modelling include:

1. Association between longitudinal information and time to an event.
2. Prediction of event based on longitudinal information.
3. Evaluating longitudinal information as a surrogate for event.

An example of the first case is in HIV/AIDS studies where a researcher could be interested in assessing how strongly associated current true level of CD4 count is with the risk for death (Mchunu, Mwambi, Reddy, Yende-Zuma, & Naidoo, 2020).

The rates of progression of medical conditions like HIV/AIDS do not just vary from one patient to another. They also do change dynamically over time for the same patient (Rizopoulos, 2012). Therefore, the true potential of a marker in describing progression of a disease and its relationship with event outcome can only be shown when the whole longitudinal history of the biomarker is incorporated into the analysis (Rizopoulos, 2012). The motivation behind joint models is to pair the event-time submodel with a suitable longitudinal submodel taking into account unique features of the marker: (Rizopoulos, 2012):

1. Measured with error due to biological variation.
2. Longitudinal information only intermittently observed (the complete history is not available).
3. Existence of the biomarker is directly related to the event outcome/failure status.

1.1.1 Survival Data

The primary aim of many clinical studies is to analyze times from origin until an occurrence of an event. Examples include time to wound healing following circumcision (Rogers, Odoyo-June, Jaoko, & Bailey, 2013; Tshimanga et al., 2017) or time to rehospitalization (Njagi, Rizopoulos, Molenberghs, Dendale, & Willekens, 2013) amongst others. Within survival analysis framework, the outcome variable is event time (failure time or survival time) (Rizopoulos, 2012). The main distinguishing characteristics of survival analysis from other standard statistical techniques is **censoring** i.e., the time to an event of interest is not fully observed for all patients under study (Rizopoulos, 2012) (Collett, 2015; Rizopoulos, 2012). Several types of censoring exist:

1. Location of the true event time with respect to the censoring time: *right, left and interval.*
2. Probabilistic relation between the true event time and the censoring time: *informative and non-informative*

In our analysis, we assumed that censoring mechanism was non-informative right censoring. Methods of analysing survival data can either be through a parametric model typically Weibull, Gamma or Exponential distributed baseline hazard function or through a semi-parametric Cox proportional hazards model (Collett, 2015). Throughout our analysis, we applied Cox proportional hazards model to determine priori selected baseline covariates (excluding CD4 count) associated with time to wound healing. In Section 3.2, we have given an in-depth review of the Cox proportional hazards model.

1.1.2 Longitudinal Data

Longitudinal data refers to measurements on one or more variables that are repeatedly collected over time for each subject (Verbeke, 1997). The key components of longitudinal data are that the repeated measurements of a variable within a subject tend to be correlated with each other. This means that there may be within-subject correlations. The second component stems from inter-subject variability. This implies that the heterogeneity between subject profiles and the measurements across subjects are often assumed to be uncorrelated (Verbeke, 1997). The measurements between two subjects can be assumed independent if the subjects are randomly selected. Further, there are some measurement errors due to biological variation especially for biomarkers like CD4 count (Verbeke, 1997). Longitudinal data can either be continuous (Verbeke, 1997) or discrete (Molenberghs & Verbeke, 2006). In longitudinal studies, we expect positive correlation between repeatedly collected measurements from the same subject. This characteristic therefore makes it inappropriate to analyse our data using classical regression models like the ordinary linear regeression models or generalized linear models that assume independence of observations (Rizopoulos, 2012). There are two popular methods for analysing longitudinal continuous data: a generalized estimating equations model (GEEs) and mixed effects regression (MER) (Garcia & Marder, 2017). We can model both time-independent covariates and time-dependent covariates (e.g., CD4 count) when we use the two models. Besides that, these two models can account for unbalanced and missing values in a covariate without the need for imputing the missing values. Despite this, GEEs are not efficient when the missingness is as a result of missing at random (MAR). On the contrary, MER models are efficient provided that the mean and variance-covariance structure are correctly specified (Garcia & Marder, 2017; Fitzmaurice, Laird, & Ware, 2012). Because of this flexibility, MERs are highly preferred to GEEs when analysing longitudinal data (Garcia & Marder, 2017). In this study we have applied MER more specifically linear mixed effects model to analyse CD4 count. Linear mixed effects model have been extensively discussed in Section 3.3.

1.2 Background

Past clinical trials by Auvert et al. (2005); Bailey et al. (2007); Gray et al. (2007) reveled that medical male circumcision (MMC) reduces HIV transmission among heterosexual persons by approximately 50%-60%. In areas where male circumcision is less practiced and epidemic mostly severe, medical male circumcision can be away of reducing HIV epidemic (Hallett et al., 2008). Based on this strong and consistent evidence, in 2007 medical male circumcision was proposed as one of the HIV prevention strategies by the Joint United Nations Programme on HIV/AIDS (UNAIDS) together with the World Health Organisation (WHO) (Njeuhmeli et al., 2011). Previously before the era of test and treat, CD4 count threshold of less than 350 cells/µL had been the most popularly used biomarker for initiation of antiretroviral therapy (ART) (Kigozi et al., 2014).

It is against this backdrop that Kigozi et al. (2014); Rogers et al. (2013); Tshimanga et al. (2017) conducted 3 different studies to investigate the relationship between CD4 count and wound healing time following circumcision. However, Kigozi et al. (2014); Rogers et al. (2013) relied on the Cox proportional hazards models approach to establish the underlying association between wound healing time and baseline CD4 count. On the other hand, Tshimanga et al. (2017) employed a binomial probability test to evaluate equivalence of proportion of patients healed by baseline CD4 count ($<$ 500 cells/µL Vs. $\geq$ 500 cells/µL).

While applying these analyses methods in their studies, Kigozi et al. (2014); Rogers et al. (2013); Tshimanga et al. (2017) only utilized CD4 count at baseline. By only utilizing the baseline CD4 count, the true potential of CD4 count in describing the progression of the disease and its relationship to wound healing was potentially not revealed.

Thus, to optimally understand the association between CD4 count and wound healing time, we used a joint modelling framework. In this framework, we will utilize patients' entire longitudinal history of CD4 count, while also properly accounting for measurement error due to biological variation and missing measurements.

1.3 Problem Statement

Three clinical studies done in the past reported consistent results on the association between CD4 count and time to wound healing among HIV positive men. A study conducted in Uganda using surgical dorsal slit reported that the median time to wound healing among HIV positive patients was 4 weeks. However, time to wound healing did not vary based on the baseline CD4 count of either < 350 cells/µL or ≥ 350 cells/µL (p-value>0.05) (Kigozi et al., 2014). Similarly, a Prepex study in Zimbabwe reported that the median time to wound healing among HIV positive patients was 42 days. Still, time to wound healing of patients with baseline CD4 count < 500 cells/µL was not different from the patients with baseline CD4 count ≥ 500 cells/µL (p = 0.66) (Tshimanga et al., 2017). Lastly, a study in Kenya by Rogers et al. (2013) reported that wound healing time did not vary by baseline CD4 count among HIV-positive patients (p = 0.20). Specifically Rogers et al. (2013) found out that the mean wound healing time was 34.5 days for those with CD4 count ≤ 350 cells/µL and 31.9 days for those with CD4 count > 350 cells/µL.

The common limitation of these three studies was that only baseline CD4 count was utilized. By only using the baseline CD4 count as opposed to the whole longitudinal history, the opportunity to assess its trajectory over time and its relationship to wound healing time was lost. In the present study, CD4 count was collected for every patient on the circumcision day and at subsequent weekly follow up visit until wound was certified to be fully healed. Indeed, in many health research, the longitudinal biomarkers are recorded together with event time of interest. If the biomarkers are repeatedly collected over time, then it is most informative to utilize all the information collected when estimating the model parameters. The present study sought to fill in these gaps by applying a joint modelling framework. In the best of our knowledge, no published study in the past has applied joint modelling technique to clearly understand the association between longitudinally collected CD4 count and wound healing time following circumcision. Joint modelling technique is an improvement over traditional Cox proportional hazards model because it utilizes data from both longitudinal and survival processes simultaneously leading to less biased estimates of the parameters.

1.4 Objectives

1.4.1 Overall Objective

To determine the association between CD4 count and wound healing time in HIV positive men following circumcision.

1.4.2 Specific Objectives

1. To determine the association between CD4 count and wound healing time, considering patients' entire CD4 count longitudinal history.
2. To determine the association between the rate of change in CD4 count and wound healing time.

1.5 Significance of Study

Our study reveals the relevance of the application of joint models to answer specific epidemiological questions in HIV research. It also paves way to explore other complex association structures between longitudinal information and time to event outcomes in medical male circumcision studies.

The remainder of our thesis is arranged as follows: In Chapter 2, we have reviewed the joint modelling framework. Chapter 3 describes statistical methods on joint models. The study then applies aforementioned methodologies to real HIV data in Chapter 4. Chapter 5 is a discussion of results and conclusions arrived at in the study.

2 Literature Review

2.1 Preliminaries

Chapter 2 gives a review of the joint modelling framework, extensions, applications, underlying assumptions, strengths and weaknesses.

2.1.1 Review of Joint Modelling

Joint models for longitudinal information and event time outcomes entails improving inferences for event time outcomes while accounting for the longitudinal information collected intermittently and with error (Njagi et al., 2013; Rizopoulos, 2012). Several studies in the past have mostly applied Cox proportional hazards model using the baseline covariates. Such an approach is only reliable when we assume that the covariates remain constant during the study period. Unfortunately, this is an unlikely case especially when dealing with biomarkers. Another approach commonly used is by incorporating the longitudinal time-dependent outcome into the Cox model (Rizopoulos, 2010). But, one needs to first determine whether the longitudinal covariate is an endogenous variable (also known as internal) or an exogenous variable (also known as external) (Rizopoulos, 2012). Endogenous variables are subject dependent i.e., their future existence is directly related to the failure status of the subject; are intermittently collected and contaminated with error. However, measurement errors are not the distinguishing characteristics between endogenous and exogenous variables. This is because an exogenous variables like air can also be error contaminated. The main distinguishing characteristic of the endogenous variables is that thy are intermittently collected i.e., their complete history is not available (Rizopoulos, 2012). This means that the levels of the biomarkers for a subject are only known for particular times that a subject visits a study site to provide data, and not in between the visit times. Endogenous variables include, for example, CD4 count for HIV-positive patients, the prothrombin index for patients with liver cirrhosis and serum creatinine level for patients with primary biliary cirrhosis (Rizopoulos, 2012).

Nonetheless, exogenous variables (e.g., air temperature) are not subject specific meaning that their values are not affected by the subject's failure status (Rizopoulos, 2012). Their values can be predicted too, meaning that their values at any time point is known infinitesimally before the same time point (Rizopoulos, 2012). In some instances the complete history of external measurements are predetermined from the beginning of the study.

Therefore extended Cox models or time-dependent Cox models are only reliable when dealing with exogenous time-dependent variables (Rizopoulos, 2012). This is because exogenous variables are neither intermittently collected nor subject dependent (Rizopoulos, 2012). The two approaches also lack the flexibility to incorporate measurement errors synonymous with longitudinal outcomes. As a result, parameter estimates obtained tend to be biased (Rizopoulos, 2012).

To address these limitations of the time-dependent Cox model and extended Cox model, Faucett and Thomas (1996); Tsiatis, Degruttola, and Wulfsohn (1995); Wulfsohn and Tsiatis (1997) proposed a framework in joint models for repeated measures and event time outcomes. In the last 20 years, joint models have grown in popularity (Alsefri, Sudell, García-Fiñana, & Kolamunnage-Dona, 2020). This popularity has been occasioned by the ability of joint models to give less biased results compared to other classical analysis techniques used in survival analysis (Rizopoulos, 2012). Joint models give less biased estimates of the parameters by accounting for the association between the repeated marker and the survival process (Hickey, Philipson, Jorgensen, & Kolamunnage-Dona, 2016).

The development of joint models started off as a simple LVCF (last value carried forward) method, followed by the two-stage method and finally to the shared random effects method (Henderson, Diggle, & Dobson, 2000; Lawrence Gould et al., 2015; Rizopoulos, 2012; Tsiatis & Davidian, 2004; Verbeke, Molenberghs, & Rizopoulos, 2010; Wulfsohn & Tsiatis, 1997). Tsiatis and Davidian (2004) have given an in-depth look into the original works on the approaches used in joint models for the repeated marker and survival processes, including those of Henderson et al. (2000); Wang and Taylor (2001); Wulfsohn and Tsiatis (1997), just to mention a few.

