Diagnostics for joint models for longitudinal and survival data

Isaac Luwinga Singini

A thesis presented for the degree of
DOCTOR OF PHILOSOPHY

Department of Statistical Sciences
University of Cape Town
Cape Town, South Africa

Supervisor: Associate Professor Freedom Gumedze
Co-Supervisor: Professor Henry Mwambi

July, 2021

Declaration

I have read and understood the School's definition of plagiarism and cheating given in the Research Degrees Handbook. I declare that this thesis is my own work. I have acknowledged all results and quotations from the published or unpublished work of other people.

Mr. Isaac Luwinga Singini

Signature:...Signed by candidate... Date: 20th July 2021

Dedication

To my anchors and hope for strength:
To my wife Ruth, son Wongani, daughter Silandera and son Chimango whose endurance and patience have been tested through the years of my research. This unfaltering dedication will never offset the sacrifices you have made to support me during this research.
To the memory of my beloved mother - Margret Singini (Nee Jere, RIP)

Table of contents

List of Figures

List of Tables

List of Definitions

Abstract

Joint models for longitudinal and survival data are a class of models that jointly analyse an outcome repeatedly observed over time such as a bio-marker and associated event times. These models are useful in two practical applications; firstly focusing on survival outcome whilst accounting for time varying covariates measured with error and secondly focusing on the longitudinal outcome while controlling for informative censoring. Interest on the estimation of these joint models has grown in the past two and half decades. However, minimal effort has been directed towards developing diagnostic assessment tools for these models. The available diagnostic tools have mainly been based on separate analysis of residuals for the longitudinal and survival sub-models which could be sub-optimal. In this thesis we make four contributions towards the body of knowledge. We first developed influence diagnostics for the shared parameter joint model for longitudinal and survival data based on Cook's statistics. We evaluated the performance of the diagnostics using simulation studies under different scenarios. We then illustrated these diagnostics using real data set from a multi-center clinical trial on TB pericarditis (IMPI). The second contribution was to implement a variance shift outlier model (VSOM) in the two-stage joint survival model. This was achieved by identifying outlying subjects in the longitudinal sub-model and down-weighting before the second stage of the joint model. The third contribution was to develop influence diagnostics for the multivariate joint model for longitudinal and survival data. In this setting we considered two longitudinal outcomes, square root CD4 cell count which was Gaussian in nature and antiretroviral therapy (ART) uptake which was binary. We achieved this by extending the univariate case

based on Cook's statistics for all parameters. The fourth contribution was to implement influence diagnostics in joint models for longitudinal and survival data with multiple failure types (competing risk). Using IMPI data set we considered two competing events in the joint model; death and constrictive pericarditis. Using simulation studies and IMPI dataset the developed diagnostics identified influential subjects as well as observations. The performance of the diagnostics was over 98% in simulation studies. We further conducted sensitivity analyses to check the impact of influential subjects and/or observations on parameter estimates by excluding them and re-fitting the joint model. We observed subtle differences, overall in the parameter estimates, which gives confidence that the initial inferences are credible and can be relied on. We illustrated case deletion diagnostics using the IMPI trial setting, these diagnostics can also be applied to clinical trials with similar settings. We therefore make a strong recommendation to analysts to conduct influence diagnostics in the joint model for longitudinal and survival data to ascertain the reliability of parameter estimates. We also recommend the implementation of VSOM in the longitudinal part of the two-stage joint model before the second stage.

Key words: Joint model, diagnostics, Cook's distance, two-stage, VSOM.

List of publications

Part of this thesis has been published:

Paper 1 (Chapter 5) **Singini, Mwambi, and Gumedze (2018)**; Diagnostics for detecting influential observations in joint survival models, In: Annual Proceedings of the South African Statistical Association Conference. Vol. 2018, Congress 1. South African Statistical Association (SASA) pp. 49-56.

Paper 2 (Chapter 4) **Singini I, Mwambi H, Gumedze F (2020)**; Diagnostics for a two-stage joint survival model, In: Communications in Statistics - Simulation and Computation. Manuscript ID is LSSP-2019-0828, Reviewed and resubmitted for publication.

Acknowledgments

My deepest appreciation goes to Associate Professor Freedom Gumedze (Department of Statistical Sciences, University of Cape Town) for his expert guidance and encouragement during this research. Associate Professor Gumedze not only provided the research topic but also provided support to secure funding for the last one and a half years when core funding for this research was unavailable due to the closure of the grant circle. It was during this period of uncertainty that he even went further to patiently guide and support this research in every way possible.

My sincere appreciation goes to Professor Henry Mwambi (School of Mathematics, Statistics and Computer Science, University of KwaZulu Natal, South Africa) who co-supervised this research and provided invaluable statistical input and guidance through-out the tenure of this research.

My heartfelt thanks to Associate Professor Francesca Little for the refresher biostatistics courses in longitudinal data analysis and survival data analysis which are the basic building blocks for this research.

I, gratefully, acknowledge Professor Jane Hutton (Department of Statistics, University of Warwick, UK) of The Newton Fund, The Academy of Medical Sciences, UK for supporting my attendance at the Academy for PhD Training in Statistics (APTS - summer 2017) as well as providing the much needed guidance on survival analysis within the joint modeling framework.

I am grateful and indebted to the academic and non-academic staff, and post-graduate stu-

dents of the Department of Statistical Sciences for their friendliness and provision of a conducive environment to conduct this research. This research has been made possible by the financial support from Professor Taha Taha (PI) & Team of The Training in HIV-related Non-communicable Disease Complications Program in Malawi (NIH Fogarty grant D43TW009609) for the first three years and Professor Graeme Meintjes (PI) & Team of The Fogarty International Center of the National Institutes for Health, under Award Number D43TW010559 for the last one and a half years.

I am further indebted to The Mayosi Research Group, Department of Medicine, University of Cape Town for providing the data for this research.

Computations for Chapter six and seven were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team: hpc.uct.ac.za

Definitions

An observation could be unusual with respect to its y-value or x-value in a simple linear regression model. The unusual observations are formally categorised as **outlier, leverage, influential points, and residual** according to their impact on the regression model, i.e. with respect to x- or y- values.

Definition 1. *Outlier:* an outlier is defined as an unusual observation with respect to either x-value or y-value. An x-outlier makes the scope of the regression estimates too broad, which is usually considered less accurate.

Definition 2. *Leverage:* Leverage – a data point whose x-value (independent) is unusual, y-value follows the predicted regression line though. A leverage point may look okay as it sits on the predicted regression line. However, a leverage point will inflate the strength of the regression relationship by both the statistical significance (reducing the p-value to increase the chance of a significant relationship) and the practical significance (increasing r-square). Unfortunately, leverage points have no impact on the coefficients because the point follows the predicted regression line.

Definition 3. *Influence:* influential point– a data point that unduly influences the regression coefficients (estimates). A point is considered influential if its exclusion causes major changes in the fitted regression function. Depending on the location of the point, it may affect all statistics, including the p-value, r-square, coefficients, and intercept.

Definition 4. *Residual:* a residual is the difference between the observed value and the mean value that the model predicts for that observation. Residual values are especially useful in regression because they indicate the extent to which a model accounts for the variation in the

observed data

Chapter 1

Introduction

In medical research it is usually the case where data are typically recorded repeatedly over planned time points. Study designs that collect or record data repeatedly on an outcome variable for each subject at some selected time points are not only common in medical research but also other scientific disciplines (Rizopoulos 2012). The main focus of such study designs on the response variable is to either evaluate population average longitudinal trajectory or subject specific trajectory over time. In the process of collecting data for the response variable other covariates that are considered to have an effect on the response variable may also be recorded. In longitudinal study designs it is common to observe events that may be of interest to the investigator especially in medical research. As an example while the investigator could be recording biomarkers such as serum bilirubin levels among patients who are treated for liver cirrhosis there could be deaths or relapse of the disease among other events that may be of interest. Statistical models are useful in many fields such as engineering, finance, space science, biology, pharmacology and medicine just to mention but a few.

In the past three decades, the longitudinal (measurement) and the event process were analysed separately until recently. There has been statistical methodology that has been developed that allows the two processes to be jointly analysed using a modeling framework commonly known as joint modeling for longitudinal and event-time (survival) data (Henderson, Diggle, and Dobson 2000; Hsieh, Tseng, and Wang 2006; Rizopoulos 2012; Tsiatis, Degruttola, and

Wulfsohn 1995; Wulfsohn and Tsiatis 1997) through semi-parametric maximum likelihood while the Bayesian approach using markov chain monte carlo (MCMC) techniques have also been developed (Chi and Ibrahim 2006; R Brown and G Ibrahim 2003; Wang and Taylor 2001; Xu and Zeger 2001).

There are different classes of joint models namely; latent class joint models which are primarily useful for predicting the occurrence of an event (Proust-Lima et al. 2009, 2014) which are handy in personalised medicine and shared parameters joint models which assess the strength of association between a bio-marker and risk of an event (Henderson, Diggle, and Dobson 2000; Rizopoulos 2012; Tsiatis and Davidian 2004), this is the main focus of this thesis.

This statistical modeling framework has the advantage that it accounts for non-random drop out due to the occurrence of an event in the measurement process. In the event process with an endogenous time-dependent covariate, it accounts for measurement error in the bio-marker as well as assessing the strength of association between the time-dependent covariate and risk for an event (Rizopoulos 2012). Parameter estimation of joint models for longitudinal and survival data and their respective scenarios are presented in Chapter 5 Section 5.1.

An equally important step in statistical modeling is to validate the model for different aspects such as assumptions made, measures of influence and model adequacy. There are different approaches that have been developed with the most common being based on residual analysis. The standard practice is to plot these residuals and then visually see the performance of the model. There are, notwithstanding, tests based on hypothesis for assessing model fit. There are also other techniques that are used such as robust estimation and influence diagnostics that have been developed as tools to assess specific aspects of statistical models.

An assessment of joint modeling mentioned above is done using residuals from the separate

sub-models. Influence diagnostics of the joint models have not been fully developed (Rizopoulos 2012; Rizopoulos, Verbeke, and Molenberghs 2010). The technique for influence diagnostics in the joint modeling framework has been using qualitative graphical displays to isolate departures from model assumptions and influential subjects (Dobson and Henderson 2003). It was observed by Dobson and Henderson (2003) that lack of computing power to do a full case deletion and re-estimation for each subject which was not practical at the time.

There has been work on model assessment tools using multiple imputation of residuals by Rizopoulos, Verbeke, and Molenberghs (2010). This is motivated by the fact that there is attrition in the longitudinal outcome due to occurrence of an event. This non-random drop out makes the use of standard goodness of fit measures from residuals unattainable because there is no reference distribution available and hence residual plots based on observed data alone can be misleading because they do not have standard properties. The main idea in this work is to augment the observed data with randomly imputed data points that would otherwise have been provided had the subject not dropped out. This creates complete data from which residuals can be calculated and plots produced, this is based on Bayesian techniques. Analysis of these residuals is again through separate sub-models as reported by Rizopoulos, Verbeke, and Molenberghs (2010). There has also been a Bayesian approach to assessing influence measures for the joint model by Zhu et al. (2012) motivated by a multi-center clinical trial of breast cancer. This work involves independently perturbing the joint model, subjects and random effects for both sub-models. These perturbations are done to priors and all components of the Bayesian joint model. These are finally combined for all the three perturbations to the prior which creates a sampling distribution from which data is obtained. These data permit an assessment on simultaneous sensitivity of all components of the Bayesian analysis. The

proposed methods are able to detect outliers, influential observations and assess the sensitivity of inference to various unverifiable assumptions on the Bayesian analysis of joint models. Another Bayesian approach that has been studied using Q-functions for case-deletion is by Tang, Tang, and Zhu (2017) with applications to skew-normal distributions. In this work a modified likelihood displacement diagnostic statistic is used. This is because it is difficult to evaluate a likelihood displacement of observed data using Q-functions. A quantity that is a Q-function, is calculated which evaluates the effect of the i^{th} observation on the maximum Q-function after excluding the i^{th} observation from the estimation of $\boldsymbol{\theta}$.

An assessment of goodness of fit has been done for joint models using Akaike information criterion (AIC) and the Bayesian information criterion (BIC) (Baghfalaki, Ganjali, and Verbeke 2017; Park and Qiu 2014; Zhang et al. 2014). The work of Baghfalaki, Ganjali, and Verbeke (2017) mainly focuses on distributional assumptions which has the potential of misleading inferences if there is heterogeneity of random effects in the shared parameter joint model, while Park and Qiu (2014) focus on separate goodness of fit assessment and diagnostics for the two sub-models. The goodness of fit by the two authors is based on both extensions to AIC and BIC which accounts for the number of parameters and sample size. On the other hand, Zhang et al. (2014), developed a quantifiable change in AIC and BIC for fitting survival data with and without the longitudinal data. This is done by decomposing the AIC and BIC into additive components from which the goodness of fit is derived. Using this approach, it is possible to compare longitudinal outcomes in terms of their contribution to the overall fit of the survival data. Model assessment is done in two parts based on likelihood theory by using the conditional density of the random effects given the longitudinal part which is replicated in the survival part of the model. This motivates an idea to decompose Cook's statistics (Cook

1977; Cook and Weisberg 1982) for diagnostics of the joint model from the joint conditional density given the random effects.

There has been little work on developing influence diagnostics for joint models based on the full conditional density of the joint model (Rizopoulos 2012) despite availability of software packages to fit these joint models and an increase in computer processing power. In this thesis we develop case-deletion (influence) diagnostics for the joint model based on Cook's statistics (Cook and Weisberg 1982) on the full parameter set which includes fixed effects, variance components and association parameter of the joint model. We approach this by decomposing and calculating Cook's distance for the joint model parameters. Case-deletion diagnostics are possible to implement because computing power has increased in the last decade and software is readily available to do a full case-deletion diagnostics on the joint model.

1.1 Contributions

In this thesis we have developed influence diagnostics for joint models for longitudinal and survival data and made the following contributions towards the body of knowledge:

(i) We have developed case deletion diagnostics for shared parameter joint models based on Cook's statistics. This is important because influential subjects and/or observations have an impact on statistical inferences hence need to investigate them. We have conducted simulation studies for the case deletion diagnostics to evaluate their performance and results are published in Singini, Mwambi, and Gumedze (2018) paper (see Chapter 5)

(ii) We have implemented the variance shift outlier model (VSOM) for the two-stage joint survival model. The VSOM is important in medical research especially laboratory based

when the data points are not enough to guarantee the required statistical power after excluding influential or outlying data points in the modeling process. It is known that two-stage joint model is straight forward to implement and it is not computationally intensive (Mauff et al. 2018, 2020), hence the VSOM is ideal to augment in the longitudinal process at the discretion of the researcher.

(iii) We have performed case deletion diagnostics in a multivariate joint model involving two longitudinal outcomes using the IMPI data set. We have demonstrated the developed case deletion diagnostics using CD4 cell count which is measured on a continuous scale and ART uptake which is binary. This is important in clinical trials, since it is common to have more than one longitudinal outcome from some known family of distributions and these outcomes are likely correlated, which is dealt with by fitting a multivariate joint model conditional on random effects. It is, therefore, equally worthwhile to investigate influential subjects and/or observations in the multivariate joint model for longitudinal and survival data.

(iv) We have illustrated case deletion diagnostics in joint model for longitudinal and survival data with multiple failure types. In clinical trials it is typical to observe competing events which could be correlated, therefore, if the joint modeling is considered for analysis then case deletion diagnostics should be performed to evaluate the impact of influential subjects on model estimates in the competing risk setting.

1.2 Rationale

It is a well-known fact that outliers and/or influential data points can lead to making statistical inferences that could be misleading in clinical trials. There is need for guidelines on design, conduct and statistical analysis of clinical trials that have the potential of having outliers and/or influential data points. When drawing inferences from a joint model for longitudinal and survival data, it may lead to overestimation of treatment effects especially in clinical trials. It is, imperative, therefore to conduct a sensitivity analysis after fitting the shared parameter joint model conditional on random effects. The sensitivity analysis is based on case deletion diagnostics evaluated from the Cook's statistics. This gives confidence about the estimates of treatment effect if they remain subtly the same with and without influential data points. Case deletion diagnostics is the main aim of this thesis on joint model for longitudinal and survival data. We illustrated these diagnostics on longitudinal CD4 cell count and ART uptake as the outcomes of interest with constrictive pericarditis and death as the time-to event outcomes (Mayosi et al. 2014). We also designed simulation studies to evaluate these diagnostics before illustrating them on real data set. We further provide a discussion and recommendation on the implementation of these case deletion diagnostics using JM Package and JMBayes Package in R statistical software.

1.3 Objectives

The following are the objectives of this study:

(i) Implement case deletion diagnostics for joint models for longitudinal and survival data,

evaluate their performance in simulation studies and illustrate them on real data set from a multi-centre clinical trial (IMPI).

(ii) Implement a variance shift outlier model in a two-stage joint model for longitudinal and survival data.

(iii) Implement case deletion diagnostics for a multivariate joint model for longitudinal and survival data.

(iv) Implement influence diagnostics for joint models with multiple failure types (competing risk).

1.4 Thesis outline

In this Chapter we have introduced the thesis, rationale and its main objectives. This Chapter under this Section also provides an outline of the thesis which is divided into eight chapters.

Chapter 2 reviews linear mixed models (Laird and Ware 1982), Cox model (Collett 2015; Cox 1972b) and their respective diagnostics. In this Chapter, we also describe the motivating data set (Mayosi et al. 2014) that is used to illustrate diagnostics for the joint model. This is necessary because it provides insights about the variables that are used in the joint model as outcomes and predictors.

In Chapter 3, we review joint model for longitudinal and survival data (Henderson, Diggle, and Dobson 2000; Rizopoulos 2012; Tsiatis, Degruttola, and Wulfsohn 1995). This review is necessary because it lays the foundation for understanding joint models for longitudinal and

survival data which are the basic building blocks for case deletion diagnostics. A review of existing model assessment tools and diagnostics is also given in this Chapter.

In Chapter 4, we implement case deletion diagnostics for the two-stage joint model using the variance shift outlier model (VSOM) (Gumedze and Chatora 2014; Gumedze et al. 2010). We begin by introducing the two-stage joint model and its formulation, then the VSOM with its hypothesis tests. This implementation is illustrated using the IMPI data set.

Chapter 5 introduces the linear mixed model and its estimation, the survival model (Cox 1972a) and its estimation. We then introduce the shared parameter joint model, its formulation and estimation after which we develop case deletion diagnostics for this joint model based on Cook's statistics (Cook and Weisberg 1982). These diagnostics are evaluated in simulations studies with different scenarios before being illustrated in real data set from a cardiology multi-centre clinical trial (IMPI) by the Mayosi Research Group (Mayosi et al. 2014).

Chapter 6 introduces the multivariate joint model for longitudinal and survival data, its formulation and estimation (Rizopoulos 2012). We develop influence diagnostics for this model and illustrate the diagnostics on real data set (IMPI).

In Chapter 7, we apply case deletion diagnostics to joint models for longitudinal and survival data with multiple failure types i.e., competing risk. We first introduce the joint model with multiple failure types, its formulation and estimation after which we illustrate case deletion diagnostics using the IMPI data set.

In Chapter 8, we provide a discussion, draw conclusions and identify potential areas for further research.

Chapter 2

Literature review

The structure of this review is as follows: Section 2.1.1 is the review of linear mixed model and diagnostics for linear mixed model. Section 2.3 is the review of extended Cox model and diagnostics for the survival model.

2.1 The linear mixed model

Introduction

In this Chapter we introduce and review the linear mixed model which is a sub-model in the joint models for longitudinal and survival data. This is important because it lays the foundation for understanding the estimation of the joint model and subsequently will lead to developing diagnostics for the joint model.

We review the joint model for longitudinal and survival data by firstly reviewing its history and evolution. This involves starting with time to event models with time dependent covariates and its extension to the two staged approach to joint modeling longitudinal and survival data. Diagnostics based on this approach are derived from the sub-models since the two staged strategy is based on fitting the linear mixed model and extracting the fitted values from the linear mixed model to be used as time dependent covariates in the Cox survival model.

2.1.1 Review of the linear mixed model

The goal of this Section is to review the general structure of longitudinal data, review standard methods used for analysing such data (model fitting) and the techniques that are available for model checking (diagnostics). This is important because it will set the scene into the joint model formulation as well as joint model diagnostics. The motivating data set has longitudinal measurements for both discrete outcomes for example ART uptake in HIV positive subjects and Gaussian outcomes for example (square root CD4 cell count) in HIV positive subjects.

Study designs that collect longitudinal measurements have two main aims under investigation and these are: an assessment of a cross-sectional effect and longitudinal effect. The cross-sectional effect investigates how the effect of treatment differs on average at specific time points, while the longitudinal effect investigates if there exists differences in the treatment effect (Rizopoulos 2012, p.16).

It is well known that longitudinal measurements from each subject over time are positively correlated. This, therefore, means that standard statistical tools which are based on the independence assumption such *t*-test and simple linear regression based on ordinary least squares estimation (OLS) can not be used for longitudinal data analysis. In order to provide valid inferences from the analysis of longitudinal outcomes, there is need to account for correlation. There exists techniques in literature for standard longitudinal data analysis that accounts for this including marginal models such generalised estimation equations (GEE). In the past decades linear mixed models (LMM) have been extensively studied and developed in the last decades (Diggle, Liang, and Zeger 1994; Fitzmaurice, Laird, and Ware 2012; Molenberghs and Verbeke 2000) among others. Suffice to say that the optimal technique to analyse lon-

gitudinal data is using linear mixed effects modeling which accounts for both the within and between subject correlation structure through random effects.

The linear mixed model has design matrices for both fixed and random effects; fixed effects are essentially population level coefficients while random effects adjust for individual differences in response to an effect.

The random effects and random errors are assumed to be independently and identically distributed. Interpretation of the fixed effects is exactly the same as in the simple linear regression setting. This is interpreted as; assuming that there are covariates in the design matrix $\mathbf{X}$, the coefficients $\boldsymbol{\beta}$ for each covariate is the change in the responses $\mathbf{Y}$ on average for a one unit increase in the corresponding covariate $\mathbf{X}$ whilst adjusting for all other predictors in the model. Interpretation of the random effects $\mathbf{b}$ is from the view point of assessing how regression parameters for the i^{th} subject departs from the population. The chief advantage of mixed effects models is that besides estimating parameters that describe how the mean response changes in the population, it is also possible to predict subject-specific response profiles over time. This is the main reason for their use in the joint modeling for longitudinal and survival data framework (Rizopoulos 2012). In addition, these models handle data that is not balanced and random effects account for correlation between repeated measurements for each subject parsimoniously. These models assume that the longitudinal responses are independent conditional on random effects.

Models with few random effects may not sufficiently capture the correlation structure in the data, and, therefore, linear mixed model that allows for appropriate covariance matrix for the subject-specific error component that depends only on one subject are used. Several proposals have been made in literature on modeling the covariance matrix which results in different se-

rial correlation functions, with first order autoregressive, Gaussian and exponential structures being the most frequently used (Molenberghs and Verbeke 2000; Pinheiro and Bates 2000). Parameter estimation of linear mixed models is based on maximum likelihood principles, where the integral of log likelihood function of the observed data given full parameter vector has a closed form solution. This is a conditional distribution of longitudinal responses given random effects; which is normally distributed with mean and covariance matrix. The log likelihood estimator of the fixed effects vector after maximising the log likelihood function conditional on the covariance matrix can be solved and corresponds to ordinary least squares estimators. When sample sizes are small, maximising the log likelihood function for fixed effects gives biased results (Molenberghs and Verbeke 2000). To address this problem, the theory of restricted maximum likelihood (REML) developed by Harville (1974) has been widely adopted. The reasoning behind this estimation procedure is to separate the part of the data used for the estimation of the variance covariance matrix from the part used for the estimation of coefficients and this corrects for the fact that coefficients have also been estimated. Notably, neither the maximum likelihood nor the restricted maximum likelihood estimators can not be written in closed form; therefore, to obtain the estimate of the variance covariance matrix, numerical optimisation has to be done. The two popular algorithms that have had wide implementation are Expectation-Maximization (EM) (Dempster, Laird, and Rubin 1977) and Newton-Raphson algorithm (Lange and Hunter 2004).

2.1.2 The formulation of the linear mixed model

The linear mixed model which a powerful and flexible tool for the analysis of correlated data observed over time has the following formulation, using matrix notation:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon} \tag{2.1}$$

where $\mathbf{Y}$ is the $N \times 1$ vector of responses, $\mathbf{X}$ is the $N \times p$ design matrix for the fixed effects; $\boldsymbol{\beta}$ which is a $p \times 1$ vector of parameters and $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_b]$, where $\mathbf{Z}_i$ is the $p \times q$ design matrix for the i^{th} random effects factor $\mathbf{b} = [\mathbf{b}_i', \ldots, \mathbf{b}_b']$ with $\mathbf{b}_i$ being the $q_i \times 1$ random effects vector such that $q = \sum_{i=1}^{b} q_i$, and $\mathbf{e}$ is the $N \times 1$ vector of random errors. It is assumed that $\mathrm{E}(\mathbf{b}) = 0$ and $\mathrm{E}(\boldsymbol{\epsilon}) = 0$ in addition that $\mathbf{b}$ and $\mathbf{e}$ are independent and multivariate normally distributed as follows:

$$\begin{bmatrix} \mathbf{b} \\ \boldsymbol{\epsilon} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad \sigma^2 \begin{bmatrix} \mathbf{G}(\boldsymbol{\gamma}) & \mathbf{0} \\ \mathbf{0} & \mathbf{R}(\boldsymbol{\rho}) \end{bmatrix} \right), \tag{2.2}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\rho}$ are respectively $r \times 1$ and $s \times 1$ $\left(\text{with } s \leq \dfrac{n(n+1)}{2} \right)$ vectors of unknown variance parameters corresponding to $\mathbf{b}$ and $\boldsymbol{\epsilon}$. If the random terms are correlated then the dimensions of $\boldsymbol{\gamma}$ may exceed q i.e. $\boldsymbol{\gamma}$ may be of dimensions $r \leq \dfrac{q(q+1)}{2}$. The variance-covariance matrix of the data, $\mathbf{Y}$, can be written as follows using the formulation by Patterson and Thompson (1971):

$$\mathrm{Var}(\mathbf{Y}) = \sigma^2(\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}) = \sigma^2\mathbf{H} \tag{2.3}$$

where

$$\mathbf{H} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}. \tag{2.4}$$

The parametrisation above (2.3) is appeling because once the residual variance σ^2 is fac-

tored out the variance matrix for the data; it reduces to t-dimensional REML log-likelihood maximization problem to unity. In this case $t = r + s + 1$, which is the number of variance parameters in model (2.1).

The matrix $\mathbf{H}$ is composed of two items that are used to model correlation and heteroscedasticity i.e. the random effects $\mathbf{ZGZ}'$ and the within-group component $\mathbf{R}$. The within-group component $\mathbf{R}$ can sometimes be used to model directly the variance-covariance matrix of the data without the need to incorporate random effects to adjust for correlation among observations.

2.1.3 Joint estimation of fixed and random effects

In this Sub-Section we briefly review how parameters are estimated in the linear mixed model. Parameter estimation has mainly been done using least squares in the case of simple linear model, which postulates that a value that has the minimum variance is the best linear estimator. This principle can be extended to the linear mixed modeling framework. A lot of investigations have been done that have showed that this approach is sub-optimal (Laird and Ware 1982). The most widely estimation technique in linear models is based on maximising the likelihood on a log scale.

In the following Section we will first deal with the joint estimation of fixed effects $(\boldsymbol{\beta})$ and random effects $\mathbf{b}$ and then proceed with the estimation of variance parameters $\boldsymbol{\gamma}, \boldsymbol{\rho}$ and σ^2. There are many methods for estimating both fixed effects and random effects simultaneously, which include Henderson (1950)'s mixed model equations. One of the earlier proposals was by Henderson (1950) and Henderson et al. (1959), who assumed that $\mathbf{b}$ and $\mathbf{Y}$ are Gaussian

distributed as

$$\begin{bmatrix} \mathbf{b} \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{X}\boldsymbol{\beta} \end{bmatrix}, \quad \sigma^2 \begin{bmatrix} \mathbf{G} & \mathbf{G}\mathbf{Z}' \\ \mathbf{Z}\mathbf{G} & \mathbf{H} \end{bmatrix} \right). \tag{2.5}$$

In this case $\mathbf{Y}$ has the marginal probability distribution function $N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H})$, where $\mathbf{H}$ is as defined in model (2.4) with $\mathbf{G}$ and $\mathbf{R}$ assumed to be known.

In the above formulation (Henderson 1950; Henderson et al. 1959) maximised the log-joint distribution of $(\mathbf{Y}, \mathbf{b})$ to obtain estimators of $\boldsymbol{\beta}$ and $\mathbf{b}$. This logarithmic function is not a log-likelihood function since $\mathbf{b}$ is not observed. The marginal distribution $\mathbf{b}$ from 2.5 is thus

$$\mathbf{b} \sim \mathcal{N}(0, \mathbf{G})$$

and the conditional distribution of $\mathbf{Y}$ given $\mathbf{b}$ is

$$\mathbf{Y}|\mathbf{b} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \sigma^2\mathbf{R}).$$

Therefore the log-joint distribution of $(\mathbf{Y}, \mathbf{b})$ is given by

$$\begin{aligned}
\log f(\mathbf{Y}, \mathbf{b}) &= \log f(\mathbf{Y}|\mathbf{b}) + \log f(\mathbf{b}) \\
&= -\frac{1}{2}\Big\{ n\log\sigma^2 + \log\mathbf{R} + (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})/\sigma^2 \Big\} \\
&- \frac{1}{2}\Big\{ q\log\sigma^2 + \log\mathbf{G} + \mathbf{b}'\mathbf{G}^{-\mathbf{1}}\mathbf{b}/\sigma^2 \Big\}. \\
&= -\frac{1}{2}\Big\{ \log\sigma^2 + \log\mathbf{R} + \log\mathbf{G} + (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})/\sigma^2 \Big\} \\
&- \frac{1}{2\sigma^2}\Big\{ \mathbf{b}'(\mathbf{Z}\mathbf{R}^{-1}\mathbf{Z}' + \mathbf{G}^{-1})\mathbf{b} - 2(\mathbf{Y} - \mathbf{X}\mathbf{B})'\mathbf{R}^{-\mathbf{1}}\mathbf{Z}\mathbf{b} \Big\}.
\end{aligned}$$

Estimates of $\boldsymbol{\beta}$ and $\mathbf{b}$ are obtained by solving score equations

$$\mathbf{X}^{-1}\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\hat{\mathbf{b}}) = \mathbf{0}$$

$$\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) - (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\hat{\mathbf{b}} = \mathbf{0}.$$

These equations are commonly refereed to as mixed model equations (MMEs) proposed by Henderson (1950). The compact representation of these equations is of the following form;

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathrm{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}. \tag{2.6}$$

The above model (2.6) was re-written by Gilmour, Thompson, and Cullis (1995) as

$$\mathbf{C}\psi = \mathbf{W}'\mathbf{R}^{-\mathbf{1}}\mathbf{Y},$$

where $\mathbf{W} = [\mathbf{X} \quad \mathbf{Z}], \boldsymbol{\psi} = (\boldsymbol{\beta}', \mathrm{b}')'$ and $\mathbf{C} = \mathbf{W}'\mathbf{R}^{-1}\mathbf{W} + \mathbf{G}^{*+}$ with $\mathbf{G}^{*} = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G} \end{bmatrix}$ and $\mathbf{G}^{*+} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{bmatrix}$, where the '+ 'sign is the Moore-Penrose inverse.

In the model (2.1) there is $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{Var}(\mathbf{Y}) = \sigma^2\mathbf{H}$, where $\mathbf{H}$ is the conditional variance of the error term given $\mathbf{X}$ is a known nonsingular covariance matrix. If $\mathbf{H}$ is known, the fixed effects parameter $\boldsymbol{\beta}$ can be estimated using generalised least squares (GLS) to get

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{H}^{-1}\mathbf{Y}. \tag{2.7}$$

the best linear unbiased estimator of $\boldsymbol{\beta}$. Suppose that $\mathbf{X}$ is not full rank, the generalised inverse $(\mathbf{X}'\mathbf{H}^{-\mathbf{1}}\mathbf{X})^{-1}$ is used instead in the place of $(\mathbf{X}'\mathbf{H}^{-\mathbf{1}}\mathbf{X})^{-1}$ to get a solution for the fixed effects $\boldsymbol{\beta}$ which is not unique let alone unbiased. On the other hand $\mathbf{X}\hat{\boldsymbol{\beta}}$ is unique and unbiased for $\mathbf{X}\boldsymbol{\beta}$. Estimates for $\boldsymbol{\beta}$ and b are obtained by solving score equations

$$\mathbf{X}'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) - \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\tilde{\mathbf{b}} = 0 \tag{2.8}$$

$$\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} = 0 \tag{2.9}$$

These equations are called the mixed model equations (MMEs) as proposed by Henderson (1950) and Henderson et al. (1959).

Using GLS in estimating $\boldsymbol{\beta}$ is computer intensive since it requires the inverse of $\mathbf{H}$ which is the $n \times n$ matrix. The joint estimators of $\boldsymbol{\beta}$ and $\mathbf{b}$ can be obtained by solving (2.6) or

$$\hat{\boldsymbol{\psi}} = \mathbf{C}^{-1}\mathbf{W}'\mathbf{R}^{-1}\mathbf{Y}, \tag{2.10}$$

where $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\beta}}', \hat{\mathbf{b}})'$. It must be noted that equation (2.10) just requires the inversion of $\mathbf{C}$, a $(p+q) \times (p+q)$ matrix which is less computer intensive compared to $\mathbf{H}$ and $\mathbf{R}^{-1}$ which is also an $n \times n$ matrix which is computationally efficient.

2.1.4 Variance parameter estimation

There have been several approaches discussed by Searle, Casella, and McCulloch (1992) for estimating variance components. Some of the approaches include analysis of variance (ANOVA) which are optimal in balanced data settings only. This techniques uses the expected mean squares which is difficult to apply in unbalanced data or modeling variation using complex variance structures. A general discussion about challenges encountered by the above approach for estimating variance components has been provided by Searle (1995) for unbalanced data.

There has been a lot of interest in the estimation of variance components in the past (Henderson 1953; Rao 1971). In his earlier work Henderson (1953), a three technique to estimating variance components was proposed and these are called Henderson I, II and III. There is a detailed account provided by Searle, Casella, and McCulloch (1992) about these techniques.

After his investigation Rao (1971) proposed a minimum norm quadratic estimation (MINQUE) method for estimating variance components which results in estimates that are translation invariant in cases where there is unbiased quadratic forms of observations.

In recent times the standard methods for estimating variance components are based on max-

imum likelihood (ML) and residual maximum likelihood (REML) for both balanced and unbalanced data. These estimation methods besides working for both balanced and unbalanced data, they handle a wide range variance models compared to simple variance components.

It has been demonstrated that maximum likelihood variance estimators are biased downwards mostly in small sample sizes. This is the case because no adjustment is done to estimates due to the degrees of freedom after estimating fixed effects (Lin and McAllister 1984; Swallow and Monahan 1984). It is for this reason that REML estimation techniques for variance components are a preferred choice. There is an elaborate discussion provided by Searle, Casella, and McCulloch (1992) about advantages and disadvantages between ML and REML. Maximum likelihood estimation for variance components alone has been an attractive research area in the past four decades or so (Hartley and Rao 1967; Jennrich and Sampson 1976; Lindstrom and Bates 1988; Molenberghs and Verbeke 2000).

In the next Section a review of variance components estimation is presented which is limited to REML methods.

2.1.5 Residual maximum likelihood (REML)

REML maximises the likelihood of linearly independent error contrasts which hide in $\mathbf{H}$ as presented in equation (2.7) when maximum likelihood estimation of variance components is done (Anderson and Bancroft 1952; Patterson and Thompson 1971). These error contrasts are a linear combination of the data $\mathbf{Y}$ which is orthogonal to the design matrix $\mathbf{X}$. In REML the linear combination are chosen as $\mathbf{K}'\mathbf{Y}$ so that $\mathbf{K}'\mathbf{Y}$ is full rank but not linked to fixed effects $\boldsymbol{\beta}$. These are residuals obtained after fitting fixed effects ($\boldsymbol{\beta}'$s) hence $\mathrm{E}(\mathbf{K}'\mathbf{Y}) = 0$ if and only if $\mathbf{K}'\mathbf{X} = 0$ holds. This quantity suggests performing maximum likelihood on $\mathbf{K}'\mathbf{Y}$ instead of

$\mathbf{Y}$.