The assumption in the the basic joint models is that time to an endpoint of a particular interest is dependant on the current true level of the repeated marker at the same timepoint *t*. However, it will not always be true that this form of association structure will always be the most appropriate for determining the association between the two processes (Rizopoulos, 2012). We risk making incorrect inferences if we misspecify the parametrization structure between these two processes. To overcome this challenge, some alternative parametrization structures have been developed. They include: the time-dependent slopes parametrization ; the shared random-effects parametrization; cumulative effects parametrization; lagged effects parametrization and interaction effects parametrization (see for example Cekic, Aichele, Brandmaier, Köhncke, and Ghisletta (2019); He and Luo (2016); Lawrence Gould et al. (2015); Papageorgiou, Mauff, Tomer, and Rizopoulos (2019); Rizopoulos (2012)). Further, for highly nonlinear longitudinal profiles of subjects, it is appropriate to use splines, or higher order polynomials to describe the evolution patterns of individuals' profiles (Fitzmaurice et al., 2012; Rizopoulos, 2012; Verbeke, 1997).

To date, there have been myriad extensions of the standard joint models. Brown, Ibrahim, and DeGruttola (2005); R. Brown and G. Ibrahim (2003), for example broadened linear growth curve models to less restrictive non-parametric models for the repeated biomarkers using cubic B-splines. M. Yu, Law, Taylor, and Sandler (2004) examined a cure model in analysing survival data in the prostate cancer study. R. M. Elashoff, Li, and Li (2008); Huang, Dagne, and Wu (2011) expanded the Cox proportional hazards model to competing risks. Njagi et al. (2016) combined conjugate and normally distributed random effects of repeated measures and event time outcomes in joint models to improve on the model fitness. Baart, Boersma, and Rizopoulos (2019) extended the application of joint models in case-cohort study design. Some recent applications of joint models in HIV/AIDS include: Buta, Goshu, and Worku (2015); Dessiso and Goshu (2017); Erango, Goshu, Buta, and Dessisoa (2017); Mchunu et al. (2020); Seyoum and Temesgen (2017); Temesgen, Gurmesa, and Getchew (2018); T. Yu, Wu, and Gilbert (2019).

Recently, Rizopoulos has made good input in joint modeling framework, first by giving some detailed theoretical and practical overview of the joint modeling framework (Rizopoulos, 2012) and then proceeding to develop two packages in R for fitting the joint models: a Bayesian approach, JMbayes package (Rizopoulos, 2014) and a maximum likelihood approach, JM package (Rizopoulos, 2010).

The only challenging task with joint models is that it can be very computationally cumbersome to obtain maximum likelihood estimates of the parameters as the number of random effects increases (Bernhardt, Zhang, & Wang, 2015; Njagi et al., 2013; Rizopoulos, Verbeke, & Lesaffre, 2009; Rizopoulos, 2012).

3 Methodology

3.1 Introduction

Joint models is used primarily when the key interest is to explore the relationship of longitudinal and event time processes. The underlying idea is to connect the two processes using a common latent structure (Sène, Bellera, & Proust-Lima, 2013). The discussions on survival analysis, longitudinal analysis and joint models are in Sections 3.2, 3.3 and 3.4 respectively.

3.2 Survival Analysis

The goal of many longitudinal studies is to follow patients in the study until a prespecified endpoint is observed. Within the survival analysis, the key response or outcome variable is event time outcome also known as survival time or failure time (Rizopoulos, 2012). Models applied in survival analysis can simultaneously analyse multiple independent prognostic factors in addition to studying differences in treatments while controlling for the heterogeneity and imbalanced baseline factors (Rizopoulos, 2012). Survival analysis relies on the distribution of survival times, and two ways of illustrating the survival times are by either the survival function or the hazard function (Moore, 2016).

3.2.1 The Survival and Hazard Functions

Assuming T^* to be the true survival time, then we can write the survival function as:

$$\text{S(t)} = \text{Pr}(T^* > t) = 1 - \text{F(t)} = \int_t^{\infty} p(u)\,du, \qquad \textbf{(3.2.1)}$$
$$F(t) = p(T^* \leq t),$$

where $S(t)$ the probability of subject *i* surviving up to time-point *t*.

The hazard function, $h(t)$ is defined as the instantaneous rate for an event occurring in the time interval *[t, t + dt)* given the subject survives up to time *t* (Rizopoulos, 2012). We can formulate it as follows:

$$\begin{aligned}
\text{h(t)} &= \lim_{\delta t \to 0} \frac{\Pr\left(t \leq T^* < t + \delta t | T^* \geq t\right)}{\delta t}, t > 0 \\
&= \lim_{\delta t \to 0} \frac{\Pr\left(t \leq T^* < t + \delta t\right)}{\delta t \Pr(T^* > t)} \\
&= \lim_{\delta t \to 0} \frac{F\left(t + \delta t\right) - F(t)}{\delta t \Pr(T^* > t)} \\
&= \frac{f(t)}{S(t)}
\end{aligned} \tag{3.2.2}$$

Cumulative hazard function, H(t) is another major quantity in survival analysis. It describes the accumulated hazard up until time *t*. It can also be interpreted as the expected observed events by time t (Rizopoulos, 2012). It is written in the following:

$$\text{H(t)} = \int_0^t h(u)\, du \tag{3.2.3}$$

The rest of the functions can be derived from S(t), h(t) or H(t):

$$\begin{aligned}
\text{h(t)} &= -\frac{d}{dt}\text{log}\big(S(t)\big) \\
\text{H(t)} &= -\text{log}\big(S(t)\big) \\
\text{S(t)} &= \text{exp}\big(-H(t)\big) \\
&= \text{exp}\Big\{-\int_0^t h(u)du\Big\}
\end{aligned} \tag{3.2.4}$$

For Equation 3.2.1 or Equation 3.2.2 or any other characteristic of the survival time to hold, we must take into consideration non-informative right censoring (Rizopoulos, 2012). We begin by assuming that C_i is the censoring time for the i^{th} individual in the study. Also assume that δ_i is the event indicator expressed as:

$$\delta_i = \begin{cases} 1, & \text{event } (T_i^* \leq C_i) \\ 0, & \text{censored } (T_i^* > C_i) \end{cases}$$

it then follows that the observed time for the i^{th} individual is given by:

$$T_i = min(T_i^*, C_i).$$

3.2.2 Estimation of the Survival Function

To estimate the survival function, S(t), we use a Kaplan-Meier (K-M) estimate. K-M estimator is a non-parametric statistical method of assessing the number of events occurring at each particular time point. Let d_i represents the number of subjects with observed event time t_i for the i^{th} observation and r_i the number of subjects at risk before time t_i, then the K-M estimate is defined as:

$$\hat{S}(t) = \begin{cases} 1 & if \quad t_j < t_1 \\ \prod_{i=1}^{n}\left[1 - \frac{d_i}{r_i}\right] & if \quad t_i \leq t_j \leq t_n \end{cases} \tag{3.2.5}$$

3.2.3 Cox Proportional Hazards Model

It is a popular semi-parametric regression model used in characterising the association between survival times and baseline covariates (R. Elashoff, Li, et al., 2016). This model allows us to test if survival times between two or more groups are different after adjusting for other covariates. The model also assumes that baseline covariates have multiplicative effects on the risk of an event. Note that some books also refers Cox proportional hazards models as relative risk regression model or relative hazards model since it assumes a multiplicative effect of covariates on the hazard scale. For the i^{th} subject, we formulate the model as:

$$h_i(t|\mathbf{w_i}) = \mathbf{h_0}(\mathbf{t})\text{exp}\{\gamma^{\mathbf{T}}\mathbf{w_i}\}, \tag{3.2.6}$$

where $h_i(t)$ denotes hazard of an event for patient i at time t, $w_i^T = (w_{i1}, w_{i2}, \cdots, w_{ip})$ a set of baseline covariates, γ the parameter vector of the baseline coariates and $h_0(t)$ the unspecified baseline hazard.

We can further express $\text{exp}(\gamma^T w_i)$ as:

$$\lambda_i = \gamma_1 w_{i1} + \gamma_2 w_{i2} + \cdots + \gamma_p w_{ip},$$

where λ_i is the linear combination of the p covariates (Collett, 2015).

The general Cox proportional hazard model can then be expressed in log scale as:

$$\log\big(h_i(t|wi)\big) = \log\big(h_0(t)\big) + \gamma_1 w_{i1} + \gamma_2 w_{i2} + \cdots + \gamma_p w_{ip},$$

adjusted γ_j explains the magnitude of adjustments in the log hazard for a unit increase in $\mathbf{w}_j$ holding other covariates constant. Consequently, $\text{exp}(\gamma_j)$ indicates the hazards ratio at any time t for two groups of subjects. We can write it as follows:

$$\frac{h_i(t|w_i)}{h_k(t|w_k)} = \exp\left\{\gamma^T(w_i - w_k)\right\}$$

3.2.4 Estimating Parameters in Cox Proportional Hazards Model

Regression coefficients of interest, namely γ, are estimated through a *partial likelihood* function (R. Elashoff et al., 2016). The partial likelihood for γ according to Cox (1972) is defined as:

$$\mathrm{L}(\gamma) = \prod_{1=1}^{n} \frac{\exp(\gamma^T w_{(g)})}{\sum_{l \in R(t_{(g)})} \exp(\gamma^T w_i)}, \tag{3.2.7}$$

where distinct n ordered observed failure times are expressed as: $t_1 < t_2 < \cdots < t_n$; $w_{(g)}$ is the vector of the baseline covariates for subjects who fail at the g^{th} ordered event time. $R(t_{(g)})$ is the risk set and denotes a set of subjects who are at risk at time $t_{(g)}$.

Subjects who have failure times constitute the product in the likelihood function in Equation 3.2.7. Subjects who are at risk only contribute to the denominator of the likelihood function.

Assuming non-informative censoring, the likelihood function shown in Equation 3.2.7 can take the form:

$$\mathrm{L}(\gamma) = \prod_{1=1}^{n} \left\{ \frac{\exp(\gamma^T w_i)}{\sum_{l \in R(t_i)} \exp(\gamma^T w_i)} \right\}^{\delta_i}, \tag{3.2.8}$$

Equation 3.2.8 can further be expressed as:

$$\log \mathrm{L}(\gamma) = \sum_{i=1}^{n} \delta_i \left[\gamma^T w_i - \log\left\{ \sum_{T_j \geq T_i} \exp(\gamma^T w_j) \right\} \right] \tag{3.2.9}$$

MLE of the parameter γ can be computed by maximizing 3.2.9 using Newton-Raphson procedure (Collett, 2015).

3.3 Linear Mixed Effects Model

Some literature refers to it as random-effects model since it extends from the classical linear regression models by introducing the random effects terms in the model. LME is used in modelling longitudinal information.

A general LME model for the normal longitudinal responses y_i is expressed as:

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i, \tag{3.3.1}$$

where $y_i = (y_{i1}, y_{i2}, \cdots, y_{im})^T$ denotes a $m_i \times 1$ vector of longitudinal information for subject i, β denotes a $p \times 1$ vector of fixed parameters while X_i is a $m_i \times p$ known design matrix of explanatory variables. b_i is a $q \times 1$ vector of random effects that completes the characterization between subject variation. Further, Z_i is a $m_i \times q$ known design matrix corresponding to random effects b_i, with $q \leq p$, and lastly $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_i m_i)^T$ represents the $m_i \times 1$ vector of measurement or sampling errors that completes the characterization of within subject-variation. Further, let b_i be multivariate normally distributed with mean 0 and the covariance matrix D i.e., $b_i \sim N(0, D)$, (Rizopoulos, 2012). The sampling errors, $\varepsilon_i(t)$, are also assumed to be independent of b_i, normally distributed with mean 0 and variance $\sigma^2 I_{m_i}$ i.e., $\text{Cov}(b_i, \varepsilon_i) = 0$ and $\varepsilon_i \sim N(0, \sigma^2 I_{m_i})$.