In their work Lee and Der Werf (2006) have shown that REML log-likelihood can be viewed as a conditional likelihood by assuming asymptotic Gaussian distribution for $\boldsymbol{\beta}$ given variance components values while Cook (1987) showed that the REML log-likelihood coincides with the conditional profile likelihood for $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{H})$ and $\mathbf{K}'\mathbf{X} = 0$ we have $\mathbf{K}'\mathbf{Y} \sim \mathcal{N}(0, \sigma^2\mathbf{K}'\mathbf{H}\mathbf{K})$ and the residual REML log-likelihood function is

$$\begin{aligned} \ell_R(\boldsymbol{\phi}; \mathbf{K}'\mathbf{Y}) = -\frac{1}{2}\Big\{ & (n-p)\log(2\pi) + (n-p)\log\sigma^2 + \log|\mathbf{K}'\mathbf{H}\mathbf{K}| \\ & + \frac{1}{\sigma^2}\mathbf{Y}'\mathbf{K}(\mathbf{K}'\mathbf{H})^{-\mathbf{1}}\mathbf{K}'\mathbf{Y}, \Big\} \end{aligned} \tag{2.11}$$

where $\boldsymbol{\phi} = (\boldsymbol{\kappa}'\boldsymbol{\sigma}^2)'$, and $\boldsymbol{\kappa} = (\boldsymbol{\gamma}', \boldsymbol{\rho})'$. The derivation of the probability distribution of $\mathbf{K}'\mathbf{Y}$ having chosen carefully $\mathbf{K}'$ as an $(n-p) \times n$ matrix whose rows are $(n-p)$ linear independent rows $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$ has been studied by Patterson and Thompson (1971). Due to the fact that $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$ is symmetric, idempotent with rank $n-p$ it can be written as $\mathbf{K}\mathbf{K}'$ such that $\mathbf{K}'\mathbf{K} = \mathbf{I}$. In their argument Patterson and Thompson (1971) say that since $\mathrm{E}(\mathbf{K}'\mathbf{Y}) = 0$ then $\mathbf{K}'\mathbf{Y}$ lies in the error space where there is no information about fixed effects but does have information about variance components.

The constants are ignored and the REML log-likelihood for the model is thus

$$\begin{aligned} \ell_R(\boldsymbol{\phi}; \mathbf{Y}) = -\frac{1}{2}\Big\{ & (n-p)\log\sigma^2 + \log|\mathbf{H}| + \log|\mathbf{X}'\mathbf{H}^{-\mathbf{1}}\mathbf{X}| \qquad (2.12) \\ & + \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{H}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{\sigma^2} \Big\} \\ = -\frac{1}{2}\Big\{ & (n-p)\log\sigma^2 + \log|\mathbf{H}| + \log|\mathbf{X}'\mathbf{H}^{-\mathbf{1}}\mathbf{X}| + \frac{\mathbf{Y}'\mathbf{P}\mathbf{Y}}{\sigma^2} \qquad (2.13) \end{aligned}$$

where $\hat{\beta}$ is the GLS estimate of $\boldsymbol{\beta}$ and $\mathbf{P} = \mathbf{H}^{-\mathbf{1}} - \mathbf{H}^{-\mathbf{1}}\mathbf{X}(\mathbf{X}'\mathbf{H}^{-\mathbf{1}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{H}^{-\mathbf{1}}$. It has been shown by Khatri (1966) and Searle, Casella, and McCulloch (1992) that if $\mathbf{K}'\mathbf{Y} = 0$ where $\mathbf{K}'$

has maximum row rank, and $\mathbf{H}$ positive definite, then

$$\mathbf{K}(\mathbf{K}'\mathbf{H}^{-1}\mathbf{K})^{-1}\mathbf{K}' = \mathbf{P},$$

so that (2.11) and (2.13) are identical.

According to Gilmour, Thompson, and Cullis (1995) taking the derivative of REML log-likelihood function of (2.13) with respect to σ^2 and κ_j, for $j = 1, ..., r+s$ gives

$$\frac{\partial \ell_R(\boldsymbol{\phi}; \boldsymbol{Y})}{\partial \sigma^2} = -\frac{n-p}{2\sigma^2} + \frac{\mathbf{Y}'\mathbf{P}\mathbf{Y}}{2\sigma^4} \tag{2.14a}$$

$$\frac{\partial \ell_R(\boldsymbol{\phi}; \boldsymbol{Y})}{\partial \kappa_j} = -\frac{1}{2}\Big\{\text{tr}(\mathbf{P}\mathbf{H}_i) - \frac{1}{\sigma^2}\mathbf{P}\mathbf{H}_i\mathbf{P}\mathbf{Y}\Big\}. \tag{2.14b}$$

Setting equation (2.14a) equal to zero and solving gives the REML estimator for the error variance as

$$\hat{\sigma^2} = \frac{\mathbf{Y}'\hat{\mathbf{P}}\mathbf{Y}}{n-p}, \tag{2.15}$$

which is computed iteratively because it depends on $\hat{\boldsymbol{\kappa}}$ through $\boldsymbol{P}$. The REML estimate for κ is also computed iteratively. The iterative scheme for obtaining REML estimates based on variance components has been given by Searle, Casella, and McCulloch (1992).

2.2 Model checking and diagnostics for linear mixed model

This Sub-Section reviews existing model checking and diagnostic tools for linear mixed models using standard model checking techniques which include residual analysis and influence diagnostics to mention but a few. In general statistical models are assessed for model fit through a number of ways which include, but not limited to, visual check of the estimated and sample means over time, comparing the estimated and sample covariance structure, statistical comparison of the selected model with higher order (saturated) models and diagnostics based on analysis of residuals which is basically the overall mis-specification relative to the data

itself as opposed to model comparison.

It is conventional practice during statistical analysis to examine residuals which help in assessing model fit and/or adequacy. A residual is defined as the difference between the observed value and the value predicted by the model. There are several methods that exist in literature that have been proposed to assess linear mixed models based on residual analysis.

In the simple case of a linear regression these residuals are uncorrelated with predicted values and with covariate. Plots of these residuals against predicted values and against predictor variables should appear random, symmetric about zero, with constant variance and no evidence of outliers for a good model.

For data that is correlated a vector of residuals for each subject can be defined;

$$r_i = Y_i - X_i'\hat{\beta}. \tag{2.16}$$

This vector of residual has mean zero and provides an estimate of the of errors,

$$e_i = Y - X_{ij}'\hat{\beta}. \tag{2.17}$$

In the longitudinal data setting, residuals in linear mixed models are vectors which can be plotted against the predicted mean response. For example a scatter plot of residuals,

$$r_{ij} = Y_{ij} - X_{ij}'\hat{\beta}. \tag{2.18}$$

against the predicted mean response

$$\hat{\mu}_{ij} = X_{ij}'\hat{\beta}, \tag{2.19}$$

can be examined for the appearance of a systematic trend. If the mean model fits the data well, plots of these residuals against the predicted mean response should not show systematic trend about the constant mean of zero, the plot should also be useful for detecting outliers. This should also be true for scatter-plots of residuals against selected covariates for the mean model

(Fitzmaurice, Laird, and Ware 2012, Chap.5.3.4) and (Diggle 2002, Chap. 10). The fitting of a smooth curve (e.g. a lowess curve) to the scatter-plot often helps in judging if curvature exists. The problem with these residuals is that they may not display standard properties, which is constant variance due to correlation. The covariance of these residuals is not identical to the covariance of (true) errors.

$$Cov(\mathbf{r_i}) \approx Cov(\mathbf{e_i}) = \Sigma_i. \tag{2.20}$$

As a result of this, a plot of residuals versus predicted values may not show homogeneity of the variance and it can further be shown that residuals and covariates may be correlated hence plots may show systematic trend.

To work around this problem, these residuals are transformed so that they mimic residuals from the standard linear regression, effectively forming residuals that have constant variance and zero correlation. The transformation is implemented through the Cholesky decomposition of the lower triangular matrix which renders these residuals to be uncorrelated and have variance that is unity. In a similar way, predicted values can be transformed giving means equivalent to those used in a standard linear regression.

Plots of both decomposed residuals against predicted values should show a random symmetric scatter around mean zero with constant variance and no outliers, the same applies to plots of these residuals at observation level against transformed covariate values. These plots are visually useful if a lowess curve is incorporated for ease of evaluating systematic trend.

2.2.1 Case-deletion diagnostics for all parameters

This Section reviews some existing case deletion diagnostics for linear mixed effects models. Measures of influence are the target of this Section and case deletion diagnostics are reviewed using variance components and fixed effects. These have been chosen because they are based on Cook's distance.

Extensive research has been done on outliers in linear mixed models to identify influential observations. The developed techniques lead to case-deletion methods proposed by Cook (1977), Cook and Weisberg (1982). Application of these ideas have also been adopted in the analysis of time-to-event data (Hall, Rogers, and Pregibon 1982; Storer and Crowley 1985) as well as in logistic regression (Hosmer, Jovanovic, and Lemeshow 1989; Pregibon 1981).

Case-deletion diagnostics assumes $\boldsymbol{\theta}$ to be a parameter vector of interest in the model under consideration and $\hat{\boldsymbol{\theta}}$ an estimate of $\boldsymbol{\theta}$ using appropriate techniques usually maximum likelihood in the frequentist paradigm (Cook and Weisberg 1982). If one excludes the i^{th} observation from the full set of observations; refitting the model under the same principle produces $\hat{\boldsymbol{\theta}}_{[-i]}$ of $\boldsymbol{\theta}$, which gives a measure of the difference between $\hat{\boldsymbol{\theta}}_{[-i]}$ and $\hat{\boldsymbol{\theta}}$, which is used to evaluate how the i^{th} observation impacts the estimate $\hat{\boldsymbol{\theta}}$ and subsequent inferences drawn from the model and this measure is called the Cook's distance.

There has been a lot of effort to investigate influence diagnostics in linear mixed models. In their work, Demidenko and Stukel (2005) extended a number of regression diagnostic approaches that are in use in linear regression to linear mixed effects model such as case-deletion diagnostics. Their techniques included explicitly defined functions that do not require re-estimation of the model. They adopted Cook's statistics based on a generalization of Miller's

simple update formula for case-deletion in linear models.

Using a slightly different technique, Zewotir and Galpin (2005) developed a routine diagnostic tools, which is not computer intensive. These diagnostics are derived from studentized residuals, error contrast matrix, and the inverse of the response variable covariance matrix. The computation of basic building blocks is done only once from the complete data analysis and provide information on the influence of the data on different aspects of the model fit.

On their part, Pan, Fei, and Foster (2014) developed a case-deletion diagnostics tool using Q-function, which is a conditional expectation of the logarithm of the joint-likelihood between random effects and an outcome of interest. The diagnostics tool is able to identify influential subjects and influential observations in linear mixed models.

The downside of case-deletion diagnostics is that, they are often times affected by swamping and masking. When an observation is wrongly identified as influential as a result of the presence of another observation it is generally referred to as swamped per description by Barnett and Lewis (1995); on the other hand, if an observation that is influential does not get detected it is referred to as masked (Atkinson 1985; Chatterjee and Hadi 1988). To avoid the masking effect, Rousseeuw and Van Zomeren (1990) compute distances based on robust estimates of location and covariance. These robust distances efficiently expose outliers which then circumvents masking.

One of the common approaches in model assessments is case-deletion (Cook and Weisberg 1982). If we let σ_b^2 to be the variance component due to random effects and σ_ϵ^2 to be the random error and hence consider a full parameter vector $\boldsymbol{\theta}' = (\boldsymbol{\beta}, \sigma_b^2, \sigma_\epsilon^2)'$ using diagnostic measures and individual parameter $\boldsymbol{\beta}, \sigma_b^2, \sigma_\epsilon^2$ in model (2.1), one can calculate Cook's distances using these parameters. The components of the vector parameter space under case-deletion investi-

gation include fixed effects ($\boldsymbol{\beta}$) and variance components (σ_b^2 and σ_ϵ^2). Separate statistics are computed for the fixed effects and the covariance parameters. The Cook's statistic is presented as a case-deletion measure, which is used to identify influential subjects.

Let's us suppose that $\boldsymbol{\theta}$ is a parameter vector of interest for fixed effects in the model and $\hat{\boldsymbol{\theta}}$ is a likelihood based estimate of $\boldsymbol{\theta}$ (MLE or REML). If the i^{th} subject or observation is removed from the full set $\mathbf{Y} = Y_1, Y_2, ..., Y_n$ refitting the model using the exact estimation principles as before results in the estimate $\hat{\boldsymbol{\theta}}_{[-i]}$ of $\boldsymbol{\theta}$. A measure of the difference between $\hat{\boldsymbol{\theta}}_{[-i]}$ and $\hat{\boldsymbol{\theta}}$ can be used to quantify how the i^{th} subject influences the estimate $\hat{\boldsymbol{\theta}}$ and its subsequent inferences. The Cook's distance defined below is a typical example;

$$CD(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}})' \mathbf{M}^{-1} (\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}}), \tag{2.21}$$

where $\mathbf{M}$ is the covariance matrix for all parameter estimates of $\hat{\boldsymbol{\theta}}$ or $\hat{\boldsymbol{\theta}}_{[-i]}$. In the simple regression setting, $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ with $\mathrm{Var}(\mathbf{Y}) = \sigma^2\mathbf{H}$, where $\mathbf{H}$ is the known covariance matrix, leverage $\mathbf{H}_i = \frac{\partial \hat{\mathbf{Y}}_i}{\partial \mathbf{Y}_i} = \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i$ where $\hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}'$, the matrix $\mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$ is often referred to as the hat matrix.

The case-deletion diagnostic measure for regression coefficients β can be formed by taking $\mathbf{M} = \mathbf{X}'\mathbf{H}^{-1}\mathbf{X}$ or $\mathbf{M} = \mathbf{X}'_{[-\mathbf{i}]}\mathbf{H}^{-1}_{[\mathbf{i}]}\mathbf{X}_{[-\mathbf{i}]}$. In this case $\mathbf{X}_{[-i]}$ and $\mathbf{H}_{[-i]}$ are matrices $\mathbf{X}$ and $\mathbf{H}$ after deleting the i^{th} subject. This leads to the Cook's statistic, which is extended to the linear mixed model, where the estimation of $\boldsymbol{\beta}$ is likelihood based and $\mathbf{H} = \mathbf{Z}_i'\mathbf{G}\mathbf{Z}_i + \boldsymbol{R}_i$.

The same model is adopted for evaluating Cook's statistics for fixed effects parameters as follows

$$CD(\boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}}_{[-i]} - \hat{\boldsymbol{\beta}})' \mathbf{M}_1^{-1} (\hat{\boldsymbol{\beta}}_{[-i]} - \hat{\boldsymbol{\beta}}), \tag{2.22}$$

where $\mathbf{M}_1$ is the variance covariance matrix for fixed effects.

2.2.2 Case-deletion diagnostics for variance components

Using the Cook's statistic in our diagnostics approach based on model (2.1) above, proposed by Christensen, Pearson, and Johnson (1992) and Cook (1977). Given a full parameter vector $\boldsymbol{\theta}' = (\boldsymbol{\beta}, \sigma_b, \sigma_\epsilon)'$ using the diagnostics measures and individual parameter σ_b^2 and σ_ϵ^2 from this model. The variance components of the vector parameter space under case-deletion investigation are (σ_b^2 and σ_ϵ^2).

The case-deletion diagnostics based on the Cook's statistic for the variance components is formulated as follows:

$$CD(\boldsymbol{\gamma}) = (\hat{\boldsymbol{\gamma}}_{[-i]} - \hat{\boldsymbol{\gamma}})' \mathbf{M}_2^{-1} (\hat{\boldsymbol{\gamma}}_{[-i]} - \hat{\boldsymbol{\gamma}}), \tag{2.23}$$

where $\mathbf{M}_2$ is the covariance of the vector of variance components while $\hat{\boldsymbol{\gamma}}' = (\sigma_b^2, \sigma_{\epsilon)}^2$ is the variance component vector from the linear mixed model and $\hat{\boldsymbol{\gamma}}_{[-i]}$ is the variance component vector after excluding the i^{th} subject from the model. Large values of $CD(\boldsymbol{\gamma})$ indicate that the i^{th} subject needs to be investigated for being a potential outlier or exerting influence on the estimation of parameters as shown by Zewotir and Galpin (2005).

2.3 Review of the survival model

This Section introduces the standard Cox survival model, which is widely used in the analysis of time-to-event data. This, then, is extended to the model with time-varying covariates (Gill, Keiding, and Andersen 1993) that are often encountered in longitudinal studies investigating the relationship between a bio-marker and risk of an event. There is extensive literature on estimation, inference and diagnostics on these models which will be reviewed.

In epidemiology and clinical trials, the main aim of the study is to analyse the time until an event of interest occurs. The dependent variable under such circumstances is the time to the occurrence of an event often referred to as survival time. The original idea of survival data was to estimate the probability of surviving, including making simple group comparisons of survival distributions under varied conditions. This idea has evolved with time to statistical modeling (regression) of survival data, which offers three basic advantages;

- simultaneously analysing multiple independent prognostic risk factors.
- assessing treatment differences while controlling for baseline characteristics.
- better calibration of predictions can be made.

The Cox model has been popular in the analysis of survival data including extensions to handling time-dependent covariates. The underlying assumptions of the model are that, covariates have a multiplicative effect on the risk (hazard) of an event (Cox 1972a; Cox 1972b).

2.3.1 Estimation of parameters in censored data

If we let T^* to be the random variable of survival times under consideration, the distribution of T^* is described by a survival function, which in essence expresses the probability that an event occurs after time t. The probability density can be defined of surviving up to time t. A description of an instantaneous risk of the event in the interval $[t, t+dt]$ is the hazard function which is an important function and has a central role because it has a unique relationship with the survival function to be described in the following Sub-Section. A cumulative hazard function describing the accumulated risk up until time t is interpreted as the expected number of events observed by time t.

Censoring is an important concept in survival analysis, which means that survival times of some individuals might not be fully observed due to a number of reasons. In clinical trials, this might happen when the study stops before the full survival times of all individuals can be observed, or a person withdrawals consent for further participation in the study, or for studies that have long follow up time, participants have the potential of not being traced and hence lost to follow up. In such cases, the study participants survives beyond the time of the study, and the exact survival time is unknown. This is called right censoring.