Next, we let $\sum_i$ to represent the diagonal matrix, $\sigma^2 I_{m_i}$, with an $m_i \times m_i$ identity matrix, I_{m_i}, the covariance of y_i is then written as follows:

$$\begin{aligned}
\text{Cov}(y_i) \equiv V_i &= \text{Cov}(Z_i b_i) + \text{Cov}(\varepsilon_i) \\
&= Z_i \text{Cov} Z_i^T + \text{Cov}(\varepsilon_i) \\
&= Z_i D Z_i^T + \sigma^2 I_{m_i} \\
&= Z_i D Z_i^T + \Sigma_i
\end{aligned}$$

Implying that

$$y_i \sim N\left(X_i \beta, Z_i D Z_i^T + \Sigma_i\right)$$

3.3.1 Estimating LME Models

To estimate the parameters in the LME, we will apply the the principles of maximum likelihood. We let the marginal density of the observed outcome for the i^{th} subject to be expressed as:

$$p(y_i) = \int p(y_i|b_i)\ p(b_i)\ db_i.$$

If we assume independence across subjects, then the log-likelihood of the LME takes the form:

$$\begin{aligned}
\text{l}(\theta) &= \sum_{i=1}^{n} \log p(y_i; \theta) \\
&= \log \int p(y_i|b_i; \beta, \sigma^2) p(b_i; \theta_b)\ db_i,
\end{aligned} \tag{3.3.2}$$

where θ represents a full parameter vector decomposed into the sub vectors, $\theta^T = (\beta^T, \sigma^2, \theta_b^T)$, with $\theta_b = vech(D)$ and

$$p(y_i;\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log|V_i| - \frac{1}{2}\Big\{\sum_{i=1}^{n}(y_i - X_i\beta)^T V_i^{-1}(y_i - X_i\beta)\Big\} \quad \textbf{(3.3.3)}$$

$|V_i|$ is the determinant of square matrix V_i. Minimizing Equation 3.3.4 is similar to maximizing the log-likelihood with respect to β

$$\sum_{i=1}^{n}(y_i - X_i\beta)^T V_i^{-1}(y_i - X_i\beta) \quad \textbf{(3.3.4)}$$

We proceed with getting an estimator of β that minimizes 3.3.4. It takes the form:

$$\hat{\beta} = \Big\{\sum_{i=1}^{n}(X^T V_i^{-1} X_i)^{-1}\Big\}\sum_{i=1}^{n}(X^T V_i^{-1} y_i) \quad \textbf{(3.3.5)}$$

The formulation on equation 3.3.5 relies on the assumption that V_i, the covariance matrix, is known (Fitzmaurice et al., 2012). When V_i is known, then for any selection of V_i the GLS (generalise least squares) estimate of β is unbiased (Fitzmaurice et al., 2012): i.e.,

$$\mathrm{E}(\hat{\beta}) = \beta \quad \textbf{(3.3.6)}$$

The sampling distribution of $\hat{\beta}$ is multivariate normal with mean, β and covariance matrix, $\Big\{\sum_{i=1}^{n}(X^T \hat{V}_i^{-1} X_i)\Big\}^{-1}$ in substantially large samples (or asymptotically). The covariance matrix of $\hat{\beta}$ can be written as:

$$\mathrm{Cov}(\hat{\beta}) = \Big\{\sum_{i=1}^{n}(X^T \hat{V}_i^{-1} X_i)\Big\}^{-1} \quad \textbf{(3.3.7)}$$

However, we usually do not know V_i therefore we typically estimate V_i from the data at hand (Fitzmaurice et al., 2012). To obtain the ML estimate of V_i, we maximize the log-likelihood of $l(\theta_b, \sigma^2)$ for a given β value (Rizopoulos, 2012). Once we get the ML estimate of V_i, we then proceed by substituting the estimate into the generalised least squares (GLS) estimator of β given by 3.3.5 to obtain MLE of β:

$$\hat{\beta} = \Big\{\sum_{i=1}^{n}(X^T \hat{V}_i^{-1} X_i)^{-1}\Big\}\sum_{i=1}^{n}(X^T \hat{V}_i^{-1} y_i) \quad \textbf{(3.3.8)}$$

Interestingly in large samples (or asymptotically) , the estimator of β that substitutes the ML estimate of V_i has all the same properties as when V_i is actually known. That is equations 3.3.6 and 3.3.7 hold (Fitzmaurice et al., 2012). However we risk getting biased ML estimate of V_i under finite samples. Specifically, the diagonal elements of V_i are underestimated under ML estimate in finite samples (Fitzmaurice et al., 2012).

In order to address the problem of ML estimator in a general case of multivariate regression when estimating matrix V_i, we use restricted maximum likelihood (REML) estimation

(Fitzmaurice et al., 2012; Verbeke, 1997). In REML estimation of V_i, the likelihood does not contain β and is defined only in terms of V_i (Fitzmaurice et al., 2012; Rizopoulos, 2012). When we maximize the slightly modified log-likelihood function we get:

$$\begin{aligned} \mathrm{l}\left(\theta_b, \sigma^2\right) &= -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{n} X_i^T X_i\right| - \frac{1}{2}\log\left|\sum_{i=1}^{n} X_i^T V_i^{-1} X_i\right| \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}\left\{\log|V_i| + \left(y_i - X_i\hat{\beta}\right)^T V_i^{-1}\left(y_i - X_i\hat{\beta}\right)\right\} \\ &\propto -\frac{1}{2}\sum_{i=1}^{n}\log|V_i| - \frac{1}{2}\sum_{i=1}^{n}\left(y_i - X_i\hat{\beta}\right)^T V_i^{-1}\left(y_i - X_i\hat{\beta}\right) \\ &\quad -\frac{1}{2}\log\left|\sum_{i=1}^{n} X_i^T V_i^{-1} X_i\right| \end{aligned}$$

The estimate $\hat{V}_i$ obtained by maximization of the modified log-likelihood takes into account the fact that parameter β has also been estimated (Fitzmaurice et al., 2012). However, neither the MLE nor the REML estimator for the unique parameters in V_i has a closed form solution hence need for numerical optimization algorithm (Rizopoulos, 2012). The two frequently used numerical optimization algorithms are the E-M (Expectation-Maximization) algorithm and the Newton-Raphson algorithm (Rizopoulos, 2012).

Generally, the restricted maximum-likelihood estimator is less seriously biased than the maximum-likelihood estimator of V_i. It is therefore more appropriate to use REML for estimation of V_i (Fitzmaurice et al., 2012). However, the difference between ML estimation and REML estimation becomes less important when n is substantially larger than p i.e. sample size is substantially larger than the dimension of β (Fitzmaurice et al., 2012).

3.4 Joint Modelling

In joint modelling, the survival submodel is coupled with the longitudinal submodel through a shared latent structure (Njagi et al., 2013). In the next three sub-sections, we will be looking into the survival submodel, the longitudinal submodel and finally the joint model.

3.4.1 The Survival Submodel

For the i^{th} subject, we assume that T_i^* and C_i are the true event time and censoring time respectively. Also assuming that $\delta_i = I(T_i^* \leq C_i)$ is the event indicator and $T_i = min(T_i^*, C_i)$ the observed time, then the relative risk model can be expressed as:

$$h_i(t|\mathrm{M}_i(t), \mathbf{w}_i) = \mathrm{h}_0(t)\exp\left\{\gamma^T \mathbf{w}_i + \alpha \mathrm{m}_i(t)\right\},\ \mathrm{t>0}, \qquad \mathbf{(3.4.1)}$$

where $\mathrm{M}_i(t) = \left\{\mathrm{m}_i(\mathrm{s}), 0 \leq \mathrm{s} < \mathrm{t}\right\}$ denotes history of the true unobserved longitudinal value up to time t; $\mathrm{h}_0(\cdot)$ indicates the unspecified baseline hazard function and $\mathbf{w}_i$ is the baseline covariates with associated vector of regression coefficients γ. Finally α quantifies the association between the longitudinal marker and the risk of an event.

Equation 3.4.1 is the basic joint model described by the current value parametrization. The basic joint model can be extended to account for complex association structures. As highlighted in Chapter 2, one of these complex association structures can be captured through time-dependent slopes parametrization. The relative risk survival submodel for the time-dependent slopes parametrization then takes the form:

$$\begin{cases} h_i(t) & = h_0(t)\exp\left\{\gamma^T \mathbf{w}_i + \alpha_1 m_i(t) + \alpha_2 m_i^{'}(t)\right\}, \\ m_i^{'}(t) & = \frac{d}{dt} m_i(t). \end{cases} \tag{3.4.2}$$

Under time-dependent slopes parametrization, we allow hazard to depend on both the current true level of the marker, $m_i(t)$ at time point t and its true rate of change, $m_i^{'}(t)$ at the same time point t. Time-dependent slopes parametrization is very useful for distinguishing between subjects who have, at a specific time point, same level of the marker but who differ in the slopes (e.g., the rate of change in the marker of one subject at a particular time point could be increasing while for the other subject it could be decreasing at the same time point).

In joint modelling framework, it is advisable to use parametric but flexible models for the baseline hazard. Leaving it unspecified could lead to the standard errors of the parameters being underestimated. More often, piecewise-constant and regression splines options work efficiently. The piecewise-constant model baseline risk function is written as (Rizopoulos, 2012):

$$h_0(t) = \sum_{q=1}^{Q} \xi_q I(v_{q-1} < t \leq v_q), \tag{3.4.3}$$

where $0 = v_0 < v_1 < \cdots < v_q$ represents a split of the time scale.

Under regression spline model, the log baseline risk function ($\log h_0(t)$) is extended into B-spline basis functions for cubic spline, i.e., (Rizopoulos, 2012):

$$\log h_0(t) = \gamma h_0, 0 + \sum_{q=1}^{Q} \gamma h_{0,q} B_q(t, v), \tag{3.4.4}$$

where $B_q(t, v)$ denotes the q-th basis function of the spline with knots $v_1, \cdots, v_Q$ and γh_0 a vector of spline coefficients. More literature on the the choice and usage of these these two baseline risk functions can be found in Rizopoulos (2012). In the present study we have employed a piece-wise constant baseline risk function.

3.4.2 The Longitudinal Submodel

Since longitudinal measurements are error contaminated and collected intermittently at some few time points t_{ij}, we need to estimate the true unobserved measurement $m_i(t)$. This is done by applying a suitable model to describe the evolution of the marker over time for each subject. Assuming that longitudinal measurements are normally distributed, the model takes the form:

$$\begin{cases} y_i(t) = m_i(t) + \varepsilon_i(t) = \mathbf{x}_i^T(t)\beta + \mathbf{z}_i^T(t)\mathbf{b}_i + \varepsilon_i(t), \\ b_i \sim N(0, D) \\ \varepsilon_i(t) \sim N(0, \sigma^2) \\ \varepsilon_i(t) \text{ is independent of } b_i \end{cases} \tag{3.4.5}$$

with $\mathbf{x}_i^T(t)$ the design matrix vector for the fixed effects β while $\mathbf{z}_i^T(t)$ the design matrix vector for the random effects $\mathbf{b}_i$. β is the parameter vector, $\varepsilon_i(t) \sim N(0, \sigma^2)$. The measurement error, $\varepsilon_i(t)$, is assumed independent of the random effects b_i, with $b_i \sim N(0, D)$.

Since survival function, $S_i(t)$ depends on the whole history of the biomarker, it is important to correctly specify the time structure of the time in $\mathbf{x}_i(t)$ and $\mathbf{z}_i(t)$ and possibly the interaction terms between the time structure and the baseline covariates. If subjects exhibit highly non-linear longitudinal profiles, then consider using high-order polynomials or splines.

3.4.3 Estimation of Joint Models

To estimate parameters in the joint models, these three methods can be applied: (Wu, Liu, Yi, & Huang, 2012):

1. Two-stage methods
2. Bayesian Markov Chain Monte Carlo (MCMC) method
3. Likelihood methods

However, our current study strictly applied the likelihood methods to estimate parameters.

Two-stage methods

A two-stage method is designed as follows:

1. Step I: The estimates of the parameter and predictions are calculated for longitudinal outcome without taking into account survival outcome.
2. Step II: The survival model is fitted by utilizing the predicted longitudinal values as true covariates.

Whereas the two-stage approach is pretty easy to be implemented with the existing software, it often produces biased results (Wu et al., 2012) as shown in some simulation studies like Dafni and Tsiatis (1998); Sweeting and Thompson (2011); Tsiatis and Davidian (2001). Following Wu et al. (2012), this approach leads to unbiased results because it does not utilize information from the longitudinal and the survival process simultaneously in each model fitting step.

Bayesian Markov Chain Monte Carlo (MCMC) method

Alternatively, a Bayesian method incorporates both types of outcomes and simultaneously estimates model parameters (Yang, Yu, & Gao, 2016). This method employs the principle of Markov chain Monte Carlo (MCMC) sampling algorithms (Alsefri et al., 2020; Yang et al., 2016). Alsefri et al. (2020) gave a comprehensive review on Bayesian MCMC method under both univariate and multivariate joint models. Bayesian approach is more flexible when it comes to parameter estimation. Also, estimates from Bayesian approach are less biased as it utilizes related historical information (Alsefri et al., 2020). However, overdependence on the prior specification can sometimes lead to invalid estimates (Yang et al., 2016). Furthermore, autocorrelation and convergence can be a problem when handling many parameters in complex models. Bayesian framework has recently been used by Baart et al. (2019); Mauff, Steyerberg, Kardys, Boersma, and Rizopoulos (2020) in their work.

Likelihood methods

Semi-parametric maximum likelihood method proposed by (Henderson et al., 2000; Hsieh, Tseng, & Wang, 2006; Wulfsohn & Tsiatis, 1997) is the popular estimation method for joint models (Rizopoulos, 2012). Within the likelihood method, we assume full conditional independence. This implies that the random effects represented by $\mathbf{b}_i$ takes into account both the association between the longitudinal and time-to-event outcomes, and the correlation between the repeated measures in the longitudinal process. We can illustrate the joint distribution as follows:

$$p(T_i, \delta_i, y_i | b_i; \theta) = p(T_i, \delta_i | b_i; \theta) p(y_i | b_i; \theta) \tag{3.4.6}$$

$$p(y_i | b_i; \theta) = \prod_j p\{y_i(t_{ij}) | b_i; \theta\} \tag{3.4.7}$$

Further, assuming non-informative censoring and visiting processes (censoring, timing, and measurement processes depend only on the observed history and latent random effects and not on the future risk time itself), the log-likelihood contribution for the i^{th} subject is formulated as:

$$\begin{aligned} \log p(T_i, \delta_i, y_i; \theta) &= \log \int p(T_i, \delta_i, y_i, b_i; \theta)\, db_i \\ &= \log \int p(T_i, \delta_i | b_i; \theta) \Big[\prod_j p\{y_i(t_{ij}) | b_i; \theta\} \Big] p(b_i; \theta)\, db_i, \end{aligned} \tag{3.4.8}$$

with θ the parameter vector, y_i the longitudinal information of the i^{th} subject, δ_i the event indicator, and

$$\begin{aligned} p(T_i, \delta_i | b_i; \theta) &= h_i(T_i | \mathrm{M}_i(T_i); \theta)^{\delta_i} S_i(T_i | \mathrm{M}_i(T_i); \theta) \\ &= \Big[h_0(T_i) \exp\{\gamma^T \mathbf{w}_i + \alpha m_i(T_i)\} \Big]^{\delta_i} \exp\left(- \int_0^{T_i} \Big[h_0(t) \exp\{\gamma^T \mathbf{w}_i + \alpha m_i(t)\} \Big]\, dt \right). \end{aligned} \tag{3.4.9}$$

The joint density for the longitudinal outcome together with the random effects is of the form:

$$\begin{aligned} p(y_i | b_i; \theta) p(b_i; \theta) \quad &= \quad \prod_j p\{y_i(t_{ij}) | b_i; \theta\} p(b_i; \theta) \\ & \qquad = (2\pi\sigma^2)^{-n_i/2} \exp\Big\{ - \frac{\|yi - X_i\beta - Z_i b_i\|^2}{2\sigma^2} \Big\} \\ & \qquad \times \quad (2\pi)^{-q_b/2} \det(\mathrm{D})^{-1/2} \exp\left(-b_i^T D^{-1} b_i / 2 \right), \end{aligned}$$

where q_b is the dimensionality of the random-effects vector, and $\|x\| = \{\Sigma_i x_i^2\}^{1/2}$ indicates the Euclidean vector norm.

Maximizing the log-likelihood function $\mathrm{l}(\theta) = \sum_i \log \mathrm{p}(T_i, \delta_i, y_i; \theta)$ with respect to θ can be achieved using either Expectation-Maximization (E-M) algorithm or the Newton-Raphson (Rizopoulos, 2012). However, E-M algorithm has been traditionally advocated for in the joint modelling literature (treating $\mathbf{b}_i$ as missing data), mainly due to the closed-form solution of the parameters (Rizopoulos, 2012). Unfortunately, the main weakness of the E-M algorithm is slow convergence at the maximum (Rizopoulos, 2012).

Implementation of Expectation-Maximization Algorithm

Two steps exist in this process: the Expectation-step also known as the E-step and the Maximization-step also known as the M-step (Rizopoulos, 2012). In the Expectation-step, we fill the missing data using the observed data and current parameters estimates by conditional expectation; in the Maximization-step, we maximize the conditional expectation from step one. The observed log-likelihood function can be expressed as:

$$\log(\theta) = \sum_{i=1}^{n} \{\log p(T_i, \delta_i | b_i; \theta, \beta) + \log p(yi | b_i; \theta) + \log p(b_i; \theta)\}$$

In the E-step the expected value of the complete log-likelihood function given the conditional distribution of $\mathbf{b}_i$ is:

$$\mathrm{Q}(\theta | \theta^m) = \sum_{i=1}^{n} \int \{\log p(T_i, \delta_i | b_i; \theta, \beta) + \log p(yi | b_i; \theta) + \log p(b_i; \theta)\} p(b_i | T_i, \delta_i, y_i; \theta^m) \, db_i,$$

The two integrals in $\mathrm{Q}(\theta|\theta^m)$ need to be solved numerically using Gaussian quadrature rules or Monte Carlo sampling.

For the M-step, closed-form solutions can be obtained for the variance and covariance matrix of residuals and random effects respectively. However, the fixed effects for every longitudinal model and parameters in the time-to-event model has to be solved numerically. The main steps for these parameters are as follows:

Step 1: Estimate the variance of residuals of each longitudinal model by closed form expressions:

$$\begin{aligned}\hat{\sigma^2} &= \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} (y_i - X_i \beta^m)^T \left(y_i - X_i^T \beta^m - 2 Z_i^T \mathrm{E}(b_i | T_i, \delta_i, y_i; \theta^m)\right) \\ &\quad + \mathrm{tr}(Z_i^T Z_i \mathrm{Var}(b_i | T_i, \delta_i, y_i; \theta^m)) + \\ &\quad \mathrm{E}(b_i | T_i, \delta_i, y_i; \theta^m)^T Z_i^T Z_i \mathrm{E}(b_i | T_i, \delta_i, y_i; \theta^m)\end{aligned}$$

where tr is the trace function of a matrix, and E, the expectation function.

Step 2: Estimation of variance-covariance matrix of random effects b_i by

$$\begin{aligned}\hat{D} &= \frac{1}{n}\sum_{i}^{n}\mathsf{Var}(b_i|T_i,\delta_i,y_i;\theta^m)\\ &\mathsf{E}(b_i|T_i,\delta_i,y_i;\theta^m)\mathsf{E}(b_i|T_i,\delta_i,y_i;\theta^m)^T\end{aligned}$$

Step 3: Since the parameters of the event time model (θ) and score equations for the fixed effect coefficient (β) lack closed form solutions, we proceed and implement the one-step Newton-Raphson algorithm for these parameters:

$$\begin{aligned}\hat{\beta}^{m+1} &= \hat{\beta}^{m}-\left(\frac{\partial S(\hat{\beta}^{m})}{\partial\beta}\right)^{-1}S(\hat{\beta}^{m}),\\ \hat{\theta}^{m+1} &= \hat{\theta}^{m}-\left(\frac{\partial S(\hat{\theta}^{m})}{\partial\theta}\right)^{-1}S(\hat{\theta}^{m}),\end{aligned}$$

where $\hat{\beta}^m$ and $\hat{\theta}^m$ are values of β and θ at the present iteration, and $\frac{\partial S(\hat{\beta}^m)}{\partial\beta}$ and $\frac{\partial S(\hat{\theta}^m)}{\partial\theta}$ are the corresponding blocks of the Hessian matrix (H-matrix) calculated at $\hat{\beta}^m$ and $\hat{\theta}^m$. The score vector's components then take the form:

$$\begin{aligned}S(\beta) &= \sum_{i=1}^{n}\frac{1}{\sigma^2}X_i^T\left(y_i-X_i^T\beta-Z_i^T\mathsf{E}(b_i|T_i,\delta_i,y_i;\theta^m)\right)+\delta_i\alpha x_i(T_i)\\ &-\exp(\gamma^T w_i)\int\int_0^{T_i}h_0(u)\alpha x_i(s)\exp\left[\alpha\{x_i^T(s)\beta+z_i^T(s)b_i\}\right]\\ &\times p(b_i|T_i,\delta_i,y_i;\theta)\,du\,db_i,\end{aligned}$$

Step 4: In this step, the survival model parameters could be updated too using the Newton-Raphson algorithm. Similarly the baseline hazard function is estimated non-parametrically using piecewise-constant function. Score equations used in the Newton-Raphson algorithm are:

$$
\begin{aligned}
S(\gamma) &= \sum_{i=1}^{n} w_i \Big[\delta_i - \exp(\gamma^T w_i) \int \int_0^{T_i} h_0(u) \exp\big[\alpha\{x_i^T(s)\beta + z_i^T(s)b_i\}\big] \\
&\quad \times p(b_i|T_i, \delta_i, y_i; \theta)\, du\, db_i \Big],
\end{aligned}
$$

$$
\begin{aligned}
S(\alpha) &= \sum_{i=1}^{n} \delta_i \{(x_i^T(T_i)\beta + z_i^T(T_i)\mathsf{E}(b_i|T_i, \delta_i, y_i; \theta^m)\} \\
&\quad - \exp(\gamma^T w_i) \int \int_0^{T_i} h_0(u) \exp\big[\alpha\{x_i^T(s)\beta + z_i^T(s)b_i\}\big] \\
&\quad \times p(b_i|T_i, \delta_i, y_i; \theta)\, du\, db_i,
\end{aligned}
$$

$$
\begin{aligned}
S(\theta_{h_0}) &= \sum_{i=1}^{n} \delta_i \frac{\partial h_0(s; \theta_{h_0})}{\partial \theta_{h_0}^T} \\
&\quad - \exp(\gamma^T w_i) \int \int_0^{T_i} \frac{\partial h_0(T_i; \theta_{h_0})}{\partial \theta_{h_0}^T} \exp\big[\alpha\{x_i^T(s)\beta + z_i^T(s)b_i\}\big] \\
&\quad \times p(b_i|T_i, \delta_i, y_i; \theta)\, du\, db_i.
\end{aligned}
$$

The corresponding blocks of the H-matrix for $\frac{\partial S(\hat{\beta})}{\partial \beta}$ and $\frac{\partial S(\hat{\theta})}{\partial \theta}$ can be calculated by central difference approximation.

To get a solution of the expected likelihood function, a pseudo-adaptive Gaussian-Hermit quadrature rule can be utilized in approximating the integrals Rizopoulos (2012). The Expectation-step and the Maximization-step iterate to a pre-specified convergence criterion.

4 Results

4.1 Introduction

The sections in this chapter are divided as follows: Section 4.2 gives a description of the data used in the study. In Section 4.3, we present the results on the demographic, clinical and behavioral characteristics of patients enrolled in the study. Section 4.4 are results for longitudinal analysis of CD4 count using LME model. Section 4.5 are results for time to wound healing using Cox proportional hazards model. In Section 4.6, we give results on joint modelling of the two processes. We used the JM package (Rizopoulos, 2010) to fit the joint models. Other analyses were also done using R software, version 3.5.2.

4.2 Data description and study population

These data were from a prospective cohort study conducted in Kisumu by Nyanza Reproductive Health Society (NRHS). The study's goals were:- ***(i)***- *to determine complete wound healing time.* ***(ii)***- *document any adverse events as a result of Prepex circumcision.* ***(iii)***- *to study evolution of CD4 counts, viral loads and viral shedding in time following Prepex circumcision.* Each of the 119 patients was planned for a return visit after every 7 days post-circumcision until wound was certified fully healed.

4.3 Demographic, clinical and behavioral characteristics

The mean age of patients in the Prepex study was 35.8 years (95% CI: 34.48 , 37.14) and about three quarter (73.11%) of them reported to be on antiretroviral treatment (ART). The mean CD4 count at baseline visit was 482.3 cells/μL (95% CI: 437.98-526.63) and the median was 437 cells/μL (IQR:298-596). 79.99% of the patients were married and only 2.52% reported not having education. Majority of patients (93.28%) reported having allergy. 82.35% of patients reported to have had sex in the 6 months preceding circumcision out of which 69.39% (68/98) used condoms at their last sexual intercourse. In addition, 8.51% (8/94) of patients had 2 or more sexual partners in the 6 months preceding circumcision. Finally, 15 (12.61%) patients reported consistent alcohol use for 3 or more days before circumcision. The baseline characteristics of patients enrolled into the study are as shown in Table 1 and Table 2.