There are, therefore, two possibilities in survival studies either the participant is observed to fail at time T, or the observation on that participants ceases at time C. Then the observation is min(T, C) and an indicator variable I shows if the individual is censored or not. The calculations for hazard and survival functions must be adjusted to account for censoring.

When the survival function is assumed to be of parametric form, estimation of parameters of interest is based on maximum likelihood methods. Since censoring has to be accounted for in survival analysis, the construction of the likelihood function proceeds by combining information from the censored as well as un-censored observations. The formulation of the log-likelihood must have the characteristic that all subjects contribute an amount to the log-likelihood equal to the negative of the cumulative hazard function evaluated at their observed event time T_i. Once the log-likelihood has been formulated, iterative optimisation procedures, such as Newton-Raphson algorithm (Lange and Hunter 2004) can be used to locate the maximum likelihood estimate for $\hat{\theta}$. Inference proceeds under classical asymptotic maximum likelihood theory (Cox and Hinkley 1974).

Extensions to the standard Cox proportional hazards model have been investigated, these particularly handle “endogenous (internal) time varying covariates using the counting process

formulation", which is a relative risk model with longitudinal history (Andersen and D 1982; Fleming and Harrington 1991; Gill, Keiding, and Andersen 1993). Parameters are estimated by maximising the partial log-likelihood function through iterative procedures, just like in the previous Section. The interpretation for regression coefficients are the same as those for standard survival setting. Special mention has to be made of the association parameter, which is interpreted as the relative increase in the risk of an event at time t that results from one unit increase in the time varying covariate denoted $\mathrm{Y}_i(t)$ at the same time point.

2.4 The Cox model and extended Cox model

Introduction

In this Section, we review the standard time-to-event model proposed by Cox (1972b), which will include model formulation and parameter estimation. Extensions to the Cox proportional hazards model, which handle time-varying covariates, will be reviewed along in this Section and we also review diagnostic approaches that exist for these models. This will lay the ground for developing case deletion diagnostics in the joint modeling framework.

2.5 The standard Cox model

We review the Cox model, which is popular in the analysis of survival data including extensions to handling time-dependent covariates. The Cox proportional hazards model (Cox 1972b) is as follows:

$$\log\left(\frac{h_i(t)}{h_{0i}(t)}\right) = \mathbf{X}_i'\boldsymbol{\beta},$$

where $h_i(t)$ denotes the hazard function for the i^{th} subject at time t, $h_{0i}(t)$ denotes the baseline hazard function and $\mathbf{X}$ is a vector of covariates, while $\boldsymbol{\beta}$ is a vector of log hazard ratios.

2.5.1 Estimation of parameters in censored data

If we let T^* to be the random variable of survival times under consideration, the probability of surviving up to time t is given by:

$$S(t) = P(T^* > t) = \int_t^\infty p(s)ds, \tag{2.24}$$

where $p(.)$ is the probability density function.

The description of an instantaneous risk of the event in the interval $[t, t+dt]$ is the hazard function;

$$h(t) = \lim_{dt \to 0} \frac{Pr(t \leq T^* < t+dt | T^* \geq t)}{dt}, t > 0$$

and it relates to the survival function as follows:

$$S(t) = \exp\{-H(t)\} = \exp\{-\int_0^t h(s)ds\}, \tag{2.25}$$

where $H(.)$ is the cumulative hazard function describing the accumulated risk up until time t. The interpretation of $H(t)$ would be, the expected number of events observed by time t.

2.6 Influence diagnostics for the Cox model

Case-deletion diagnostics in the Cox model are not the same as in the linear mixed model because the likelihood function for the Cox regression model is not an expression of the sum of a number of terms. In linear mixed model, these terms contribute to the log-likelihood made by each observation. In the Cox model, excluding an observation in the model affects

the risk set over, which quantities of the form $\exp(\mathbf{X}_i\boldsymbol{\beta}')$ are summed. This therefore renders influence diagnostics for the Cox model difficult to derive (Collett 2015).

2.6.1 Delta betas

Informally, case-deletion diagnostics identifies influential observation(s) by iteratively excluding an i^{th} observation $(n-1)$ in turn during the estimation process. The effect of excluding an observation in parameter estimation can then be calculated. It has to be noted that, this procedure is computer intensive when the number of observation is large, in which case an approximation to the degree by which $\boldsymbol{\beta}_j$ changes when the i^{th} observation is excluded for $i = 1, 2, \ldots, n$ is employed. If we denote the value of the j^{th} parameter estimate after excluding the i^{th} observation as $\hat{\beta}_{(i)}$; Cain and Lange (1984) showed that an approximation to $\hat{\beta} - \hat{\beta}_{(i)}$ is based on score residuals (2.26).

Score residuals are used to identify influential subjects obtained by differentiating the logarithm of the partial likelihood function with respect to the parameters β_j, $j = 1, 2, ..., p$ (Collett 2015). The i^{th} score residual, $i = 1, 2, ..., n$ for the p^{th} explanatory variable in the model, X_p is defined as

$$r_{S_{pi}} = \delta_i\Big(x_{pi} - \hat{a}_{pi}\Big) + \exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}') \sum_{tr \le t_i} \frac{(\hat{a}_{pr} - x_{pi})\delta_r}{\sum_{l \in R(t_r)} \exp(\mathbf{X}_l\hat{\boldsymbol{\beta}}')}, \tag{2.26}$$

where x_{pi} is the i^{th} value of the p^{th} explanatory variable, δ_{r} is the event indicator, R_{tr} is the risk set at time t_{r} and $a_{\text{pr}} = \dfrac{\sum_l x_{pl}\exp(\boldsymbol{\beta}'\mathbf{x}_l)}{\sum_l \exp(\boldsymbol{\beta}'\mathbf{x}_l)}$. Let $\mathbf{r}_{s_i}$ be the vector of values of the score residuals for the i^{th} observation, so that $\mathbf{r}'_{s_i} = (r_{s_{1i}}, r_{s_{2i}}, \ldots, r_{s_{pi}})$, where $r_{s_{ji}}, j = 1, 2, \ldots, p$ is the i^{th} score residual defined as follows

$$r_{s_{ji}} = r_{p_{ji}} + \exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}') \sum_{tr \le t_i} \frac{(\hat{a}_{jr} - x_{ji})\delta_r}{\sum_{l \in R(t_r)} \exp(\mathbf{X}_l\hat{\boldsymbol{\beta}}')}. \tag{2.27}$$

An approximation to $\hat{\beta} - \hat{\beta}_{(i)}$, the change in $\hat{\beta}$ after excluding the i^{th} observation is the i^{th} component of the vector

$$\mathbf{r}_{s_i}\mathrm{Var}(\hat{\boldsymbol{\beta}})'$$

where $\mathrm{Var}(\hat{\boldsymbol{\beta}}$ is the variance covariance matrix of the vector of parameter estimates in the fitted Cox regression model. This i^{th} element of the said vector is referred to as delta-beta, denoted as $\Delta_i\hat{\beta}_j$, so that $\Delta_i\hat{\beta}_j \approx \hat{\beta} - \hat{\beta}_{(i)}$. Using this approximation means that values of $\Delta_i\hat{\beta}_j$ can be computed from quantities available after fitting the model to the full data-set. It can, therefore, be seen that observations that influence a particular parameter estimate are those whose values of delta-beta are large in absolute terms compared to the rest of in the data-set. Using index plots of these delta-betas for predictor variables, observations that are influential can be revealed. Plots of $\Delta_i\hat{\beta}_j$, against rank order of survival times results in information about the relationship between survival times and influence.

Use of the statistic $\Delta_i\hat{\beta}_j$ as an approximation to the actual change in the parameter after excluding the i^{th} observation from the model fit, is therefore, appropriate in this setting, this is so because observations that are considered to be influential on parameter estimates are highlighted. On the other hand, a careful scrutiny of model based inferences after excluding an observation from the fit are needed to examine the actual effect.

2.6.2 Infinitesimal jackknife statistic

In their work Cain and Lange (1984) defined case influence of subject j up on $\hat{\boldsymbol{\beta}}$ as $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}_{(j)}}$. We refer to subject as to opposed observation for consistency. This quantity, $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}_{(j)}}$, is calculated by excluding subject j and refitting the model to the data. This approach is known to be computationally intensive and thus can be estimated by a one-step approximation using

Newton-Raphson algorithm. An alternative that has been provided by Cain and Lange (1984) is the use of the first-order approximation to $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}_{(j)}}$ using weighted analysis, with subject j given weight w_j. They have demonstrated how the weights are incorporated in the likelihood function. If all subjects have, say, weight unity w_j, then considering $\hat{\boldsymbol{\beta}}$ as a function of w_j, that is, $\hat{\boldsymbol{\beta}}(w_j)$, it is noted that $\hat{\boldsymbol{\beta}}(1) = \hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}(0) = \hat{\boldsymbol{\beta}}_{(j)}$. The approximation of $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}_{(j)}}$ based on Taylor series expansion about $w_j = 1$, is thus, $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}_{(j)}} \approxeq \partial\hat{\beta}/\partial w_{(j)}$, where the derivative is evaluated at $w_{(j)} = 1$. This approximation is the same as that used in the infinitesimal jackknife statistic (Efron 1981; Miller Jr 1974) and is the estimate of case influence in that, when $w_{(j)} = 0$ effectively excludes subject j from the estimation process.

2.6.3 Perturbed / Augmented regression

Another diagnostic based on principles of augmented regression was developed by Storer and Crowley (1985). In this piece of work, Storer and Crowley (1985), derived score and information matrix requiring minor modification to accommodate more general relative-risk functions.

The score vector which they derived with respect to β is;

$$\mathbf{S}_{\mathrm{k}} = \partial\ell_{\mathrm{k}}/\partial\beta = \mathbf{X}'_{\mathrm{k}}\Big(\mathbf{y}_{\mathrm{k}} - \mathbf{p}_{\mathrm{k}}\Big), \tag{2.28}$$

where $\ell_{\mathrm{k}} = \log(L_{\mathrm{k}}) = \mathbf{Y}'_{\mathrm{k}}\theta_{\mathrm{k}} - \mathbf{a}(\theta_{\mathrm{k}})$. The vector $\mathbf{p}_{\mathrm{k}}$ is $p_{1i}, \ldots, p_{in_{\mathrm{k}}}$ and

$$p_i = \frac{\partial \mathbf{a}(\theta_{\mathrm{k}})}{\partial\theta_i} = \frac{\displaystyle\sum_{c(i)\in R_k}\left(\prod_{l=c_1}^{c_{m_k}}\exp(\theta_l)\right)}{\displaystyle\sum_{c\in R_k}\left(\prod_{l=c_1}^{c_{m_k}}\exp(\theta_l)\right)}. \tag{2.29}$$

The summation over $c(i)$ is the overall combination of m_k labels from R_k that contain the label i. The interpretation of p_i in a way, the conditional expectation of y_i, that is, the probability

of $y_i = 1$ given the covariate configuration of all subjects in the stratum and the fact that $\mathbf{y}'\mathbf{y} = m_k$. In the survival context p_i, is a function of t_k.

The resulting information matrix derived from equation (2.28) has the following formulation,

$$-\left(\partial^2 l_k/\partial\boldsymbol{\beta}^2\right) = \mathbf{X}_k'\mathbf{U}_k\mathbf{X}_k, \tag{2.30}$$

where $\mathbf{U}_k = \bar{a}\theta_k$ is a matrix with (i, j, k) entry

$$\frac{\partial^2 a(\theta_k)}{\partial\theta_i\partial\theta_j} = \frac{\sum\limits_{c(i,j)\in R_k}\left(\prod\limits_{l=c_1}^{c_{m_k}}\exp(\theta_l)\right)}{\sum\limits_{c\in R_k}\left(\prod\limits_{l=c_1}^{c_{m_k}}\exp(\theta_l)\right)} - \frac{\partial\mathbf{a}(\boldsymbol{\theta}_k)}{\partial\theta_i}\times\frac{\partial\mathbf{a}(\boldsymbol{\theta}_k)}{\partial\theta_j} \tag{2.31}$$

$$= p_{ij} - p_i \times p_j. \tag{2.32}$$

As earlier said these derivations require minor modification to accommodate more general relative risk functions of the form $f(X_i'\beta)$. The analytical expressions of the score and information matrix are thus, $\mathbf{S}(\boldsymbol{\beta}) = \bar{\mathbf{X}}'(\mathbf{y} - \mathbf{p})$ and $\mathbf{I}(\boldsymbol{\beta}) = \bar{\mathbf{X}}'\mathbf{U}\mathbf{X} - \mathbf{Q}$ respectively, where $\bar{\mathbf{X}} = {}^{\partial\boldsymbol{\theta}}/_{\partial\boldsymbol{\beta}}$ and $\mathbf{Q} = \sum(y_i - p_i) \times {}^{\partial^2\boldsymbol{\theta}}/_{\partial\boldsymbol{\beta}^2}$ (Storer and Crowley 1985).

2.6.4 Likelihood displacement

There are times when the structure of the fitted model is sensitive to at least one observation in the data set. In this case, it seems reasonable to identify such observations that exert influence using a diagnostic tool that is designed to pick these observations given a complete set of parameter estimates in the linear predictor. Such diagnostics have the capability to reflect influence that individual observations have on the risk score. This is, thus, an extension on information provided by delta betas (Collett 2015).

It is worth noting, that excluding a given observation from the data set may not be noticeable on

some parameter estimates, as such, delta betas can not pick them as influential. An assessment of changes in a set of estimated parameters might well be that, a form of the estimated hazard function or values of summary statistics given the fitted model change markedly when that observation is omitted from the model fitting. The other advantage for statistics that assess influential observations on a set of parameter estimates is that, a single value of the diagnostic is for each observation, which makes them easier to use than diagnostics like delta betas.

We had considered two diagnostics in this review for assessing influence primarily because of their mathematical derivation and analogy with the linear mixed model diagnostics. This, therefore, fits in well with implementation and development of diagnostics in the joint models for longitudinal and survival data. Firstly, a commonly used approach for detecting influential observations on the overall fit of the model that has been developed makes use of examining the amount of change by which the value minus twice the logarithm of the maximised partial likelihood, $2\log\hat{L}$ under the fitted model changes when each observation is in turn excluded from the model fitting.

Let $-2\log L(\hat{\boldsymbol{\beta}})$ be the value of the maximised log-likelihood when the model is fitted to the full data set and $-2\log L(\hat{\boldsymbol{\beta}}_{(i)})$ the value of the maximised log-likelihood after excluding an i^{th} observation from the full data set during the estimation process, which, therefore, has the following diagnostic.

$$-2\{\log L(\hat{\boldsymbol{\beta}}) - \log L(\hat{\boldsymbol{\beta}}_{(i)})\}$$

which is appropriate to study the influence on the overall fit of the model.

The likelihood displacement case-deletion diagnostic is defined as:

$$\text{LD}_i = 2\left[\log L(\hat{\boldsymbol{\beta}}) - \log L(\hat{\boldsymbol{\beta}}_{[-i]})\right], \tag{2.33}$$

where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate for all subjects or observations fitted in the model and $\hat{\boldsymbol{\beta}}_{[-i]}$ is the maximum likelihood estimate with the i^{th} subject deleted. This calculates how much the log-likelihood would change after excluding subject i in the model (Escobar and Meeker Jr 1992). There has also been an investigation by Pettitt and Daud (1989) who, have shown that an approximation of the above diagnostic is attainable based on Cook and Weisberg (1982) approach as a likelihood displacement denoted as

$$\text{LD}_i = \mathbf{r}'_{s_i} \text{Var}(\hat{\boldsymbol{\beta}}) \mathbf{r}_{s_i}, \tag{2.34}$$

where $\mathbf{r}_{s_i}$ is the $p \times 1$ vector of score residuals and $\text{Var}(\hat{\boldsymbol{\beta}})$ is the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ the vector of parameter estimates. This statistic is computed in a straight forward manner, using delta betas for each predictor variable in the model. In most cases, index plots or a plot of likelihood displacement against the rank order of survival times provides an informative visual summary of values of this diagnostic. Observations that have relatively large values of the diagnostics are deemed to be influential.

Secondly, another commonly used case-deletion diagnostic for the Cox model is the DFbetas given by

$$\text{DFBetas}_{[-i]} = \frac{\hat{\boldsymbol{\beta}}_{[-i]} - \hat{\boldsymbol{\beta}}}{\text{S.E.}(\hat{\boldsymbol{\beta}}_{[-i]})}. \tag{2.35}$$

This is the difference between the estimated cox regression coefficient $\hat{\boldsymbol{\beta}}$ using all subjects and $\hat{\boldsymbol{\beta}}_{[-i]}$ that without the i^{th} subject standardised by the standard error of $\hat{\boldsymbol{\beta}}_{[-i]}$. This is useful for assessing the influence of subjects on the estimated regression coefficients (Therneau 1997).

2.7 Description of the IMPI trial dataset

In this Section we introduce the IMPI dataset that is used to illustrate the developed diagnostics for joint models for longitudinal and survival data. We also provide descriptive and exploratory statistics for this dataset. The IMPI dataset has been reported by Mayosi et al. (2014) as Phase III multi center, randomized, double-blinded clinical trial. The trial was a 2×2 factorial design for the evaluation of efficacy and safety of two treatments namely, Prednisolone (P) *Mycobacterium indicus pranii* $(M\mathrm{w})$ for the treatment of presumed and confirmed TB pericarditis. There were four treatment arms with the following allocations: $P^{+}M\mathrm{w}^{+}$, $P^{+}M\mathrm{w}^{-}$, $P^{-}M\mathrm{w}^{+}$ or $P^{-}M\mathrm{w}^{-}$ where the plus sign stands for an active agent and a minus sign stands for a placebo.

High morbidity and mortality have been associated with Tuberculosis pericarditis even if anti-tuberculosis therapy is administered as reported by Mayosi et al. (2014). This is prevalent in sub-Sahran Africa and Asia (20%), where a reduction in inflammatory outcome to TB pericarditis is assumed to improve patient's conditions. Two thirds of the participants had concomitant human immunodeficiency virus (HIV) infection, which doubles morbidity to about 40% (Mayosi et al. 2014). It is being evaluated whether the effects of adjunctive glucocorticoid therapy and *Mycobacterium Indicus pranii* immunotherapy in patients with tuberculosis pericarditis is efficacious. The primary efficacy outcome was a composite of death, cardiac tamponade requiring pericardiocentesis, or constrictive pericarditis.