4.4 Longitudinal Analysis: modelling CD4 count

CD4 count for all patients was measured at baseline visit (circumcision visit) and at subsequent common weekly follow-up visits until wound was certified fully healed. If values on CD4 count were missing for any follow-up visit, then it was considered as missing at random **(MAR)**. Hence, our analysis was based on the available data and the results are still valid under **MAR**.

Since our longitudinal analysis was done using LME for continuous outcome variable, there was need therefore for CD4 count values to be normally distributed. We checked for normality of CD4 count values by plotting histograms and overlaid normal curves for the original CD4 count values, square root transformed values and log transformed values. CD4 count values generated from square root function exhibited normal distribution compared to either log transformed values or original values. We therefore considered square root CD4 counts in our subsequent analyses. Figure 1, Figure 2 and Figure 3 show histogram results for the original CD4 count values, square root transformed values and log transformed values respectively.

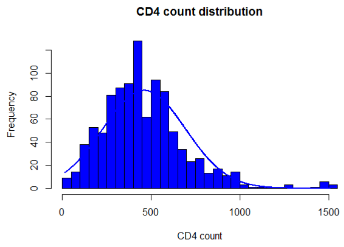


Figure 1. Histogram for CD4 count (cells/μL)

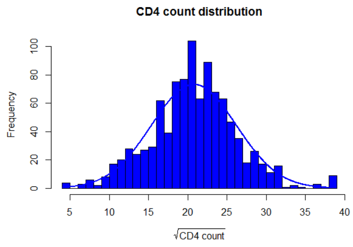


Figure 2. Histogram for square root CD4 count (cells/μL)

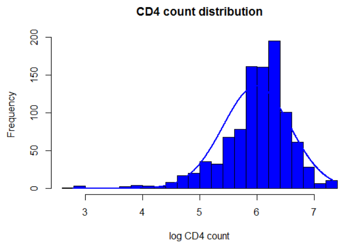


Figure 3. Histogram for Log CD4 count (cells/μL)

4.4.1 Exploring individual profiles and the mean structure

As illustrated in Figure 4, the individual profile plots for 15 randomly selected patients show that there is heterogeneity at baseline among patients enrolled in the study with respect to the square root CD4 count. Further, there is variability within and between patients over time with respect to square root CD4 count. Therefore, there is need to consider a suitable random effects structure in our analysis of square root CD4 count.

The average evolution is important as it describes how the profile for subjects being studied evolves over time and the results of the exploration helps in determining the best choice for a fixed-effects structure for the LME model (Fitzmaurice et al., 2012; Verbeke, 1997). Figure 5 shows the average profile plot with respect to square root CD4 count over time. From the average profile plot, we see that the square root CD4 count presents a quadratic trajectory over time in weeks. This is further tested in Section 4.4.2 whether or not we need the quadratic time effect in our model.

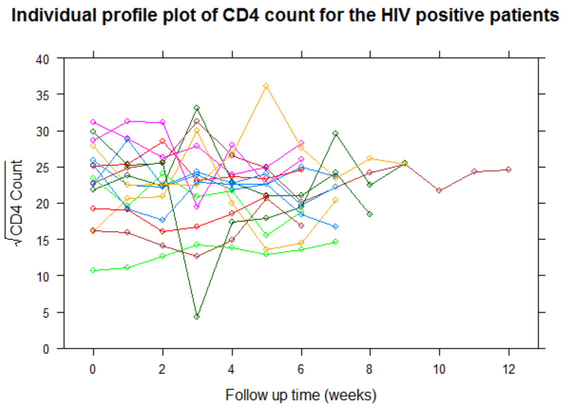


Figure 4. Individual profile plots (n=15)

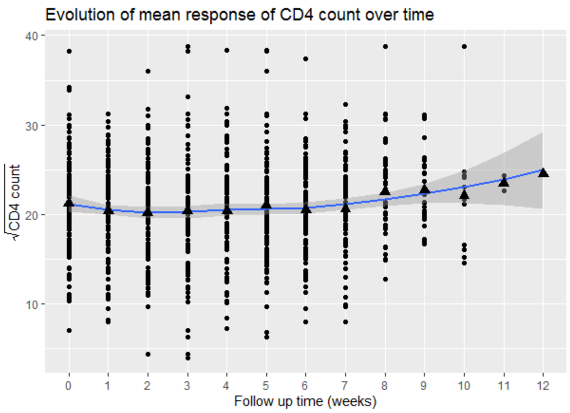


Figure 5. Mean structure

4.4.2 Tests for need for random effects

Random effects b_i denotes the variability in subject-specific intercepts and slopes, not explained by the explanatory variables in the model (Verbeke, 1997). It is important therefore to conduct a test on whether or not we should retain the random effects in our model. In this study, we conducted a hypothesis test in a hierarchial way by dropping one random effect from the model at a time starting from the highest-order time effect (Drikvandi, Verbeke, Khodadadi, & Partovi Nia, 2013; Verbeke, 1997). The results of the maximized log-likelihood is given in Table 3.

The main interest is to test for the need for the random slope for the quadratic time effect in the model as exhibited in Figure 5. The LR statistic for testing 2 versus 3 random effects is a mixture of chi-square test with equal weights 0.5 (χ^2_2 and χ^2_3) (Drikvandi et al., 2013; Verbeke, 1997). The p-value under REML of the LR test is smaller than 0.0001 and therefore the random slope for the quadratic time effect is retained in the model. Table 4 shows the likelihood ratio statistics for comparing 2 versus 3 random-effects.

4.4.3 Linear Mixed Model Results

The most satisfactory LME model to use in describing the average change in square root CD4 count over time was therefore one with the random intercept, linear and quadratic time effects according to the results obtained in Section 4.4.2. The next step was to fit a saturated model with all the explanatory variables of interest. Table 5 gives us the results of the saturated LME model.

To improve the overall fitness of the model, we used a step-wise method. First, we dropped the most insignificant terms in the complex (saturated) model and sequentially selected the term for which its removal had the least damaging effect on the model and the process terminated when any further elimination led to a poorer fit.

The AIC (Akaike information criterion) decreased from 5979.97 to 5959.33 after dropping all the statistically insignificant terms, which implies a better model fit. Table 6 illustrates the results of the final model fitted under REML estimation method with all the terms being statistically significant (p-value<0.05). The average square root CD4 count at baseline for those without allergy was 20.69. For those with allergy, it was more by 3.74.

Figure 6 shows the results of the estimated square root CD4 count over time for patients in each allergy arm. Patients with allergy (coded as 1) had consistently higher estimated square root CD4 count than patients without allergy (coded as 0).

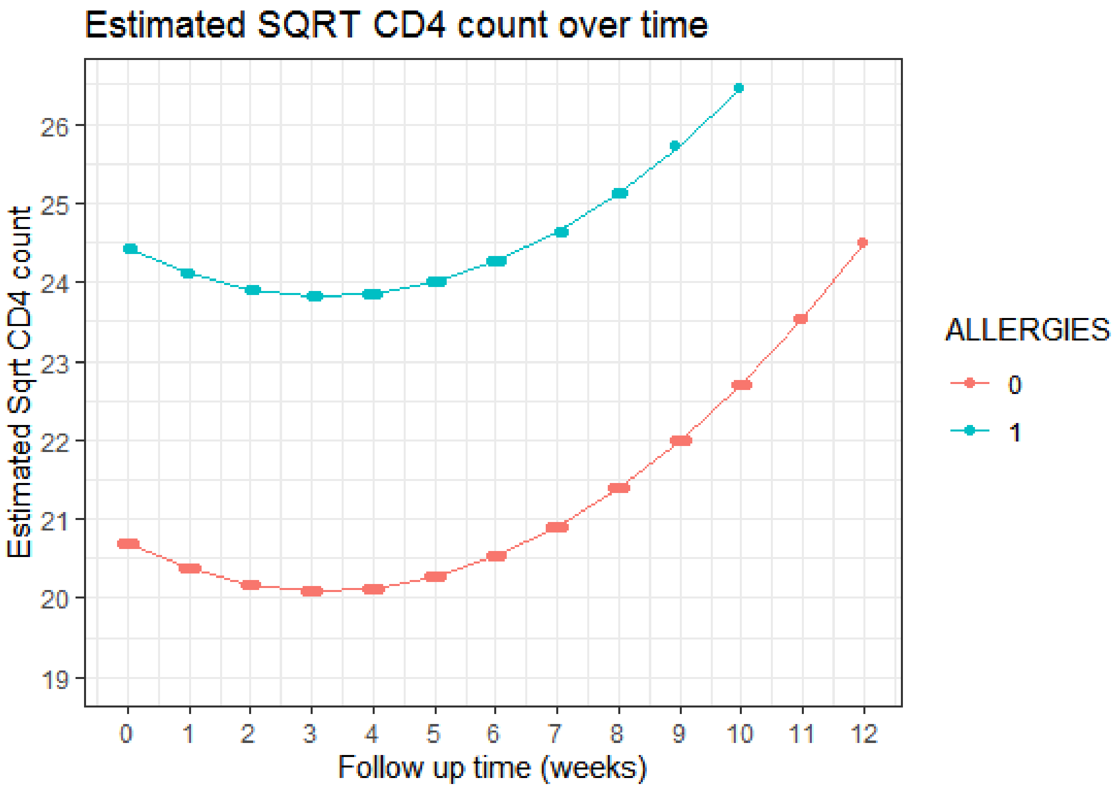


Figure 6. Estimated square root CD4 count over time

4.5 Survival analysis: modelling time to wound healing

Out of the 119 patients enrolled into the study, 115 (96.64%) were certified fully healed during follow up visit while only 4 (3.36%) were lost to follow up. The overall median complete wound healing time was 49 days (IQR:49-63). Figure 7 is a Kaplan-Meir plot for time to wound healing.

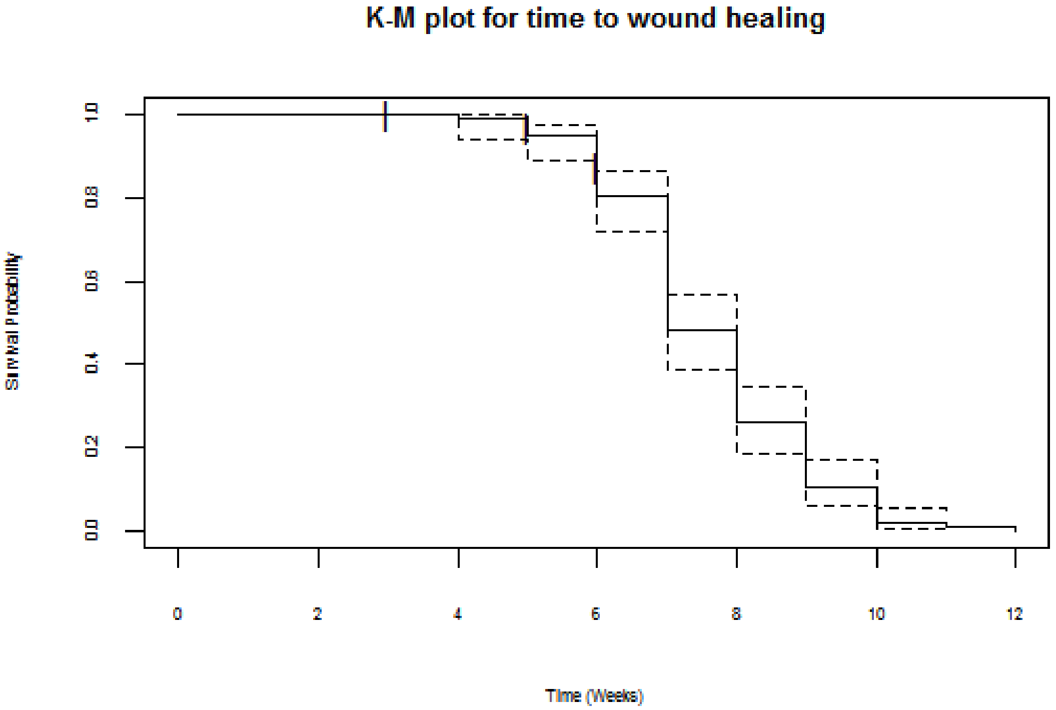


Figure 7. K-M plot for time to wound healing

To determine effects of demographic, behavioral and clinical characteristics on the risk for wound healing, we applied Cox proportional hazards models. We presented hazard ratios (HR) alongside their corresponding 95% CI (confidence intervals). The proportional hazards assumption was checked graphically besides doing a formal test on the Schoenfeld residuals (Collett, 2015; Moore, 2016).

To asses proportional hazards assumption graphically, we first plotted log(-log(survival function)) against the log of time for different priori selected covariates and the plots checked for parallelism (Moore, 2016). Figure 8, clearly shows that the log-log plots for marital status, antiretroviral treatment (ART) status, allergy and sex in the past 6 months cross more than once. Indeed, there is a problem deciding “how parallel is parallel using this approach (Moore, 2016). Also, we lack a clear way to assess statistical significance (Collett, 2015). There is need therefore to conduct a formal statistical test to aid in good judgement.

Our next step was to plot Schoenfeld residuals against time by baseline covariates to examine the assumption of proportionality of the hazards. Figure 9 are plots for Schoenfeld residuals against time. The dashed lines represent ± 2 standard-error bands around the smoothing spline fits to the plots. There is no evidence of pattern of Schoenfeld residuals with time. Consequently, the assumption of proportional hazards holds in all the baseline covariates.

We further validated our results by testing for the correlation between the Schoenfeld residuals and time-to-event. A correlation value of 0 implies that the model has met the proportional hazards assumption. Otherwise, the assumption has been violated. We utilized the function "cox.zph" from the R survival package to conduct a test for each of the baseline covariates individually and globally (model with all the covariates) (Moore, 2016; Cekic et al., 2019). The proportional hazards assumption is supported by a statistically insignificant relationship between residuals and time, and violated by a significant relationship. The results in Table 7 shows no enough proof to suggest that the assumption has been violated by any of the covariates.

Since no covariate violated the proportional hazards assumption, we proceeded to the next step of fitting Cox proportional hazards models with the baseline covariates.

First, we assessed each of the priori selected explanatory variables on their own using an un-adjusted Cox proportional hazards model to establish the reduction in -2 log likelihoods when compared to a null model. Baseline covariates significantly reducing the model deviance (p-value$<$0.05) were then included in the adjusted Cox proportional hazards model. Explanatory variables that failed to meet this set condition were dropped from the model. Only marital status was at the borderline statistical significance (p-value=0.046) while the rest of the baseline covariates were insignificantly associated with wound healing. The results are illustrated in Table 8 and Table 9.

4.6 Joint model results

We utilized the piecewise constant baseline risk function in our joint modelling framework (Rizopoulos, 2010, 2012). Since marital status was at borderline statistical significance, we never considered it in our joint modelling framework. This is mainly because of the convergence issues we were getting while running our joint models. We therefore considered a survival submodel without any of the baseline covariates in our joint modelling framework.

4.6.1 The current value parametrization

Table 10 shows the results of the joint models under current value and time-dependent slopes parametrization. The estimate of α is 0.028. This can further be expressed in hazard ratio as: exp(0.028)=1.028. Implying, a 2.8% higher hazard of wound healing at any time for a unit increase in the current true value of square root CD4 count at the same time point. However, the relationship between square root CD4 count and the hazard of wound healing was statistically insignificant at 0.05 level of significance, (p-value=0.536).

Hence, the hazard of wound healing at a given time point was not significantly associated with the true level of square root CD4 count at that time point.

4.6.2 The time-dependent slopes parametrization

The rate of change in square root CD4 count at a given time point, *t* was a much more important predictor of hazard of wound healing than the current true value of square root CD4 count at that same time point, *t*. Specifically, for patients with the same current true level of square root CD4 count, the log hazard ratio for a unit increase in the current slope of square root CD4 count trajectory was 1.514 (95% CI: 1.121; 1.908). Still under time-dependent slopes association, the association between the current true level of square root CD4 count and the hazard of wound healing was statistically insignificant, -0.08 (95% CI: -0.197; 0.029). Results are shown in Table 10.

Table 1. Demographic, clinical and behavioral characteristics

Baseline characteristics		n (%)
Ethnicity		
	Luo	119 (100.00)
Marital Status		
	Married	94 (78.99)
	Not married	25 (21.02)
Highest education level		
	None	3 (2.52)
	Primary	55(46.22)
	Secondary	49 (41.18)
	Post-secondary	12 (10.08)
ART at baseline		
	No	32 (26.89)
	Yes	87 (73.11)
Reported allergy at baseline		
	No	111 (93.28)
	Yes	8 (6.72)
Past 6 months sexual intercourse		
	No	21 (17.65)
	Yes	98 (82.35)
Past 6 months sexual partners		
	None	21 (17.65)
	One	87 (73.11)
	Two	9 (7.56)
	Three or more	2 (1.68)
Ever used a condom		
	No	18 (15.13)
	Yes	76(63.87)
	Missing	25(21.01)
Condom use at last sex		
	No	28 (23.53)
	Yes	66 (55.46)
	Missing	25(21.01)
Alcohol consumption (days per week)		
	None	71 (59.66)
	Less than or one	25 (21.01)
	One to two	10 (8.40)
	Three or more	15 (12.61)

Table 2. Descriptive statistics

Demographic/ Clinical characteristic	Mean	95% CI		Median	IQR	
		Lower Bound	Upper Bound		25%	75%
Age	35.81	34.48	37.14	36	30	42
Weight	62.22	60.52	63.91	61	56	65.7
CD4 count	482.30	437.98	526.63	437	298	596

Table 3. LME models with the associated log-likelihood value

Random effects	ML	REML
M1: 3 random effects	−2954.28	−2952.99
M2: 2 random effects	−2968.87	−2970.60

*3 random effects constitute random intercept; linear and quadratic slope

*2 random effects constitute random intercept and linear slope

Table 4. Mixture of chi-square test for comparing random-effects models

Hypothesis	$-2\ln(\lambda_N)$	Asymptotic null distribution	p-value
*M2 versus M1	35.22	$\chi^2_{2:3}$	<0.0001

*REML estimation

Table 5. Saturated LME model

Effect	Estimate	SE	p-value
Intercept	17.713	4.768	0.000
Time	−0.543	8.257	0.948
Time^2	0.297	4.103	0.942
Age	0.036	0.069	0.598
ART-Yes	1.420	1.087	0.194
Weight	0.011	0.050	0.824
Married-Yes	−0.194	1.299	0.882
Allergies-Yes	2.714	1.901	0.156
Past 6 months sexual intercourse-Yes	−0.757	1.377	0.584
Frequent alcohol use-Yes	0.599	0.985	0.544
Highest education level-Primary	1.122	3.039	0.713
Highest education level-Post-Primary	−0.120	3.084	0.969
Time:Age	−0.069	0.121	0.565
Time^2:Age	0.020	0.060	0.743
Time:ART-Yes	−0.228	1.953	0.907
Time^2:ART-Yes	−0.478	0.997	0.632
Time:Weight	0.014	0.090	0.879
Time^2:Weight	−0.004	0.046	0.924
Time:Married-Yes	3.856	2.350	0.101
Time^2:Married-Yes	−1.759	1.207	0.145
Time:Allergies-Yes	0.871	3.307	0.792
Time^2:Allergies-Yes	−0.059	1.644	0.971
Time:Past 6 months sexual intercourse-Yes	−5.112	2.524	0.043
Time^2:Past 6 months sexual intercourse-Yes	3.238	1.308	0.014
Time:Frequent alcohol use-Yes	0.797	1.718	0.643
Time^2:Frequent alcohol use-Yes	−0.503	0.858	0.558
Time:Highest education level-Primary	0.762	4.978	0.878
Time:Highest education level-Post-Primary	1.698	5.032	0.736
Time^2:Highest education level-Primary	0.010	2.358	0.997
Time^2:Highest education level-Post-Primary	−0.589	2.376	0.804

Table 6. Final LME model

Effect	Estimate	SE	p-value
Intercept	20.689	0.493	0.000
Time	−1.493	0.671	0.026
Time^2	0.920	0.300	0.002
Allergies-Yes	3.740	1.169	0.002

*Note: Model reduction is based on step-wise procedure

Table 7. An investigation to the proportional hazards assumption

Effect	rho	chi-square	p-value
Age	0.140	3.790	0.052
ART	−0.040	0.098	0.754
Weight	0.011	0.374	0.541
Married	0.031	1.405	0.236
Allergies	0.074	0.644	0.422
Past 6 months sexual intercourse	0.025	0.474	0.491
Frequent alcohol consumption	−0.088	0.427	0.513
Highest education level	−0.073	0.645	0.724
Global	−0.071	6.479	0.691

Table 8. Model selection

Effect	-2 log $\hat{L}$	p-value
null	868.903	
Age	866.988	0.166
ART	867.707	0.274
Weight	868.890	0.909
Married	864.612	0.046
Allergies	868.899	0.949
Past 6 months sexual intercourse	866.705	0.138
Frequent alcohol use	868.489	0.520
Highest education level	866.736	0.141

Table 9. An unadjusted Cox proportional hazards model

Effect	Estimate	HR (95% CI)
Age	−0.018	0.982 (0.958-1.007)
ART-Yes	−0.239	0.787 (0.517-1.199)
Weight	0.001	1.001 (0.980-1.023)
Married-Yes	−0.504	0.604 (0.384-0.951)
Allergies-Yes	0.023	1.024 (0.498-2.103)
Past 6 months sexual intercourse-Yes	−0.401	0.669 (0.403-1.112)
Frequent alcohol use-Yes	−0.123	0.884 (0.608-1.286)
Highest education level-Primary	0.662	1.940 (0.604-6.229)
Highest education level-Post-Primary	0.460	1.585 (0.496-5.067)

Table 10. Joint model results

Current value parameterization

Effect	Estimate	SE	z	p-value
Longitudinal Process				
β_0	20.690	0.491	42.124	< 0.0001
Time	−1.571	0.685	−2.292	0.0219
Time^2	0.978	0.316	3.091	0.0020
Allergies-Yes	3.727	1.157	3.220	0.0013
Event Process				
Assoct	0.028	0.045	0.620	0.5356

Time-dependent slopes parameterization

Effect	Estimate	SE	z	p-value
Longitudinal Process				
β_0	20.301	0.412	49.283	< 0.0001
Time	−0.923	0.688	−1.341	0.1798
Time^2	0.665	0.309	2.152	0.0314
Allergies-Yes	5.126	0.436	11.759	< 0.0001
Event Process				
Assoct	−0.084	0.058	−1.457	0.1450
Assoct.s	1.514	0.201	7.551	< 0.0001

*Note: Estimates are based on the piecewise constant baseline risk function

5 Discussion and Conclusion

5.1 Discussion

This study demonstrated the usefulness of joint models to establish the association between longitudinal information and time to event outcomes. Prior to the current study, there was no published work on joint modelling of CD4 count and wound healing time in HIV-positive men following circumcision.

The linear mixed effect model was used to characterise CD4 count while Cox proportional hazards model was used to model time to wound healing. To estimate the joint model parameters, we chose a piecewise constant baseline risk function. In the current study, we have only studied the association between a single event time outcome (wound healing time) and a single longitudinal outcome (CD4 count). Some CD4 count values were missing for some patients at follow-up visits. This is a missing data problem synonymous with longitudinal studies and we assumed that the missingness mechanism was missing at random (MAR). Thus, the results obtained from the joint models, are still valid under MAR.

Previous studies on wound healing time following circumcision have mainly focused on using Cox proportional hazards model to investigate the association between CD4 count and time to wound healing. The present study however applied a joint modelling framework.

Linear mixed model with random intercept, linear and quadratic slope produced a better model fit to describe the average evolutions in square root CD4 count over time. Our choice of the mean structure in the LME model aligns with Temesgen et al. (2018).

Only allergy was significantly associated with CD4 count in the current study. Our results contradict those obtained by Temesgen et al. (2018) and Mchunu et al. (2020). Temesgen et al. (2018) for instance reported that weight and functional status were significantly associated with CD4 count. In addition, Mchunu et al. (2020) reported that gender, age, log viral load and square root CD8 count were significantly associated with CD4 count. However, our current results might not be comparable to the results reported by these two studies because of other reasons like dissimilarities in characteristics of the population being studied and differences in the mean structures used in the LME models. Unlike our choice of LME model which also had a quadratic time effect, Mchunu et al. (2020) applied a linear mixed effect model with only a random intercept and a linear time effect.

In the Cox proportional hazards model, our results showed that only marital status was associated with wound healing time. This was at borderline statistical significance (p-value=0.046). On the contrary, Feldblum et al. (2016) identified older persons (25+ years), adverse events and lesser pain during device removal to be significantly associated with slow wound healing. Consistent with our results, Rogers et al. (2013) reported that age and alcohol frequency were not significantly associated with time to wound healing. Nevertheless, Rogers et al. (2013) identified early post-operative infection and evidence of tight sutures to be associated with slow wound healing. Unlike our study, Kigozi et al. (2014) reported that alcohol use among HIV-positive patients with CD4 counts ≥ 350 cells/µL was associated with wound healing at the 4^{th} week of follow up visit (p-value=0.044). Nonetheless, our insignificant results on the association between wound healing time and baseline covariates (excluding marital status) echo the results obtained by Kigozi et al. (2014). Lastly, baseline ART status was not significantly associated with wound healing time in the current study and it is consistent with the insignificant results reported by Tshimanga et al. (2017).