The trial recruited 1400 subjects from nineteen centers in eight African countries with University of Cape Town's Groote Schuur Teaching Hospital acting as a coordinating centre between January 2009 - February 2014.

In this study, participants meeting eligibility criteria were randomly assigned to receive a course of high-dose prednisolone tapered over the course of six weeks or placebo and they were further randomly assigned to receive five injections of heat-killed *Mycobacterium indicus pranii*. At the time of the study, it was a rapidly growing nontuberculous mycobacterium species generally considered to have better therapeutic effects. The primary end point for this study was a composite outcome of death, cardiac tamponade and constrictive pericarditis, whichever occurred first. Study participants were followed up at weeks two and four; thereafter at months three and six during the intervention period with a subsequent six-monthly follow up for four years (Mayosi et al. 2014).

The randomisation of trail participants was such that a total of 706 participants were randomised to the prednisolone arm, which was compared to 694 randomised to the placebo arm. In the *Mycobacterium indicus pranii* arm there were 625 participants assigned to the active treatment and another 625 assigned to the placebo arm. The comparison between *Mycobacterium indicus pranii* and placebo was stopped early for futility in February 2013 (Mayosi et al. 2014). The study power for non-adherence was 10% in the intervention groups, which was a near match with 11% non-adherence in the prednisolone arm. There was a higher non-adherence rate in the *mycobacterium indicus pranii* arm (21%), which was possibly attributed to injection-site adverse effects (Mayosi et al. 2014). The trial experienced about 18% deaths associated with TB pericarditis, HIV related death were 1.3%, other TB not related pericardial effusion (3.3%), cardiovascular (1%) and unknown causes of death (3.5%).

The primary aim of the IMPI trial was to assess safety and effectiveness of prednisolone and *Mycobacterium indicus pranii* in reducing time to the first occurrence of any of the composite outcome which were death, pericardial constriction and cardiac tamponade. The study showed

that prednisolone for six weeks and Mw for 12 weeks had no effect on the composite outcome from all causes. Both treatments were associated with increased risk of HIV related malignancy. The use of prednisolone, on the other hand, showed a reduction in the incidence of constrictive pericarditis and hospitalisation. The study also showed that the beneficial effects of prednisolone on constriction and hospitalisation were similar in HIV positive and HIV negative participants.

The focus of this thesis broadly was to assess the effect of two time varying covariates; namely CD4 cell count and ART uptake, on survival outcomes, which are death and pericardial constriction. We used the joint modelling approach to specifically develop diagnostics for assessing the effect of time varying covariates on survival outcomes. This was so because joint modelling provides better insights into mechanisms that underlie both survival and longitudinal outcomes (Rizopoulos, Verbeke, and Molenberghs 2010). There is a high co-morbidity between HIV and TB in the IMPI trial $939(67\%)$ hence interest in investigation the effect of prednisolone among HIV positive participants who had CD4 cell count measurements and ART uptake recorded repeatedly over the study duration.

We restricted this analysis to HIV positive subjects only who were in majority and presumed to be sicker and we further focused on prednisolone which showed beneficial effects (Figure 2.1). The analysis was further restricted to a six months' observation time because this is the duration where CD4 cell count measurements were mandatory for all participation sites.

We further focused on those subjects (n = 587) that had at least two CD4 cell count measurements for the joint modeling analysis. We defined an end point as the first occurrence of either death or constrictive pericarditis in the first six months of study period. The trial collected CD4 cell count measurements at baseline, two weeks, one and a half months, three months

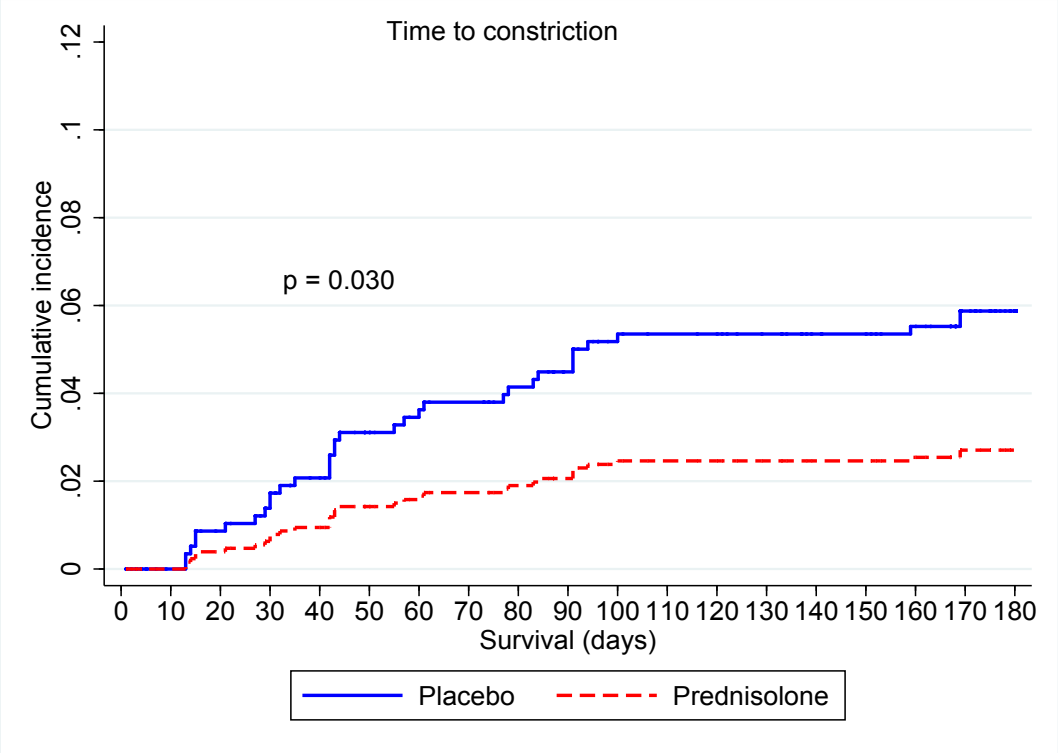


Figure 2.1: Kaplan Meier for constriction.

and six months together with information about ART up take during the scheduled time points. In most participating sites, except South Africa, CD4 cell count measurements were collected for only six months, therefore, the analysis excluded data from months 24, 36 and 48, which would likely bias results and inferences from this statistical analysis.

2.7.1 Descriptive statistics of the IMPI data set

We will consider square root CD4 cell count (Gaussian) and ART (binary) uptake as the main longitudinal outcomes in this analysis. At baseline, there were 644 (68.5%) measurements taken and 587 (42%) had, at least, two CD4 cell count measurements. The number of CD4 cell count measurements decreased steadily with time from baseline until six months. The following figure shows the spaghetti plot (Figure 2.2) and the longitudinal profile of mean square root CD4 cell count (Figure 2.2). The mean profile for the square root CD4 cell count seems

to be linear, hence consideration of the linear mixed model.

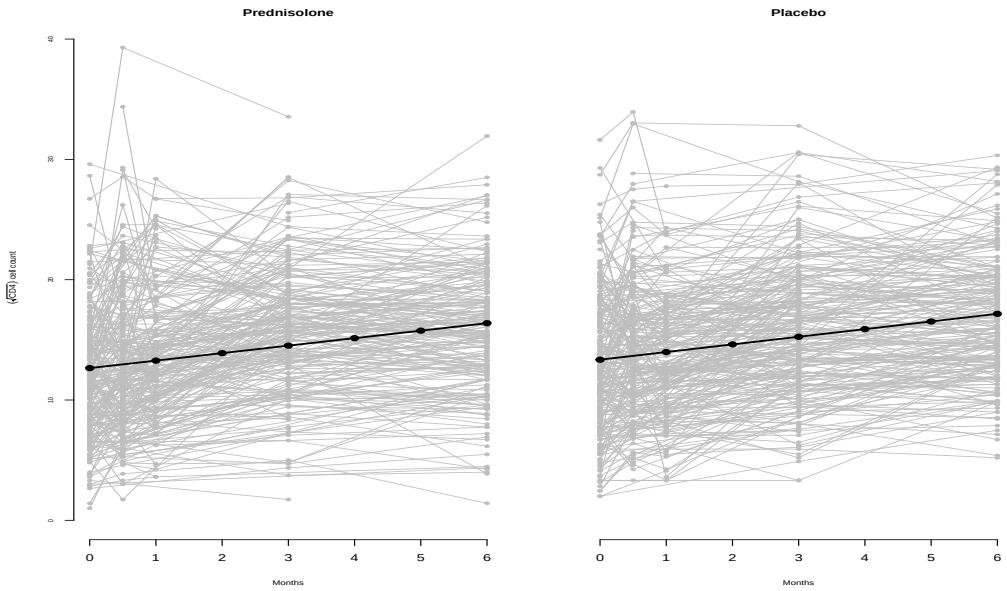


Figure 2.2: Plot of $\sqrt{CD4}$ cell count over six months.

The distribution of CD4 cell count is known to be skewed and this was consistent in the IMPI dataset, as such the square root transformation has been adopted for modeling the longitudinal CD4 cell count. From Figure 2.2, the predicted square root CD4 cell count from the model is fairly linear which prompted the implementation of the linear mixed model.

2.7.2 A linear mixed model for longitudinal CD4 cell counts

We propose a linear mixed model which assumes a linear trajectory over time for the square root CD4 for each subject, which means that it includes a single random effects term for each subject. This, therefore, posits that the average longitudinal evolutions per subject are same

but different at baseline for each subject.

$$\begin{cases} Y_{ij} = \beta_0 + \text{Pred}_i\beta_2 + t_{ij}\beta_1 + b_i + \epsilon_i \\ \\ \quad b_i \sim \mathcal{N}(0, \sigma_b^2), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \\ \\ \quad b_i \perp\!\!\!\perp \epsilon_i \end{cases} \tag{2.36}$$

where Y_{ij} is the CD4 cell count at the j^{th} visit time for the i^{th} subject, β_0 is the intercept, t_{ij} is the CD4 cell count measurement time, Pred_i is treatment effect, β_i is the fixed effect and b_i is the random effect. Results from model (2.36)are summarized in Table (2.1) below. The intercept is 13.14 which is the average square root CD4 cell count at baseline and the constant rate of increase in the square root CD4 cell count in the Prednisolone arm is 0.25 per calendar time. There is no difference between the two treatment arms with respect to increase in CD4 cell count (p-value> 0.40).

Table 2.1: Linear mixed effects models results.

	Random intercept only		Random intercept and slope	
Parameter	Estimate (Std. Error)	p-value	Estimate (Std. Error)	p-value
Intercept	13.14 (0.28)	<0.001	13.14 (0.26)	<0.001
Prednisolone	-0.25 (0.39)	0.52	-0.26 (0.37)	0.49
Time	0.16 (0.01)	<0.001	0.17 (0.01)	<0.001
Prednisolone × Time	-0.01 (0.02)	0.59	-0.01 (0.02)	0.61
σ_b^2	5.03	-	4.71	-
σ_e^2	2.76	-	2.85	-
-2 Loglikelihood	13113.82	-	13145.96	-
AIC	13129.82	-	13159.96	-
BIC	13175.68	-	13200.08	-

2.7.3 Case-deletion diagnostics for the linear mixed model for CD4 cell counts

The diagnostics considered for this model. which is the sub-model in the joint modeling framework, is based on Cook's distance. The approach hinges on re-estimating model parameters after iteratively excluding each subject or an observation. The estimation mechanism could be based on maximum likelihood, restricted maximum likelihood or any novel estimation procedure.

2.7.4 Identifying influential subjects in the IMPI dataset

We implement the Cook's distance statistic to identify subjects that are potentially influential using IMPI data as follows from model 2.21 above:

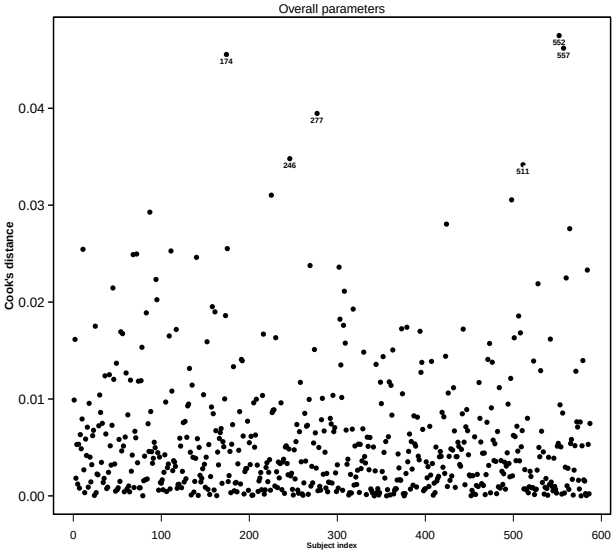


Figure 2.3: Cook's statistics for all parameters.

The following subjects: 174, 246, 277, 552 and 557 were identified as potential influential subjects on the estimation of all parameters Figure 2.3. On the other hand, a similar set of subjects as follows: 174, 246, 277, 511, 552 and 557,were identified as influential on the

estimation of fixed effects parameters as in Figure 2.4 (a). Subjects identified as potentially influential on the variance components were 87, 174, 277, 557 and 564. The Cook's statistics were implemented separately for the fixed effects and variance component isolated subjects 174, 277 and 557 as being potential subjects that required further investigation in both settings (Figure 2.4(b)).

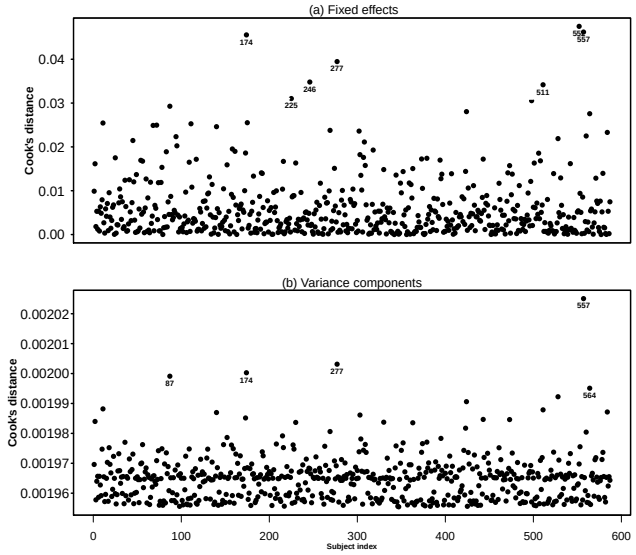


Figure 2.4: Fixed effects and variance components.

2.7.5 Identifying influentials observations

We also implemented the Cook's distance statistic at observation level using IMPI data to the following model for all parameters:

$$CD_{ij}(\theta) = (\hat{\theta}_{[-ij]} - \hat{\theta})' \mathbf{M}^{-1} (\hat{\theta}_{[-ij]} - \hat{\theta}). \tag{2.37}$$

The evaluation of the Cook's statistic was done after iteratively excluding the j^{th} observation for the i^{th} subject from the model and re-estimating the parameters. The observations in this case were square root CD4 cell count measured at the j^{th} time point.

The same approach is in model (2.37)was implemented for fixed effects at observation level.

Observations are denoted with an underscore indicating the time point at which the CD4 cell count was measured. As an example, observation 2_{-6} means that it is the square root CD4 cell count for subject number measured at 6 months' visit.

Observations that were identified as potential outliers or exerting influence on the estimation process are $235, 574, 867, 868$ and 1950 which are linked to the following subjects and their time for CD4 cell count measurement $68_{-6}, 158_{-6}, 246_{-0.5}, 246_{-6}$ and 552_{-6}. The evaluated Cook's statistic was, overall, on the entire parameter space (Figure 2.5). The same observations were identified as influential when the Cook's statistic was evaluated on the fixed effects Figure2.6 (a)

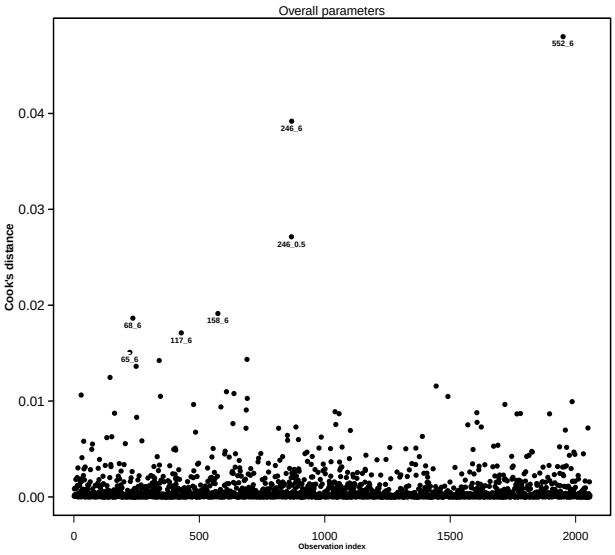


Figure 2.5: Cook's statistic for all parameters.

Those that were identified as having potential influence on the variance components are: $65_{-6}, 68_{-6}, 117_6, 158_6, 246_{-0.5}, 246_{-6}$ and 552_{-6}. In this case observations that are identified as influential on both fixed effects and variance components are $68_{-6}, 158_6, 246_{-0.5}$, 246_{-6} and 552_{-6}. These observations, therefore, need further investigation Figure 2.6(b).

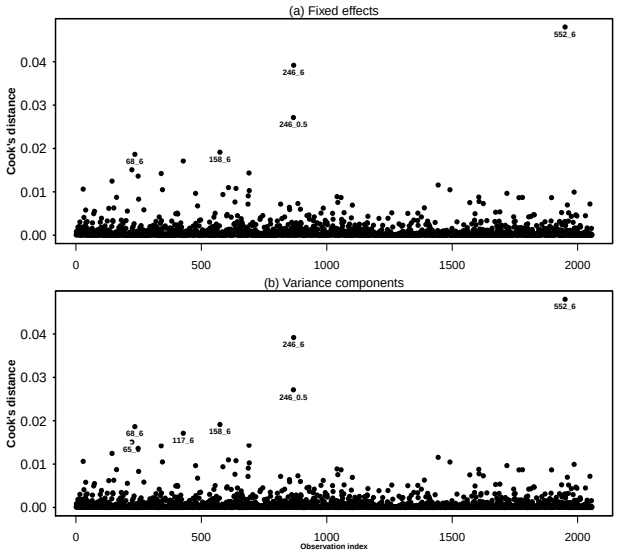


Figure 2.6: Fixed effects and variance components

The case-deletion diagnostics using analogues of models 2.21 and (2.37) for subjects and observations, respectively, in the two-stage joint model, the following subjects were identified as influential $316, 320, 468, 474, 506$ and 571 Figure 2.7 (a). Adopting the notation that 1_{-0} refers to an observation that belongs to subject identifier number one with a CD4 cell count measurement taken at visit zero the following observations were identified as influential $23_{-0}, 205_{-6}, 473_{-0}, 473_{-0}, 506_{-0}$ and 571_{-0} as in the Figure 2.7 (b).