Again our results on the association between baseline covariates and time to wound healing may not be fully comparable to other similar circumcision studies because of the differences in the study designs, circumcision device, definition of complete wound healing, population characteristics and the statistical analysis techniques employed. For example, Feldblum et al. (2016) enrolled 427 HIV-uninfected men aged 18–49 years in a prospective cohort study and further grouped them as either healed or not healed by day 42. They later on applied a logistic model in their analysis to determine statistically significant relationships between baseline factors and wound healing time. Additionally, Rogers et al. (2013) age-matched 108 HIV-positive men with the 215 HIV-negative men and therefore the Cox proportional hazards model results reported were based on the age-matched analysis of 108 HIV-positive men and 215 HIV-negative men. However, the current study enrolled 119 HIV-positive men aged 18-49 years and circumcised them using a non-surgical device (Prepex). We later on applied a Cox proportional hazards model, a clear departure from the aforementioned methods used by Feldblum et al. (2016); Rogers et al. (2013).

There was no significant association between the current true level of square root CD4 count and hazard of wound healing as shown in both current value and time-dependent slopes parameterization. However under, time-dependent slopes parametrization, hazard of wound healing was associated with the rate of change in square root CD4 count. Indeed patients with the same current true level of square root CD4 count at a given point in time t, could experience different rate of change in square root CD4 count at the same time point t leading to different progression of their wound healing.

A surgical dorsal slit study in Uganda among HIV positive patients found no significant association between CD4 count and wound healing time among patients aged ≥ 12 years old (Kigozi et al., 2014). Similarly in Zimbabwe, a Prepex study among HIV-positive patients reported no significant association between CD4 count and wound healing time (Tshimanga et al., 2017). In addition, a study of forceps-guided method in Kenya reported no significant difference in wound healing time by baseline CD4 count among HIV-positive patients (p-value=0.20) (Rogers et al., 2013). These studies however employed different analyses approaches. Kigozi et al. (2014) and Rogers et al. (2013) used the Cox proportional hazards model in order to arrive at their conclusions while Tshimanga et al. (2017) used a binomial probability test to evaluate equivalence of proportion of patients healed by baseline CD4 count (< 500 cells/µL Vs. ≥ 500 cells/µL).

The irregular evolution of square root CD4 count over time could be as a result of other factors like a rise in the viral load in HIV positive-patients few weeks after their circumcision. A study done by Baeten et al. (2010) reported that the viral load substantially increased in the forth week post-circumcision among HIV-positive patients that were ART-naive. Unfortunately in our study, we did not study the association between viral load and time to wound healing and its association to CD4 count. It would be important therefore to extend this study in future by including viral load in the analysis. Another biomarker of interest would be penile viral shedding. According to JUNE (2014), penile viral shedding peaked at the first week post-circumcision then declined to undetectable levels by the the sixth week post-circumcision. Therefore penile viral shedding too would offer some good insight for the irregularity in evolution of square root CD4 count over time.

5.2 Conclusion

We met all the objectives of the study. We found no significant association between current true level of square root CD4 count and wound healing time. However, the rate of change in square root CD4 count was a strong predictor of hazard of wound healing. In summary, circumcising HIV-positive patients with any level of square root CD4 count is not harmful to their post-circumcision wound healing. However, patients with the same current true level of square root CD4 count could exhibit different slopes of the square root CD4 count trajectory at the same time point leading to different progression of wound healing between them. The irregular trajectory in square root CD4 count could be as a result of other biomarkers like penile viral shedding and viral load that often rise in the first few weeks after circumcision.

5.3 Study Limitations

There were missing data in CD4 count and some baseline covariates. The assumption was that the data were missing at random (MAR). These results are therefore only valid under MAR.

5.4 Future Research

The present study only considered one biomarker, CD4 count. It would be important to consider other biomarkers like viral load and penile viral shedding and analyse them jointly, taking into account their association structures. Since missing data in other covariates was challenging to handle in our study, future studies should consider imputing them using a multiple imputation technique which is compatible with a joint model for longitudinal and time to event data. Lastly, future studies should also consider doing some sensitivity analysis to determine how departures from MAR assumption influence parameter estimates.

References

Alsefri, M., Sudell, M., García-Fiñana, M., & Kolamunnage-Dona, R. (2020). Bayesian joint modelling of longitudinal and time to event data: a methodological review. *BMC Medical Research Methodology*, *20*, 1–17.

Auvert, B., Taljaard, D., Lagarde, E., Sobngwi-Tambekou, J., Sitta, R., & Puren, A. (2005). Randomized, controlled intervention trial of male circumcision for reduction of hiv infection risk: the anrs 1265 trial. *PLoS medicine*, *2*(11).

Baart, S. J., Boersma, E., & Rizopoulos, D. (2019). Joint models for longitudinal and time-to-event data in a case-cohort design. *Statistics in medicine*, *38*(12), 2269–2281.

Baeten, J. M., Donnell, D., Kapiga, S. H., Ronald, A., John-Stewart, G., Inambao, M., … Celum, C. (2010). Male circumcision and risk of male-to-female hiv-1 transmission: a multinational prospective study in african hiv-1 serodiscordant couples. *AIDS (London, England)*, *24*(5), 737.

Bailey, R. C., Moses, S., Parker, C. B., Agot, K., Maclean, I., Krieger, J. N., … Ndinya-Achola, J. O. (2007). Male circumcision for hiv prevention in young men in kisumu, kenya: a randomised controlled trial. *The lancet*, *369*(9562), 643–656.

Bernhardt, P. W., Zhang, D., & Wang, H. J. (2015). A fast em algorithm for fitting joint models of a binary response and multiple longitudinal covariates subject to detection limits. *Computational statistics & data analysis*, *85*, 37–53.

Brown, E. R., Ibrahim, J. G., & DeGruttola, V. (2005). A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, *61*(1), 64–73.

Buta, G. B., Goshu, A. T., & Worku, H. M. (2015). Bayesian joint modelling of disease progression marker and time to death event of hiv/aids patients under art follow-up. *Journal of Advances in Medicine and Medical Research*, 1034–1043.

Cekic, S., Aichele, S., Brandmaier, A. M., Köhncke, Y., & Ghisletta, P. (2019). A tutorial for joint modeling of longitudinal and time-to-event data in r. *arXiv preprint arXiv:1909.05661*.

Collett, D. (2015). *Modelling survival data in medical research*. Chapman and Hall/CRC.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Dafni, U. G., & Tsiatis, A. A. (1998). Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, 1445–1462.

Dessiso, A. H., & Goshu, A. T. (2017). Bayesian joint modelling of longitudinal and survival data of hiv/aids patients: a case study at bale robe general hospital, ethiopia. *American Journal of Theoretical and Applied Statistics*, *6*(4), 182–190.

Drikvandi, R., Verbeke, G., Khodadadi, A., & Partovi Nia, V. (2013). Testing multiple variance components in linear mixed-effects models. *Biostatistics*, *14*(1), 144–159.

Elashoff, R., Li, N., et al. (2016). *Joint modeling of longitudinal and time-to-event data*. CRC Press.

Elashoff, R. M., Li, G., & Li, N. (2008). A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics*, *64*(3), 762–771.

Erango, M. A., Goshu, A. T., Buta, G. B., & Dessisoa, A. (2017). Bayesian joint modelling of survival of hiv/aids patients using accelerated failure time data and longitudinal cd4 cell counts. *Br J Med Med Res*, *20*(6), 1–12.

Faucett, C. L., & Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach. *Statistics in medicine*, *15*(15), 1663–1685.

Feldblum, P. J., Odoyo-June, E., Bailey, R. C., Lai, J. J., Weiner, D., Combes, S., … Cherutich, P. (2016). Factors associated with delayed healing in a study of the prepex device for adult male circumcision in kenya. *Journal of acquired immune deficiency syndromes (1999)*, *72*(Suppl 1), S24.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis* (Vol. 998). John Wiley & Sons.

Garcia, T. P., & Marder, K. (2017). Statistical approaches to longitudinal data analysis in neurodegenerative diseases: huntington's disease as a model. *Current neurology and neuroscience reports*, *17*(2), 14.

Gray, R. H., Kigozi, G., Serwadda, D., Makumbi, F., Watya, S., Nalugoda, F., … others (2007). Male circumcision for hiv prevention in men in rakai, uganda: a randomised trial. *The Lancet*, *369*(9562), 657–666.

Hallett, T. B., Singh, K., Smith, J. A., White, R. G., Abu-Raddad, L. J., & Garnett, G. P. (2008). Understanding the impact of male circumcision interventions on the spread of hiv in southern africa. *PloS one*, *3*(5).

He, B., & Luo, S. (2016). Joint modeling of multivariate longitudinal measurements and survival data with applications to parkinson's disease. *Statistical methods in medical research*, *25*(4), 1346–1358.

Henderson, R., Diggle, P., & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, *1*(4), 465–480.

Hickey, G. L., Philipson, P., Jorgensen, A., & Kolamunnage-Dona, R. (2016). Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC medical research methodology*, *16*(1), 117.

Hsieh, F., Tseng, Y.-K., & Wang, J.-L. (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*, *62*(4), 1037–1043.

Huang, Y., Dagne, G., & Wu, L. (2011). Bayesian inference on joint models of hiv dynamics for time-to-event and longitudinal data with skewness and covariate measurement errors. *Statistics in Medicine*, *30*(24), 2930–2946.

JUNE, E.-O. (2014). *Wound healing and resumption of heterosexual intercourse following voluntary medical circumcision of adult males in kisumu city, kenya* (Unpublished doctoral dissertation). University of Nairobi.

Kigozi, G., Musoke, R., Kighoma, N., Watya, S., Serwadda, D., Nalugoda, F., … others

(2014). Male circumcision wound healing in human immunodeficiency virus (hiv)-negative and hiv-positive men in r akai, u ganda. *BJU international*, *113*(1), 127–132.

Lawrence Gould, A., Boye, M. E., Crowther, M. J., Ibrahim, J. G., Quartey, G., Micallef, S., & Bois, F. Y. (2015). Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the dia bayesian joint modeling working group. *Statistics in medicine*, *34*(14), 2181–2195.

Mauff, K., Steyerberg, E., Kardys, I., Boersma, E., & Rizopoulos, D. (2020). Joint models with multiple longitudinal outcomes and a time-to-event outcome: a corrected two-stage approach. *Statistics and Computing*, 1–16.

Mchunu, N. N., Mwambi, H. G., Reddy, T., Yende-Zuma, N., & Naidoo, K. (2020). Joint modelling of longitudinal and time-to-event data: an illustration using cd4 count and mortality in a cohort of patients initiated on antiretroviral therapy. *BMC infectious diseases*, *20*, 1–9.

Molenberghs, G., & Verbeke, G. (2006). *Models for discrete longitudinal data.* Springer Science & Business Media.

Moore, D. F. (2016). *Applied survival analysis using r.* Springer.

Njagi, E. N., Molenberghs, G., Rizopoulos, D., Verbeke, G., Kenward, M. G., Dendale, P., & Willekens, K. (2016). A flexible joint modeling framework for longitudinal and time-to-event data with overdispersion. *Statistical methods in medical research*, *25*(4), 1661–1676.

Njagi, E. N., Rizopoulos, D., Molenberghs, G., Dendale, P., & Willekens, K. (2013). A joint survival-longitudinal modelling approach for the dynamic prediction of rehospitalization in telemonitored chronic heart failure patients. *Statistical Modelling*, *13*(3), 179–198.

Njeuhmeli, E., Forsythe, S., Reed, J., Opuni, M., Bollinger, L., Heard, N., … others (2011). Voluntary medical male circumcision: modeling the impact and cost of expanding male circumcision for hiv prevention in eastern and southern africa. *PLoS medicine*, *8*(11).

Papageorgiou, G., Mauff, K., Tomer, A., & Rizopoulos, D. (2019). An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual review of statistics and its application.*

R. Brown, E., & G. Ibrahim, J. (2003). A bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, *59*(2), 221–228.

Rizopoulos, D. (2010). Jm: An r package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software (Online)*, *35*(9), 1–33.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in r.* Chapman and Hall/CRC.

Rizopoulos, D. (2014). The r package jmbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *arXiv preprint arXiv:1404.7625.*

Rizopoulos, D., Verbeke, G., & Lesaffre, E. (2009). Fully exponential laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(3), 637–654.