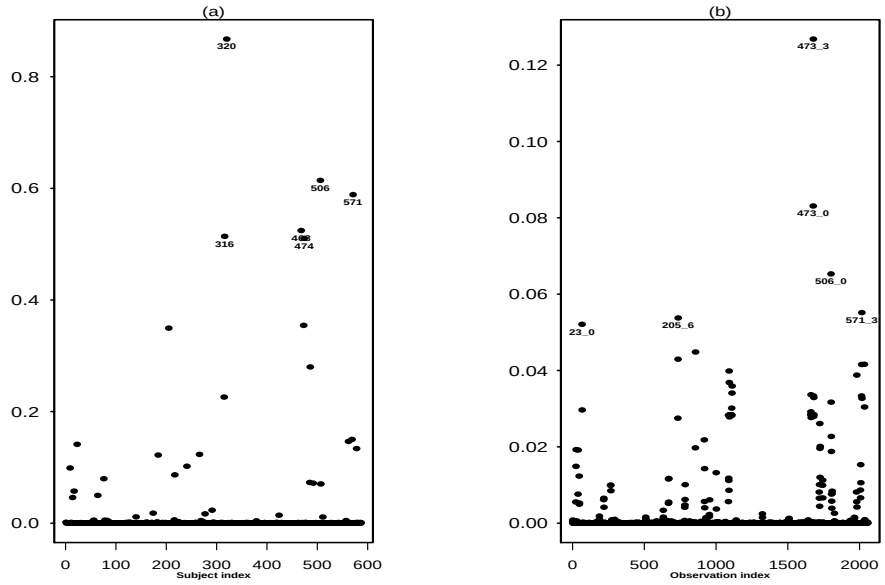


Figure 2.7: Cook's distance for the Two-stage joint model.

2.8 A generalised linear mixed model for ART uptake

There were a few missing data reported on ART uptake over the period under consideration. At baseline there was 2% missing data on ART uptake, 1% and 4% at months three and six, respectively. We modeled ART uptake using a generalised linear mixed modeling framework for ART uptake, which is then further extended into the joint model for a binary longitudinal outcome. We used the logit link function to model the binary outcome (Yes/ No), which was recorded at baseline, week two, months' one, three and six.

$$\Pr(Y_{ij} = 1|X_{ij}, \mu_i) = F(X_{ij} + \mu_i), \tag{2.38}$$

where X_{ij} are covariates which vary between and within subjects and μ_i is the random effect which represents unexplained specific covariates, which in essence adds a random intercept.

The random intercept model is as follows:

$$\text{logit}\Pr(Y_{ij}=1|X_{ij},\mu_i)=\alpha+X_{ij}b+\mu_i \tag{2.39}$$

There was an improved ART up take from baseline (19%) to about 70% at six month visit (Figure 2.8).

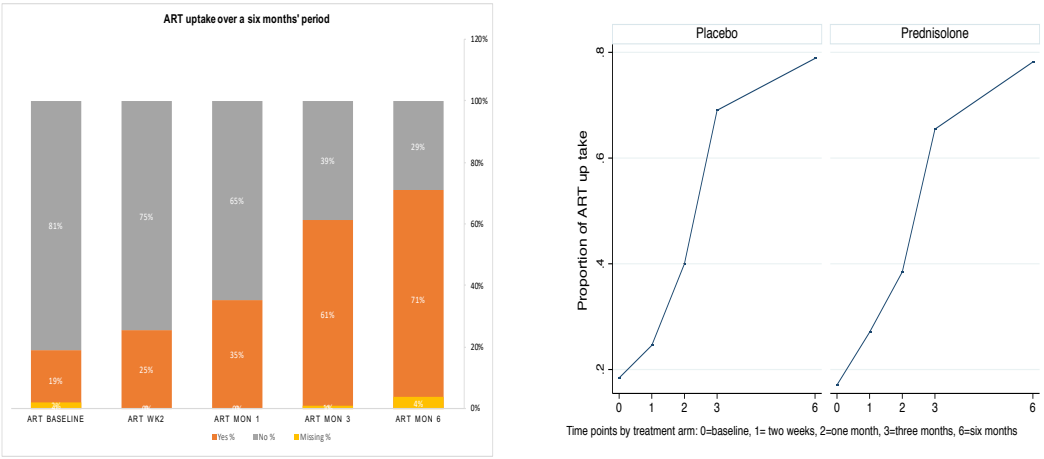


Figure 2.8: ART uptake profiles during the six months' period.

We fitted model (2.39) to the IMPI dataset and tabulated the results as follows

Table 2.2: Logistic mixed effects model results

Parameter	Estimate (SE)	p-value
Intercept	-6.36 (1.08)	<0.001
Treatment	-0.16 (0.76)	0.84
Time	2.19 (0.31)	<0.001
Treatment $\times$ Time	0.01 (0.17)	0.99
σ_b^2	47.87	-
σ_e^2	31.83	-
-2 Loglikelihood	1862.80	-

To account for the subject specific nature of the trial we fit a mixed effects logistic regression model with a random effect per subject (i.e., a random intercepts model). From this model, we got the regression coefficient for treatment (Prednisolone) as -0.16. This -0.16 is the log

odds ratio between the placebo and active treatment for each subject. There is, on the hand, a statistically significant effect of time; log odds 2.19 $p-$Value < 0.001 but no significant difference between placebo and active treatment over time (interaction effect) $p-$Value $= 0.99$ (table 2.2).

2.9 Summary and discussion

In medical research, often times, repeated measurements are conducted on subjects or indeed repeated bio-marker samples drawn from subjects for analysis. At the same time interest may be on observing an event of interest during a study period. There is a rich methodology for analysing data that is correlated through repeated measurements as well as analysing time to event data independently. There are cases when interest is in assessing the strength of association between longitudinal bio-markers and risk of an event. An equally important research question could be to conduct an analysis on a survival outcome influenced by time-dependent covariates measured with error. The other aspect to be considered for analysis could be accounting for non-random drop out caused by the occurrence of an event in the longitudinal outcome. Standard methods for conducting analysis to address the above are not attainable as observed by Rizopoulos, Verbeke, and Molenberghs (2010) and Tsiatis, Degruttola, and Wulfsohn (1995). In the last two decades interest had increased to jointly estimate the two processes through semi-parametric maximum likelihood (Henderson, Diggle, and Dobson 2000; Hsieh, Tseng, and Wang 2006; Wulfsohn and Tsiatis 1997), whilst the Bayesian approach, using MCMC methods has also been done (Chi and Ibrahim 2006; R Brown and G Ibrahim 2003; Wang and Taylor 2001; Xu and Zeger 2001) among others. This joint estimation is commonly

referred to as joint models for longitudinal and survival data. Joint models for longitudinal and survival data are a class of models that jointly analyse an outcome repeatedly observed over time such as a bio-marker and associated event times. These models are useful in two practical applications; firstly, focusing on survival outcome whilst accounting for time varying covariates measured with error and secondly, focusing on the longitudinal outcome while controlling for informative censoring.

There are different classes of joint models namely; latent class joint models, which are primarily useful for predicting the occurrence of an event (Proust-Lima et al. 2009, 2014) and shared parameters joint models which assess the strength of association between a bio-marker and risk of an event (Henderson, Diggle, and Dobson 2000; Rizopoulos 2012; Tsiatis and Davidian 2004) and this is the main focus of this thesis.

Interest on the estimation of these joint models has grown in the past two and half decades. However, minimal effort has been directed towards developing diagnostic assessment tools for these models. The available diagnostic tools have mainly been based on separate analysis of residuals for the longitudinal and survival sub-models which could be sub-optimal. An assessment of joint models is done using residuals from the separate sub-models and joint models diagnostics have not been fully developed. The technique for joint model diagnostics has been using qualitative graphical displays to isolate departures from model assumptions and influential observations by using summary measures (Dobson and Henderson 2003). This has been in part due to lack of computing power to do a full case deletion and re-estimation for each subject which was not practical as observed by Dobson and Henderson (2003).

As can be evidently seen in Section 2.7.2 of this Chapter, in Figures 2.3 and 2.5, with their respective parameter based Cook's statistics, there exist influential subjects or observations in

the longitudinal as well as survival sub model. These influential subjects or outliers have an effect of either over or under estimating treatment effects thereby leading to unrealistic statistical inferences if not properly investigated.

In this thesis we develop Cook's statistics for detecting influential observations or subjects for joint model estimates in univariate (Singini, Mwambi, and Gumedze 2018) and multivariate setting as well as implementing a variance shift outlier models in the longitudinal part of the two-stage joint model. We conducted simulation studies to evaluate these diagnostic statistics. These diagnostic techniques are further illustrated using data from a multi-center clinical trial on TB pericarditis.

Chapter 3

Review of the joint models for longitudinal and survival data

The purpose of this Chapter is two-fold; we will, firstly, review an introduction to the joint model for longitudinal and survival data. This will lay the foundation for understanding the building blocks for these joint models which is crucial in estimation and inference as well as model checking, validity and influence diagnostics. This foundation will be used in the development of diagnostic tools. Secondly, we will review existing model checking and diagnostic tools for the joint model for longitudinal and survival data.

Estimation techniques for joint models have grown in the last two decades as reported by Rizopoulos, Verbeke, and Molenberghs (2010). These methods are largely semi-parametric maximum likelihood based on work done by Henderson, Diggle, and Dobson (2000), Hsieh, Tseng, and Wang (2006), Tsiatis and Davidian (2004), and Wulfsohn and Tsiatis (1997) and Bayes using MCMC methods by Chi and Ibrahim (2006), R Brown and G Ibrahim (2003), Wang and Taylor (2001), and Xu and Zeger (2001).

There are different classes of joint models as previously stated; latent class joint models as described in Section 2.9, and among others, shared parameter joint models, which is the main focus of this thesis.

3.1 Models for time varying covariates

The extended Cox survival model handles time-dependent covariate, instead of using baseline measurements of a biomarker for prognosis in clinical research (Hippisley-Cox et al. 2007). This type of approach ignores longitudinal measurements for prognosis, which may not be optimal because useful longitudinal information is practically thrown way and also assumes that there is no measurement error. The extended Cox model 3.4 below (Gill, Keiding, and Andersen 1993), on the other hand, incorporates the observed longitudinal measurements but also assumes that the covariate does not change between measurements, thereby, ignoring measurement error, which may not be true.

3.1.1 The extended Cox model

In the analysis of time to an event; if the survival function is assumed to be of parametric form, estimation of parameters of interest is based on maximum likelihood methods. Maximum likelihood methods are used to obtain estimates of coefficients (βs) in the Cox proportional hazards model. Like all likelihood estimations, the first step is to get the likelihood from the sample data. The Cox model is unique in that it is a function of the survival times that have been observed and unknown parameters (β). Maximum likelihood estimates are thus those that are more likely given the observed data, which in essence maximises the likelihood function. The standard practice has been to maximise the logarithm of the likelihood function due to computational challenges. It has been shown by Cox (1972b) that the likelihood function

of the proportional hazards model is as follows:

$$\ell(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\mathbf{X}_j \boldsymbol{\beta}')}{\sum_{l \in R(t_j)} \exp(\mathbf{X}_l \boldsymbol{\beta}')}, \tag{3.1}$$

where $\mathbf{X}_j$ is a vector of covariates for events at the j^{th} event time, t_j and the summation in the denominator is the sum of values $\exp^{(\mathbf{X}_l \boldsymbol{\beta}')}$ over all individuals who are at risk at time t_j.

Since censoring has to be accounted for in survival analysis, the construction of the likelihood function proceeds by combining information from the censored as well as un-censored subjects, which results in the log-likelihood of the following form:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \log p(T_i; \boldsymbol{\theta}) + (1 - \delta_i) \log S_i(T_i; \boldsymbol{\theta}), \tag{3.2}$$

where δ_i is an indicator of whether the i^{th} subject has an event or not, and hence takes value one for an event and zero otherwise. Using the relationship $S(t) = \exp\{-H(t)\}$ the log-likelihood can be re-written as follows:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \log h_i(T_i; \boldsymbol{\theta}) - \int_0^{T_i} h_i(S_i; \boldsymbol{\theta}) ds, \tag{3.3}$$

This formulation has the characteristic that all subjects contribute an amount to the log-likelihood equal to the negative of the cumulative hazard function evaluated at their observed event time T_i.

Once the log-likelihood has been formulated, iterative optimisation procedures such as Newton-Raphson algorithm Lange and Hunter (2004) can be used to locate the maximum likelihood estimate for parameter $\boldsymbol{\theta}$, say. Inference proceeds under classical asymptotic maximum likelihood theory Cox and Hinkley (1974).

An extension to model (2.5)can be formulated to handle "endogenous (internal) time varying covariates using the counting process formulation" (Rizopoulos 2012), which is a relative risk model with longitudinal history (Andersen and D 1982; Fleming and Harrington 1991; Gill,

Keiding, and Andersen 1993). This is also called the extended Cox model.

$$h_i(t|Y_i(t), x_i) = h_0(t)R_i(t)\exp\{\mathbf{X}_i\boldsymbol{\beta}' + \alpha Y_i(t)\}, \tag{3.4}$$

where $\mathbf{X}_i$ denotes a vector of baseline covariates and $Y_i(t)$ denotes the time varying covariates and $R_i(t)$ is the risk set. The idea behind this model is to imagine that the occurrence of events as a slow Poisson process (Rizopoulos 2012). The notation used in the counting process is $\{N_i(t), R_i(t)\}$, which is the event process for the i^{th} subject with $N_i(t)$ denoting the number of events for subject i by time t and $R_i(t)$ is an indicator of the at risk process. $R_i(t) = 1$ if subject i is at risk at time t and $R_i(t) = 0$ otherwise.

Parameters $\boldsymbol{\beta}$ and α are estimated based on the partial log-likelihood function written as (Rizopoulos 2012):

$$\begin{aligned} p\ell(\boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n} \int_0^{\infty} \Bigg\{ & R_i(t)\exp\{\mathbf{X_i}\beta' + \alpha Y_i(t)\} \\ & - \log\left[\sum_j R_j(t)\mathbf{X_i}\beta' + \alpha Y_j(t)\}\right] \Bigg\} dN_i(t), \end{aligned} \tag{3.5}$$

where R_i is the risk set, N_i is the counting process, $\boldsymbol{\beta}$ is the coefficient of baseline covariates and α is the time varying risk coefficient. The interpretation for regression coefficients are the same as those for standard survival setting. Special mention has to be made, once again, of α which is interpreted as the relative increase in the risk of an event (death or constriction) at time t that results from one unit decrease in the time varying covariate (square root CD4 cell count) at the same time point.

To get around this problem, a two-stage joint survival model is proposed, which first fits the linear mixed model for subject specific predictions and these predicted values are then used in the survival model as time-varying covariates (Sweeting and Thompson 2011). The steps for implementing this model are presented in Chapter 4 and the model formulation is 4.3.

The problem with this approach is that, since the first step is stochastic in nature, the uncertainty in the estimates has to be accounted for in the second stage, but this approach does not account for this uncertainty. This problem has been dealt within the Bayesian framework using importance sampling as reported by Mauff et al. (2018, 2020). The shared parameter joint model utilises all repeated measures, accounts for measurement error, reduces bias and maximizes efficiency. Shared parameter Joint models are useful in three ways namely:

- they adjust for non-random drop out in the analysis of the longitudinal outcome, which is not the focus of this thesis (missing data).
- they provide an effective way of conducting analysis of survival outcome influenced by time varying covariate measured with error.
- they assess the strength of association between a biomarker and risk of an event in longitudinal studies.

The shared parameter Joint model shares common the random effects between the longitudinal and survival processes as follows (Tsiatis and Davidian 2004):

$$p(\mathbf{Y}_i, \mathbf{T}_i, \boldsymbol{\delta}_i) = \int p(\mathbf{Y}_i|\mathbf{b}_i)\{H(\mathbf{T}_i|\mathbf{b}_i)^{\boldsymbol{\delta}_i} S(\mathbf{T}_i|\mathbf{b}_i)\}p(\mathbf{b}_i)d\mathbf{b}_i, \tag{3.6}$$

where $\mathbf{b}_i$ is a vector of random effects that explains the interdependencies between the longitudinal density $p(.)$ and the survival density $S(.)$. This model has been presented in Chapter 5 from estimation to inference.

3.2 Bayesian Case Influence Diagnostics for Survival Models

The work by Cho et al. (2009) proposed and developed a Bayesian case influence diagnostics for survival models which is a sub-model for joint models for longitudinal and survival data. The developed influence diagnostic tool is based on case deletion for both the joint and marginal posterior distributions using Kullback–Leibler divergence (K–L divergence). They further decomposed the expression for computing the K–L divergence between the posterior with the full data and the posterior based on single case deletion. They also investigated its relationships to the conditional predictive ordinate. All the computations for the proposed diagnostic measures were done using Markov chain Monte Carlo samples from the full data posterior distribution. In their approach Cho et al. (2009) considered the Cox model with a gamma process prior on the cumulative baseline hazard. Their theoretical relationship between case-deletion diagnostics and diagnostics based on Cox's partial likelihood are presented below supported by simulated data and two real data examples were given (Cho et al. 2009).

3.2.1 Development of the Bayesian diagnostic tool

In their development of the influence diagnostic tool Cho et al. (2009) begin by letting, D be full data and D_{-i} to be the data with the i^{th} subject deleted. They also let $\ell(\beta|D)$ denote the likelihood based on the full data and $\ell(\beta|D_{-i})$ denote the likelihood based on the data without the i^{th} subject. The posterior distributions for the full data and the i^{th} subject deleted is be defined as $p(\beta|D) \propto \ell(\beta|D)\pi(\beta)andp(\beta|D_{-i}) \propto \ell(\beta|D_{-i})\pi(\beta)$, respectively, where $\pi(\beta)$ is the prior distribution of β. A typical choice of $\pi(\beta)$ is a $\mathcal{N}p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ distribution or a uniform improper prior. Further let $K(P, P_{-i})$ denote the K–L divergence between P and

P_{-i}, where P denotes the posterior distribution of β for the full data, and P_{-i} denotes the posterior distribution of β without the i^{th} subject as follows:

$$K(P, P_{-i}) = \int p(\beta|D) \log \left\{ \frac{p(\beta|D)}{\beta|D_{-i}} \right\} d\beta \tag{3.7}$$

$K(P, P_{-i})$ measures the effect of deleting the i^{th} subject from the full data on the joint posterior distribution of $\boldsymbol{\beta}$ and in general $K(P, P_{-i}) \neq K(P_{-i}, P)$. Extensions and expression of this statistic has been explained and eraborated with two examples (Cho et al. 2009).

3.3 Review of joint model assessment tools

In this Section we review existing model assessment tools for the joint model. The main focus of the current joint model checking techniques have targeted separate analysis of residuals from the sub-models (Dobson and Henderson 2003; Rizopoulos, Verbeke, and Molenberghs 2010). There have been investigations done in both the frequentist as well as the Bayesian stream to assess different aspects of the joint model.

3.4 Influence diagnostics for joint model

The model formulation is such that; a relative risk model based on a conditional Gaussian process is assumed and for discrete dropout time, using an appropriate probability density. This model is popularly referred to as the extended Cox model. Responses and dropout are considered to be conditionally independent meaning that they are taken as a linear combination of Gaussian random effects. The main technique for model diagnostics has been using graphical displays to isolate departures from model assumptions and influential observations by using

summary measures (Dobson and Henderson 2003). A preliminary analysis to fitting the joint model is to check whether there is anything in the response, $\mathbf{Y}$, that would allow subjects to drop out for informative censoring, otherwise, the joint analysis is not worthwhile.

Letting $\mathbf{R}_i$ be the vector of residuals for subject i having removed the fixed covariate effect, at measurement time d, R_d is the risk set consisting all subjects who have, at least, $d-1$ responses, R_{id} be the subset of $\mathbf{R}_i$ corresponding to response before d. Using the two-dimensional summaries of R_{id} given as $p_{id} = (p_{id1}, p_{id2})'$ which enables a qualitative inspection of scatter plots of p_{id} given a group code at each measurement time. Letting $\bar{p}_{gd}$ be the sample mean of p_{id} for subjects in group g, the Mahalanobis distance can be calculated (McLachlan 1999). is presented as:

$$M_{jk} = (\bar{p}_{jd} - \bar{p}_{kd})' \hat{Var}(p_{id})^{-1} (\bar{p}_{jd} - \bar{p}_{kd}). \tag{3.8}$$

Permutation based tests on shuffles of group membership indicator g can be used for quantitative assessment to visually inspect the joint model for influential subjects (Dobson and Henderson 2003). The influence diagnostics are evaluated using a one-step approximation derived from the first terms of Taylor series of the log-likelihood at $\hat{\theta}_{(i)}$ expanded about $\hat{\theta}$.

$$\hat{\theta}_{(i)} - \hat{\theta} \simeq \{-I^{-1}_{(i)}(\hat{\theta}) u_{(i)}(\hat{\theta})\}, \tag{3.9}$$

where $u_{(i)}(\hat{\theta})$ and $I^{-1}_{(i)}(\hat{\theta})$ are the observed score and information after removing subject i.

This was the case because a fitted joint model was computer intensive as such full case deletion and re-estimation for each subject is not practical. In conclusion Dobson and Henderson (2003) managed to develop exploratory techniques to assess model fit and influence in the joint model setting. To assess model fit on the longitudinal sub-mode, they used the Akaike and Bayes information criterion (AIC & BIC), which fitted the data well even after a penalty

was used for complexity (Dobson and Henderson 2003).

3.5 Bayesian influence diagnostics for joint model

Influence diagnostics on the joint models have been studied by Zhu et al. (2012) using a Bayesian approach motivated by the absence of such research done on these joint models. In this Bayesian joint modeling framework, they used the shared parameter model with time dependent random effects underlying both the survival and longitudinal sub-model. The model formulation was done using the conditional independence assumption on the random effects on both the survival and longitudinal outcomes.

The use of Bayesian influence analysis in the joint modeling framework is addressed in three ways through model perturbation schemes. Individual perturbation of a subjects' longitudinal profile are done to reveal profiles that are significantly different from the rest of the subjects contributing to the evolution process. In their approach Zhu et al. (2012), also did perturbation to survival data designed to expose survival times that relatively have influential values. They went further to perturb the common random effects for both parts of the joint model. This allowed to detect survival times with influence which had low occurrence with respect to their subject specific longitudinal profiles.

The generic perturbation scheme that Zhu et al. (2012) implemented was as follows: Letting ω be a vector in a set $\Omega \subset R^W$, which represents a Euclidean space of dimension W where W is an integer. The perturbed model $M = \{p(D_0, \mathbf{b}; \boldsymbol{\theta} \quad \boldsymbol{\omega}) : \omega \in \Omega\}$ has the effect of indicating various perturbations to the specified density $p(D_0, \mathbf{b}; \boldsymbol{\theta})$ such that $\int p(D_0, \mathbf{b}; \boldsymbol{\theta}, \boldsymbol{\omega}) D_0 D\mathbf{b} = 1$ and $p(D_0, \mathbf{b}; \boldsymbol{\theta}, \boldsymbol{\omega^0}) = p(D_0, \mathbf{b}; \boldsymbol{\theta})$ for unique ω^0, which means that there is no perturbation.

This process starts with perturbing subject's individual longitudinal profile, repeated measures within each subject, covariates, survival time and finally, the censoring indicator.

Finally, Zhu et al. (2012) perturbed the prior and all components of the Bayesian joint model at the same time. They then combined the three perturbation schemes and obtained prior sampling distributions given the data. This enables to assess sensitivity in all components of the joint model while local influence measures are done using the first order and second order influence, which quantifies the magnitude of influential direction in the joint model. These proposed methods are able to detect outliers, ascertain unverifiable assumptions underlying the Bayesian joint modeling framework.

3.6 Multiple imputation-based diagnostics for joint model

There is extensive work done on the residual based diagnostics of survival and longitudinal outcomes modeled separately. The problem arises when these residuals are calculated based on the joint fitted model and the observed data. This is mainly due to the non-random dropout in the longitudinal process caused by the occurrence of events. A detailed account has been given by Rizopoulos, Verbeke, and Molenberghs (2010), which presents both response vectors for the observed and missing longitudinal part. In their attempt to address model assessment, Rizopoulos, Verbeke, and Molenberghs (2010) propose a three step simulation scheme as a result of the fact that there exists non-random dropout in the longitudinal marker. As already discussed, this arises because the missing longitudinal measurement is as a result of the occurrence of an event, it is, therefore, misleading to calculate and analyse residuals whose reference distribution does not exhibit standard properties, let alone being directly available. It is,

thus, proposed to impute the missing longitudinal data using a three step multiple imputation scheme. Formally the three step simulation scheme is presented as:

$$\begin{aligned} s1: Draw \quad & \theta^{(\ell)} \sim N\{\hat{\theta}, v\hat{a}r(\hat{\theta})\} \\ s2: Draw \quad & b_i^{(\ell)} \sim N\{b_i|y_i^o, T_i, \delta_i, \theta^{(\ell)}\} \\ s3: Draw \quad & y_i^{m(\ell)}(t_{ij}) \sim N\{\hat{m}_i^{(\ell)}(t_{ij}), \hat{\sigma}^{2(\ell)}\}. \end{aligned} \tag{3.10}$$

For respecified visit times $t_{ij} > T_i, j = 1, ..., n_i'$ that were not observed for the i^{th} subject. where $\hat{m}_i(t_{ij}) = x_i^T(t_{ij})\hat{\beta}^{(\ell)} + z_i^T(t_{ij})\hat{b_i}^{(\ell)}$

This scheme has a Bayesian back ground, where the multiple imputation comes from the repeated sampling from the posterior distribution of the missing longitudinal response given the observed data, which is then averaged over a posterior distribution of parameters. The idea behind the multiple imputation is inspired by augmenting the observed data with randomly imputed data points that would otherwise have been provided, had the subject not dropped out as a result of the occurrence of an event. This creates a complete data from which residuals can be calculated and plots produced.

Separate analysis of residuals is done for the joint model, i.e. the longitudinal part is assessed for marginal as well as subject specific residuals versus fitted values. The same applies to the survival part of the joint model, i.e. marginal survival plus cumulative hazard plots are assessed including the survival functions of Cox-Snell's plus Martingale residuals. In this calculation of residuals, it is assumed that the variance of all residuals is the same as the variance corresponding to the error terms of the sub-model, which is not a correct assumption. As noted by Rizopoulos, Verbeke, and Molenberghs 2010 the merit of this simulation approach is that the steps taken are simple and easy to perform, especially with the current computing power.

3.7 Goodness of fit joint model

An important characteristic of assessing goodness-of-fit for joint models for longitudinal and survival data is separate contribution from the model components. A novel decomposition of the Akaike information criterion (AIC) and Bayesian information criteria (BIC) into additive components that permits an assessment of goodness-of-fit is derived (Zhang et al. 2014). This is done by developing a quantifiable change in AIC and BIC for fitting the survival data with and without the longitudinal data.

We commence presenting the Akaike information criterion (AIC) which is an estimator of the relative quality of statistical models for a given data set Akaike (1974). If there are more than one model for fitting the data, AIC estimates the quality of each model, relative to each of the other. In this way AIC provides a means for model selection, and it is presented as follows. Let k be the number of estimated parameters in the model. Let $\hat{L}$ be the maximum value of the likelihood function for the model.

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})\mathrm{AIC} = 2k - 2\ln(\hat{L}) \tag{3.11}$$

In a situation where there is set of candidate models for fitting data, the preferred model is the one with the minimum AIC value. The AIC, therefore rewards goodness of fit which is assessed by the likelihood function, but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages over-fitting, which is desired because increasing the number of parameters in the model almost always improves the goodness of the fit. We then, briefly discuss the Bayesian information criterion (BIC) (Schwarz 1978) which, like the AIC, provides a means for assessing the goodness of fit from a set of

candidate models. The BIC is defined as:

$$\mathrm{BIC} = k\ln(n) - 2\ln(\widehat{L}).\ \mathrm{BIC} = k\ln(n) - 2\ln(\widehat{L}), \tag{3.12}$$

where $\hat{L}$ is the maximized value of the likelihood function of the model M; $\widehat{\boldsymbol{\theta}}$ are the parameter values that maximize the likelihood function; xx is the observed data; n is the number of data points in x otherwise the sample size; and k is the number of parameters estimated by the model (Schwarz 1978).

The approach proposed by Zhang et al. (2014), it is possible to compare the longitudinal outcomes (binary) in terms of their contribution to the overall fit of the survival data (Zhang et al. 2014). The assessment of the model is done in two parts, based on the likelihood theory using conditional density of the random effects given the longitudinal part and the same for the survival part. This quantification of change in AIC for fitting survival data with and without longitudinal data helps to compare the longitudinal outcomes for their contribution to the overall fit of the survival data.

Chapter 4

Two-stage joint model for longitudinal and survival data

Introduction

In this Chapter, we introduce the motivation behind joint models for longitudinal and survival data. There are a number of ways to model data that has time dependent covariate and these are; using the extended Cox model or using the two-stage joint model or even further using the shared parameter joint modeling of the two processes given shared common random effects. We start with the two-stage joint model since the extended Cox has been covered in the previous Chapter and move on to the shared parameter joint model. We will summarise the evolution and formulation of these joint models for longitudinal and survival data. We implemented the variance shift outlier model in the two-stage joint model which identifies outliers. The identified outliers are then down-weighted, at the discretion of a researcher, using a shift in variance. There are currently no studies that have implemented a variance shift outlier model in the two-stage joint model. Implementing a variance shift outlier model is important because all observations are used in the model, thereby, maintaining the power of the study.

4.1 The Two-stage joint model for longitudinal and survival data

There are cases where a research question addresses two components, for instance, how does a biomarker evolve over time, and is evolution of the biomaker associated with the risk of

an event? In this setting, there is clearly a longitudinal and survival component that need to be considered when addressing such a research question as previously spelt out. In the past, traditional approaches to this were to fit a survival model with baseline biomarker values as a predictor variable, which is considered sub-optimal as it does not use all the repeated measurements in the model. The alternative to this shortcoming is to fit a time-dependent survival model (extended Cox), which, then, makes use of all the repeated measurements. The time dependent covariate model introduces a stringent assumption; that the biomarker stays constant between successive measurements and that there is no measurement error.

There has been much work done by a number of authors; Dafni and Tsiatis (1998), Self and Pawitan (1992), and Tsiatis, Degruttola, and Wulfsohn (1995) to deal the problems. It had been proposed to have a two staged model, which, first, fits the linear mixed model and derive maximum likelihood estimates and best linear unbiased predictors (BLUPs). After fitting the linear mixed model, the fitted values are extracted and used in the survival model as time varying covariates (random effects) as in the following steps (Bycott and Taylor 1998):

Step 1: Fit the linear mixed model as in model (2.1).

Step 2: Extract subject specific predictions from the fitted linear mixed model.

Step 3: Use these predictions as time-dependent covariates in the survival model.

The two-stage model is presented as follows:

$$\mathbf{Y}_i(t) = \mathbf{m}_i(t) + \epsilon_i; \quad \boldsymbol{\epsilon}_i \sim N(0, \sigma^2), \tag{4.1}$$

where

$$\mathbf{m}_i(t) = \mathbf{X}'(t)\boldsymbol{\beta} + \mathbf{Z}'_i(t)\mathbf{b}_i. \tag{4.2}$$

which are predictions obtained for subjects at time t to be used as time-dependent covariate in the model below:

$$h_i(t|\mathbf{Y}_i(t), X_i) = h_0(t)\exp\{\mathbf{X}\boldsymbol{\beta} + \alpha\mathbf{m}_i(t)\}. \tag{4.3}$$

In models 4.2 and 4.3 the fixed effects are the same, including those used in the shared parameter joint model.

4.1.1 A two-stage approach for the joint model

We adopted and introduce some notation from Rizopoulos, Verbeke, and Molenberghs (2010). Let T_i^* denote the true event time for subject i, T_i is the observed event time for subject i, δ_i is the event indicator (which equals one for a true event and zero otherwise), and, finally $Y_{ij}, i = 1, 2, \ldots, n; j = 1, 2, \ldots, n_i$ are the observed longitudinal responses. The two-stage joint model was formulated in two steps: The first step was to fit the linear mixed model to the data and the fitted linear mixed model is given by;

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{4.4}$$

where $\mathbf{Y}_i$ is a vector of observed outcomes for the subject, $\mathbf{b}_i$ are random effects and assumed to be distributed as $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G})$; $\mathbf{G}$ is the covariance matrix for the random effects and residuals; $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{R})$, where $\mathbf{R} = \sigma^2\mathbf{I}_i$; $\mathbf{X}_i$ and $\mathbf{Z}_i$ are design matrices for fixed effects and random effects respectively. It is usually assumed that R $= \sigma^2\mathbf{I}_i$.

The next step was to extract subject specific predictions from the fitted linear mixed model (4.4), which takes into account measurement error and are smoothed over measurement time points. These predictions were then used as time-dependent covariates in the extended survival

model:

$$h_i(t|m_i(t)) = h_0(t)\exp\{\mathbf{X}_i'\boldsymbol{\tau} + \hat{m}_i(t)\}, \tag{4.5}$$

where $\hat{m}_i(t) = \{m_i(s), 0 \leq s < t\}$ is the fitted value for the longitudinal history of the i^{th} subject ($\mathbf{Y}_i$) and $\mathbf{X}_i$ is a matrix for fixed effects and $\boldsymbol{\tau}$ is a vector of coefficients for the fixed effects. The fixed effects $\mathbf{X}_i$ are primarily involved in the second step.

In the two-stage model, the uncertainty in the first stage is not carried through to the second stage as noted by Sweeting and Thompson (2011). The problem of uncertainty not being carried over to the second stage has been addressed by Mauff et al. (2018) using importance sampling in the Bayesian framework.

4.1.2 The variance shift outlier model (VSOM)

The objective was to identify possible outlying subjects or observations at the first stage. We adopted the variance shift outlier model of Gumedze et al. (2010) for down-weighting outliers. The variance shift outlier model does not remove outlying subjects or observations in the modelling process as do standard case-deletion diagnostics. The removal of outlying or influential observations in the estimation process may under power the study or throw away useful information, on the other hand, VSOM down-weights outliers using the shift in variance implied by outliers (Gumedze et al. 2010). A two level VSOM was implemented in the linear mixed model of the two-stage joint model to identify outliers and down-weight them before extracting predicted values to be used as time-varying covariate in the survival model. This involved fitting a VSOM at the subject level i.e. for (n) subjects and then a VSOM at the observation level i.e. n_i observations. We fitted a combined VSOM, which included outliers

identified at subject level and those identified at observation level. If a subject was identified as outlying and had simultaneously observation(s) that were identified as outlying, the corresponding observation would not be included in the fitting of the combined VSOM. The reason for this was so to avoid over-fitting the model. The fitted values ($\hat{\mathrm{Y}}_i$) were then extracted from the combined VSOM and used as time-varying covariate in the extended Cox model.

The results from the two-stage joint model, with a combined VSOM at the first stage, were contrasted with those from the standard extended Cox model using AIC and BIC. Model checking was performed using the influence diagnostics: dfbetas and likelihood displacement under a VSOM at subject level, VSOM at observation level and the combined VSOM. The results from the preceding implementation were compared.