Rogers, J. H., Odoyo-June, E., Jaoko, W., & Bailey, R. C. (2013, Apr). Time to Complete Wound Healing in HIV-Positive and HIV-Negative Men following Medical Male Circumcision in Kisumu, Kenya: A Prospective Cohort Study. *PLoS ONE*, *8*(4), e61725. doi: 10.1371/journal.pone.0061725

Sène, M., Bellera, C., & Proust-Lima, C. (2013, October). *Shared random-effect models for the joint analysis of longitudinal and time-to-event data: application to the prediction of prostate cancer recurrence.*

Seyoum, A., & Temesgen, Z. (2017). Joint longitudinal data analysis in detecting determinants of cd4 cell count change and adherence to highly active antiretroviral therapy at felege hiwot teaching and specialized hospital, north-west ethiopia (amhara region). *AIDS research and therapy*, *14*(1), 14.

Sweeting, M. J., & Thompson, S. G. (2011). Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, *53*(5), 750–763.

Temesgen, A., Gurmesa, A., & Getchew, Y. (2018). Joint modeling of longitudinal cd4 count and time-to-death of hiv/tb co-infected patients: A case of jimma university specialized hospital. *Annals of Data Science*, *5*(4), 659–678.

Tshimanga, M., Makunike-Chikwinya, B., Mangwiro, T., Tapiwa Gundidza, P., Chatikobo, P., Murenje, V., … others (2017). Safety and efficacy of the prepex device in hiv-positive men: A single-arm study in zimbabwe. *PloS one*, *12*(12), e0189146.

Tsiatis, A. A., & Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, *88*(2), 447–458.

Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 809–834.

Tsiatis, A. A., Degruttola, V., & Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American statistical association*, *90*(429), 27–37.

Verbeke, G. (1997). Linear mixed models for longitudinal data. In *Linear mixed models in practice* (pp. 63–153). Springer.

Verbeke, G., Molenberghs, G., & Rizopoulos, D. (2010). Random effects models for longitudinal data. In *Longitudinal research with latent variables* (pp. 37–96). Springer.

Wang, Y., & Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, *96*(455), 895–905.

Wu, L., Liu, W., Yi, G. Y., & Huang, Y. (2012). Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. *Journal of Probability and Statistics*, *2012*.

Wulfsohn, M. S., & Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 330–339.

Yang, L., Yu, M., & Gao, S. (2016). Joint models for multiple longitudinal processes and time-to-event outcome. *Journal of statistical computation and simulation*, *86*(18), 3682–3700.

Yu, M., Law, N. J., Taylor, J. M., & Sandler, H. M. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica*, 835–862.

Yu, T., Wu, L., & Gilbert, P. (2019). New approaches for censored longitudinal data in joint modelling of longitudinal and survival data, with application to hiv vaccine studies. *Lifetime data analysis*, *25*(2), 229–258.

A R codes used in the analysis of the data

A.1 Final linear mixed effects model

```
LMFit <- lme(SQRTCD4   (TIME+I(TIME^2))+factor(ALLERGIES),
random =   (TIME+I(TIME^2)) | SUBJECT,
data = datalong, control=ctrl)

summary(LMFit)
```

A.2 Final Cox proportional hazards model

```
CoxFit <- coxph(Surv(DURATION, STATUS)∼ 1,
data = datacox2,x = TRUE)

summary(CoxFit)
```

A.3 Current value parametrization

```
jointFit1 <- jointModel(LMFit, CoxFit, method = "piecewise-PH-aGH",
timeVar = "TIME", verbose = TRUE,
iter.EM = 500)

summary(jointFit1)
exp(confint(jointFit1,parm="Event"))
```

A.4 Time-dependent slopes parametrization

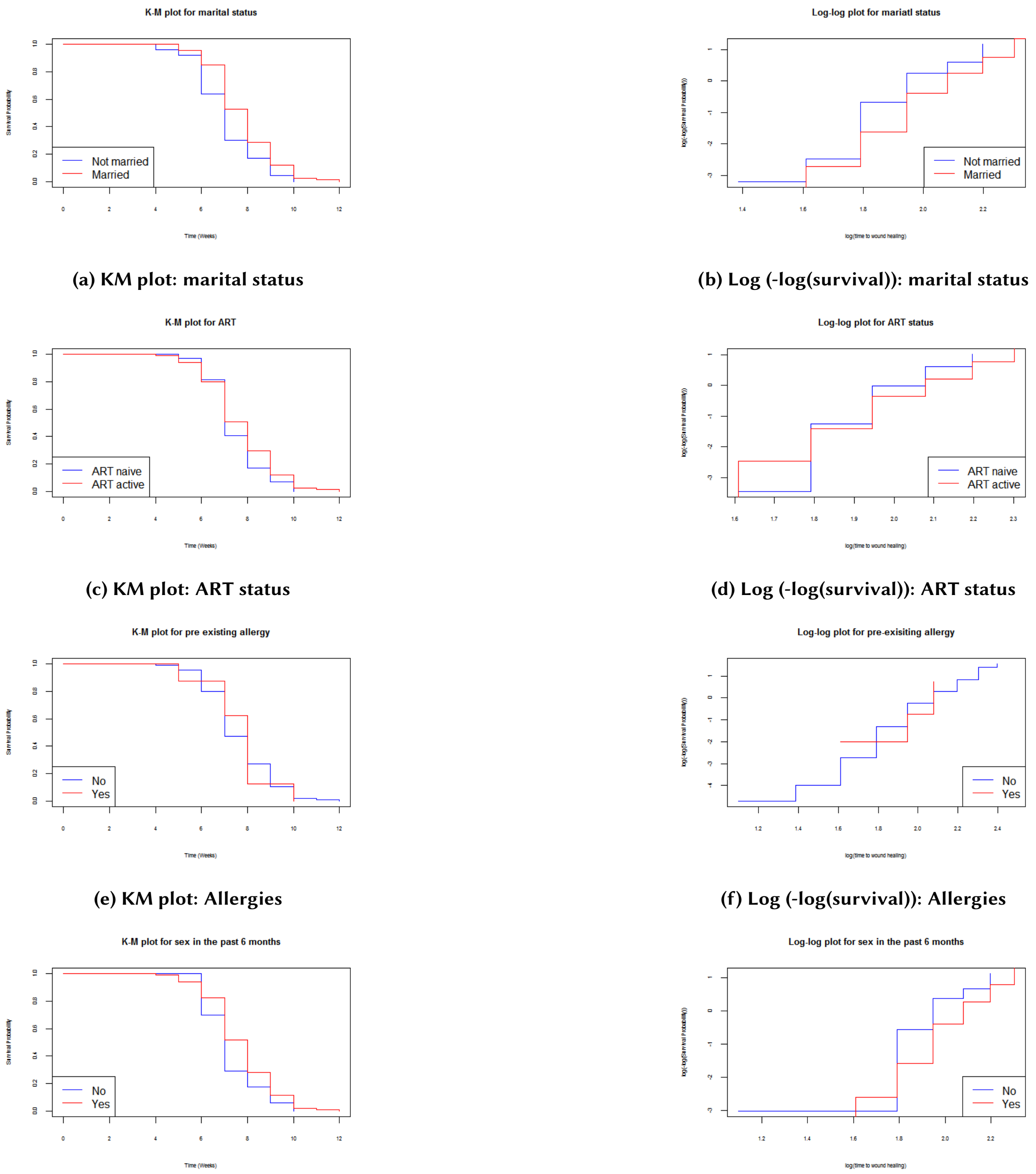
```
dform <- list(fixed =   I(2*TIME)
,indFixed = 3:4, random =  I(2*TIME), indRandom =2:3)

jointFit2 <- update(jointFit1, parameterization = "both",
derivForm = dform)

summary(jointFit2)
confint(jointFit2,parm="Event")
```

B Figures

B.1 Kaplan-Meier curves and log(-log(survival)) curves



(a) KM plot: marital status

(b) Log (-log(survival)): marital status

(c) KM plot: ART status

(d) Log (-log(survival)): ART status

(e) KM plot: Allergies

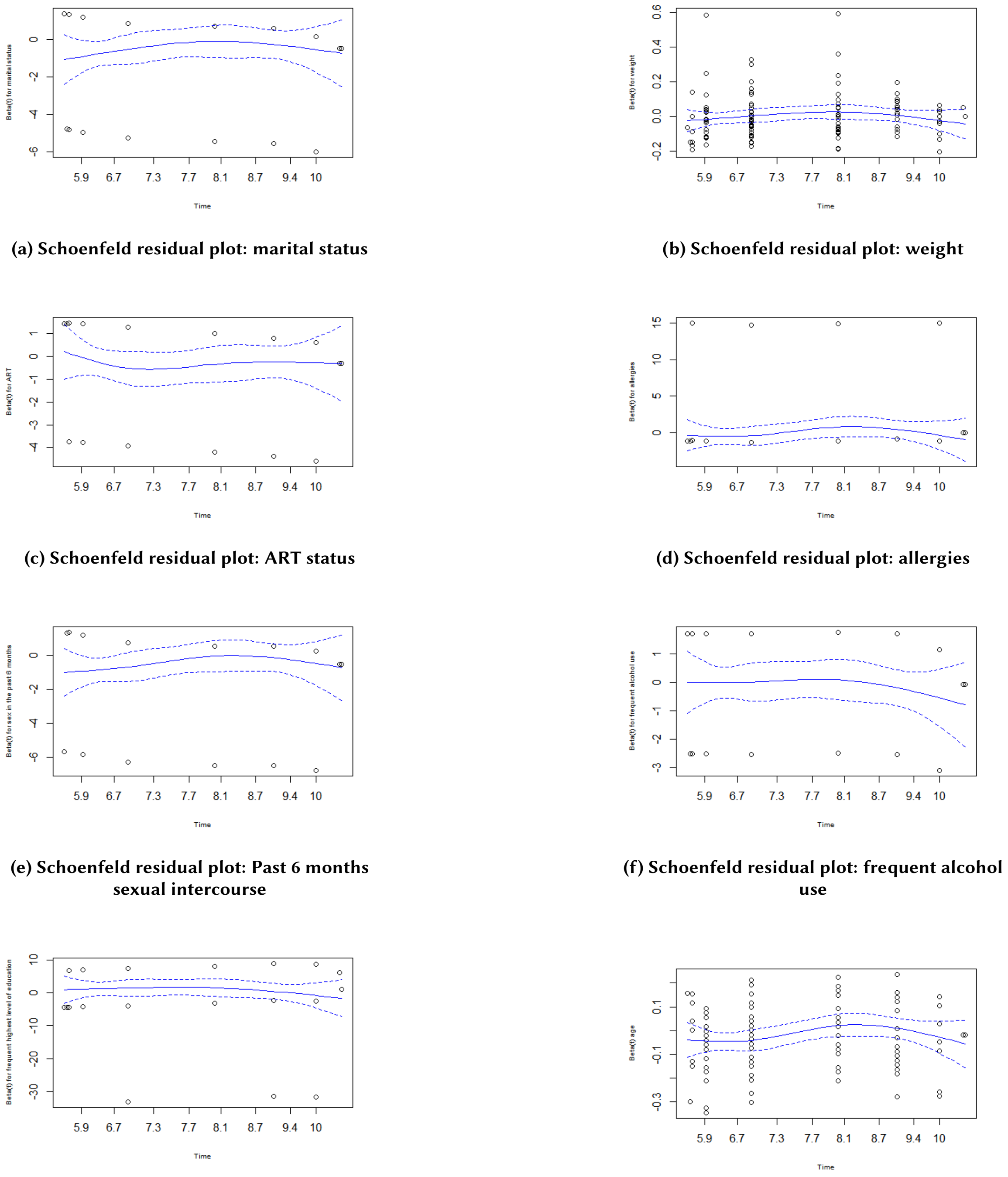
(f) Log (-log(survival)): Allergies

(g) KM plot: Past 6 months sexual intercourse

(h) Log (-log(survival)): Past 6 months sexual intercourse

Figure 8. Kaplan-Meier curves and log(-log(survival)) curves

B.2 Schoenfeld residual plots



(a) Schoenfeld residual plot: marital status

(b) Schoenfeld residual plot: weight

(c) Schoenfeld residual plot: ART status

(d) Schoenfeld residual plot: allergies

(e) Schoenfeld residual plot: Past 6 months sexual intercourse

(f) Schoenfeld residual plot: frequent alcohol use

(g) Schoenfeld residual plot: highest education level

(h) Schoenfeld residual plot: age

Figure 9. Schoenfeld residual plots