The VSOM for the j^{th} observation from the i^{th} subject is given by Gumedze et al. (2010);

$$\begin{aligned}\mathbf{Y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b} + \mathbf{d}_{ij}\zeta_{ij} + \boldsymbol{\epsilon} \\ &\sim N(\mathbf{X}\beta, \sigma^2(\mathbf{ZGZ}' + \omega_{ij}\mathbf{d}_{ij}\mathbf{d}_{ij}' + \mathbf{R})) \\ &\equiv N(\mathbf{X}\beta, \sigma^2\mathbf{V}),\end{aligned} \tag{4.6}$$

where $\mathbf{Y}, \mathbf{X}, \mathbf{Z}$ and $\boldsymbol{\epsilon}$ are as defined in equation (4.4), $\mathbf{d}_{ij}$ is the column vector of length n_i with a single non-zero entry whose value is one in the j^{th} position and ζ_{ij} is an unknown random coefficient $\zeta_{ij} \sim N(0, \omega_{ij}\sigma^2)$ for $\omega_{ij} > 0$;

$$\mathbf{V} = \mathbf{ZGZ}' + \omega_{ij}\mathbf{d}_{ij}\mathbf{d}_{ij}' + \mathbf{R}, \tag{4.7}$$

where if $R = \sigma^2\mathbf{I}$ then (4.6) can be written as;

$$\mathrm{Var}(\mathbf{Y}) = \sigma^2(\mathbf{ZGZ}' + \omega_{ij}\mathbf{d}_{ij}\mathbf{d}_{ij}' + \mathbf{I}) = \sigma^2\mathbf{V}. \tag{4.8}$$

The variance of the j^{th} observation is inflated as $\sigma^2(\mathrm{Z}_{ij}'\mathrm{G}\,\mathrm{Z}_{ij} + \omega_{ij} + R_{ij})$, where z_{ij}' is the i^{th} row of $\mathbf{Z}$ and R_{ij} is the j^{th} diagonal element in $\mathbf{R}$.

If the variance shift for the j^{th}observation is denoted ω_{ij}, which, in essence is the variance shift

for the j^{th}observation associated with the random effect b_{ij}. The magnitude of this variance shift determines whether the j^{th} observation is an outlier or not. Observations with a relatively large variance shift are considered outlying.

Extensions to non-normal data for the i^{th} subject and the j^{th} observation being outlier with inflated variance can be formulated using hierarchical generalised linear modelling techniques (HGLMs) (Gumedze and Chatora 2014).

The VSOM for the i^{th} subject based on Gumedze et al. (2010) is given by:

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \mathbf{d}_i\zeta_i + \epsilon \\ &\sim N(\mathbf{X}\beta, \sigma^2(\mathbf{ZGZ}' + \omega_i\mathbf{d}_i\mathbf{d}_i' + \mathbf{R})) \\ &\equiv N(\mathbf{X}\beta, \sigma^2\mathbf{V}),\end{aligned} \tag{4.9}$$

where $\mathbf{d}_i$ is the column vector of length n with a single non-zero entry whose value is one in the i^{th} position and ζ_i is an unknown random coefficient; $\zeta_i \sim N(0, \omega_i\sigma^2)$ for $\omega_i > 0$;

$$\mathbf{V} = \mathbf{ZGZ}' + \omega_i\mathbf{d}_i\mathbf{d}_i' + \mathbf{R}, \tag{4.10}$$

where if $R = \sigma^2\mathbf{I}$ then equation(4.9)can be written as;

$$\text{Var}(\mathbf{Y}) = \sigma^2(\mathbf{ZGZ}' + \omega_i\mathbf{d}_i\mathbf{d}_i' + \mathbf{I}) = \sigma^2\mathbf{V}. \tag{4.11}$$

The variance of the i^{th} subject is inflated as $\sigma^2(\mathbf{Z}_i'\mathbf{GZ}_i + \omega_i + \mathbf{R}_i)$, where z_i' is the i^{th} row of $\mathbf{Z}$ and $\mathbf{R}_i$ is the i^{th} block diagonal element in $\mathbf{R}$.

If the variance shift for the i^{th} subject is denoted ω_i, which is the variance shift for the i^{th} subject associated with the random effect b_i. The magnitude of this variance shift determines whether the i^{th} observation is an outlier or not. Subjects with a relatively large variance shift are considered outlying.

4.1.3 Descriptive summary of the variance shift outlier model (VSOM)

Outliers are data points that fall outside expected range of the rest of the data points i.e. these data points are far removed from the rest. Outliers are common in experimental research as a result of a number of reasons such as transcription errors or malfunction in an experimental equipment. In most cases outliers are quickly identified and action taken about them as per prescribed methods in the analysis plan. It is sometimes not clear how to deal with outliers especially when they are anomalous. There are established statistical techniques that detect outliers in linear models such as the fixed effects model with extensions in the linear mixed effects models. The traditional approach for detecting outliers is the case deletion technique which may ultimately lead into excluding outliers from the final modelling process at the discretion of the investigator. The variance shift outlier model is based on the assumption that there is a shift in the variance parameter δ_i for the i^{th} data point herein after called subject where δ_i is not known in advance but estimated from the observed data. The estimated values of δ_i represent subjects likely to have inflated variance compared to the rest of the subjects under consideration and therefore considered as outliers. The variance shift outlier model offers an alternative to include "anomalous subjects" in the analysis by down-weighting them using the variance shift estimate ω_i which in turn gives subject i an opportunity to be consistent with the rest and thereby get included in the analysis. This exercise has the advantage of maintaining the study power if the sample size is barely adequate.

4.1.4 Hypothesis tests on variance shift parameters

Hypothesis tests for the variance shift parameters are formulated using standard procedures for hypothesis testing. The null hypothesis was that the subject or observation does not have inflated variance ($H_0 : \omega_i = 0$) versus the alternative that the subject or observation has inflated variance ($H_1 : \omega_i > 0$). This was done using likelihood ratio test (LRT) statistic after accounting for the profile likelihood (Lee, Nelder, and Pawitan 2006). This statistic is presented as

$$\text{LRT}_i = -2\Big[\ell(\hat{\theta}) - \ell(\theta_0)\Big], \tag{4.12}$$

where $\ell(\hat{\theta})$ is the log-likelihood for the alternative model and $\ell(\theta_0)$ is the log-likelihood of the null model.

For the reason that under the null hypothesis, the variance shift parameter is zero, hence falls in the boundary of the parameter space; it is impossible to use asymptotic theory for the distribution of the LRT statistics. This is so because, it defies regularity conditions (Self and Liang 1987). Fitting the VSOM (Gumedze et al. 2010), showed that the $0.68\chi^2_0 + 0.32\chi^2_1$ mixture distribution performed just as well, compared to simulation studies by Pinheiro and Bates (2000), which qualified $0.65\chi^2_0 + 0.35\chi^2_1$ mixture of distribution as appropriate. To solve the problem of this unknown distribution and that of multiple testing, Gumedze et al. (2010) proposed a parametric bootstrapping approach.

To solve regularity conditions under the null hypothesis, that the variance shift parameter is zero hence lies in the boundary of the parameter space, we used the $0.65\chi^2_0 + 0.35\chi^2_1$ mixture distribution on p-values of the likelihood ratio tests. The problem of multiple testing was addressed by adjusting the p-values using the Benjamini and Hochberg (1995) (B-H) procedure.

The B-H procedure is performed in the following four steps:

1. Put individual $p-$values in ascending order: $p_1 \leq p_2 \leq \cdots \leq p_m$ corresponding to the respective null hypothesis H_i
2. Assign ranks to the $p-$values. As an example the smallest has a rank of 1, the second smallest has a rank of 2.
3. Calculate each individual $p-$value's B-H's critical value, using the formula $\left(\frac{i}{m}\right)Q$, where;
   a) i is the individual $p-$value's rank.
   b) m is total number of tests.
   c) Q is the false discovery rate which is a percentage chosen by the researcher.
4. Compare the original p-values to the critical B-H values from the previous step and find the largest $p-$value that is smaller than the critical value.

The B-H procedure is broadly valid under three main assumptions i) when tests are independent ii) when tests are independent and continuous iii) when there is positive dependence in the tests. The main limitation of applying the BH procedure on the VSOM is that the tests may not be consistent with the three assumptions.

4.1.5 Model fitting

The linear mixed model were fitted using the LME4 Package in R, while the VSOM used the function *mat.or.vec(nrow(dataframe), 1)* to create a vector of zeroes elsewhere except at the position of the subject where there was a one. The extended Cox model was implemented

using the Survival Package and all hypothesis tests on variance parameters were done in R using the procedures descried in Section 4.1.4 above.

In the two-stage model the uncertainty in the first stage is not carried through to the second stage (Sweeting and Thompson 2011) and, therefore, a model that accounts for measurement error, utilises all available repeated measurements and reduces bias with a potential of maximising efficiency, is based on the joint likelihood of the two processes and these are longitudinal and survival processes discussed in the following Section.

4.2 Diagnostics for two-stage joint survival model using variance shift outlier model

Introduction

Two-stage joint survival model is a class of models based on extended Cox survival models with time-varying covariates. The two-stage joint model is considered to be superior over standard modelling of survival outcomes with a time-varying covariate because they account for measurement error in a time-varying covariate and do not assume that the value of the covariate remains constant between any two measurements (Self and Pawitan 1992). This model is, particularly, useful because it addresses a research question that has the two components and these are longitudinal and time-to-event (Tsiatis and Davidian 2001). The validation of this model is done using standard Cox survival models, which is categorised into four parts, namely: overall fit of the model, excess survival times, proportionality assumptions and influence diagnostics (Collett 2015). There is vast literature on these four aspects of model assessment, but in this thesis, we only focus on influence diagnostics, which tis the main ob-

jective of this research.

Excess survival times are assessed using martingale residuals that have a basic interpretation of being the difference between the observed and expected number of events in the time interval $(0, t_{\mathrm{i}})$, for the i^{th} individual (Therneau, Grambsch, and Fleming 1990). These residuals highlight individuals who have experienced an event too soon or too late on the basis of the assumed model herein called excess survival times. An index plot of martingale residuals will highlight individuals whose survival time is not well fitted by the model and these subjects are considered as outliers. Plotting of these residuals against survival time, the rank order survival times, or explanatory variables, may indicate whether there are particular times, or values of the explanatory variables, where the model does not fit the data. Deviance residuals are motivated by the fact that they are components of the deviance; which is a statistic used to summarise the extent to which the fit of a model of interest deviates from that of the model, which is a perfect fit of the data (Davison, Gigli, and A 1989). The deviance residuals can be looked at as transformed martingale residuals that produce values that are symmetric about zero when the fitted model is appropriate (Therneau, Grambsch, and Fleming 1990).

DFBetas are used to identify not only subjects but also observations that exert influence on the parameter estimation. DFBetas compare the estimated parameter $\hat{\beta}$ and $\hat{\beta}_{[i]}$ where $\hat{\beta}$ is the parameter estimate from the model and $\hat{\beta}_{[i]}$ is the estimate when subject i is removed from the model. If the difference between $\hat{\beta}_{[i]}$ and $\hat{\beta}$ is large, then subject i is deemed to be exerting influence on parameter estimation. An alternative influence diagnostic to DFBetas proposed by Escobar and Meeker Jr (1992) is the likelihood displacement value and $\ell_{\max}$. An $\ell_{\max}$ value is a measure of influence resulting from deleting a subject on the overall vector of parameter. An $\ell_{\max}$ value is based on an eigen system analysis of efficient score residuals (Pettitt and

Daud 1989). Both likelihood displacement and $\ell_{\max}$ examine the subjects' influence on the entire vector of coefficients. The idea behind likelihood displacement is that, if subject i is influential, then the vector $\hat{\beta}_{[i]}$ will be different from the vector $\hat{\beta}$, when this situation arises examining the log-likelihood at such a solution results in a different log-likelihood (Pettitt and Daud 1989).

Extensive research has been done on outliers in linear mixed models to identify influential observations. The developed techniques lead to case-deletion methods proposed by Cook (1977) and Cook and Weisberg (1982). Application of these ideas have also been adopted in the analysis of time-to-event data (Hall, Rogers, and Pregibon 1982; Storer and Crowley 1985) as well as in logistic regression (Hosmer, Jovanovic, and Lemeshow 1989; Pregibon 1981).

Case-deletion diagnostics assumes $\boldsymbol{\theta}$ to be a parameter vector of interest in the model under consideration and $\hat{\boldsymbol{\theta}}$ an estimate of $\boldsymbol{\theta}$ using appropriate techniques usually maximum likelihood in the frequentist paradigm (Cook and Weisberg 1982). If one excludes the i^{th} subject from the full set of observations; refitting the model under the same principle produces $\hat{\boldsymbol{\theta}}_{[-i]}$ of $\boldsymbol{\theta}$ which gives a measure of the difference between $\hat{\boldsymbol{\theta}}_{[-i]}$ and $\hat{\boldsymbol{\theta}}$ which is used to evaluate how the i^{th} subject impacts the estimate $\hat{\boldsymbol{\theta}}$ and subsequent inferences drawn from the model and measure is called the Cook's distance.

There has been a lot of effort to investigate influence diagnostics in the linear mixed model. In their work, Atkinson (1985), Barnett and Lewis (1995), Chatterjee and Hadi (1988), Demidenko (2013), Demidenko and Stukel (2005), Pan, Fei, and Foster (2014), Rousseeuw and Van Zomeren (1990), Zewotir and Galpin (2005), Zhang et al. (2014), and Zhu et al. (2012), extended a number of regression diagnostic approaches that are in use in linear regression to linear mixed effects model, which have been reviewed in Chapter 2 Section 2.2.

In their work Schall and Dunne (1988) developed a unified approach to outliers in the general linear models, by showing that testing for the different types of outliers in the linear model $(Y, X\beta, \sigma^{\mathbf{2}}, V)$, is equivalent to testing for outliers in the linear model $(\hat{\epsilon}, 0, \sigma^{\mathbf{2}}\mathcal{N})$, where $\hat{\epsilon}$ is the vector of estimated residuals from the original model, and $\sigma^{\mathbf{2}}\mathcal{N}$ is its variance-covariance matrix.

The variance shift outlier model (VSOM) by Gumedze et al. (2010), detects outliers and down-weights influential subjects or observations using shift in variance parameters. Therefore, this model does not exclude outliers and/or influential subjects or observations in the modeling, thereby, throwing away potentially useful information. The model scans throw the data and detects subjects or observations with inflated variance and down-weights them to be constituent with the rest of the data.

It is not fully understood what impact influential subjects and/or observations in the linear mixed model would have in parameter estimates when carried through to the next stage of the two-stage joint survival model. In this thesis, we postulate the implementation of the VSOM in the linear mixed model to identify outliers and at the discretion of the researcher down-weight these outliers instead of excluding them from the analysis. This is in line with what was proposed by Gumedze and Chatora (2014) and Gumedze et al. (2010); where situations may arise that a researcher does not have enough data points, as such excluding outliers will not only throw away useful information but also under power the study. This is the main motivation for the implementation of a VSOM at the first stage of the two-stage joint survival model coupled with the fact that estimates from a two-stage joint survival model are reliable and credible when importance sampling is used (Mauff et al. 2018, 2020). It is also known that a two-stage joint survival models are not computer intensive even in the presence of an

increasing number of random effects (Sweeting and Thompson 2011) hence their adopting is widely expected with the recent development by Mauff et al. (2020).

4.2.1 Case deletion versus variance shift outlier model

Case deletion is an extreme case variance outlier model (VSOM). In case deletion parameter estimates are iteratively re-estimated after excluding the i^{th} observation hence quantifying the effect of the i^{th} observation on the fit of the model. This is done by computing the change in aspects of the model that take place after removing the i^{th} observation. This approach has roots from linear models with extensions to linear mixed effects models (Belsley, Kuh, and Welsch 1980; Chatterjee and Hadi 1988). The variance shift outlier model adopted from Cook and Weisberg (1982) by contrast; is a method for outlier detection and accommodation in both linear models as well as linear mixed effects models. The VSOM assumes the variance of all observations except for one unknown unit is σ^2 (Gumedze and Chatora 2014; Gumedze et al. 2010). The variance for the unit i for example is assumed to be $\alpha_i\sigma^2(\alpha_i \geq 1)$. The unknown parameter α_i is estimated from from the data and acts as a variance inflator in the variance for the i^{th} unit (Gumedze et al. 2010). The variance shift outlier model seeks to consider units with inflated variance as possible outliers. This emphasis is different from case deletion methods despite a direct link when the variance shift parameter for a unit goes to infinity as observed by Gumedze et al. (2010).

4.3 Illustration using VSOM - IMPI dataset

As described by Gumedze and Chatora (2014), there are two ways of implementing VSOM and these are, firstly, scanning the dataset for outliers by fitting a null model at the observation or subject level and the next stage is focused on down-weighting the outliers in the analysis. Extra care has to be taken when deciding whether outliers are due to errors in the dataset in which case corrections can be done or they (outliers) are unexplained data points hence they could be candidates for down-weighting. Both of the above decisions are made in line with the study aims as well as the application area for the analysis at the discretion of the researcher. We fitted model 4.6 and 4.9 to the data which are respectively at observation and subject level. The VSOM, ordinarily, proceeds by identifying outlying observations or subjects and then down-weighting those observations or subjects that have inflated variance and viewed as outliers. This is the second step and is done when necessary or when desired by the investigator. In this implementation, we fitted four models, one model is the linear mixed model which is considered the null model. We, then, fit the VSOM observation level, thereafter we fit VSOM at subject level and lastly we fit a combined VSOM. At the subject, level we don't include observations that are identified as outlying for the subject which in a way deals with over-fitting. The following notation was adopted, $k = 1, ..., N$, where $N = 587$ was the total number of subjects that had been indexed in the dataset. This, therefore, meant that the variance shift parameter was denoted d_k, for $k = 1, ..., 587$.

The CD4 cell count measurements were mandated at baseline, week two, month one, three and six. In South African centres, on the other hand, CD4 measurements continued beyond the mandated period. The analysis of CD4 cell count is based on patient identifies (pid) and

their respective time point at which CD4 cell count was done. We provide the following table as a guide to identifying observations in the analysis:

Table 4.1: Observation identifier guide.

Pid	CD4 visit time	Observation identifier
5	baseline (0)	5_0
5	week two (0.5)	$5_{0.5}$
5	month one (1)	5_1
5	month three (3)	5_3
5	month six (6)	5_6

It was shown that observations; $557_0, 557_{0.5}, 552_6,$ and $564_{0.5}$ were outlying in all the three statistics used, namely variance shift (ω_i), residuals (σ_e^2) and the likelihood ratio test (Figure 4.1). Observation 45_0 was outlying for residuals and likelihood ratio test statistics.

The definition of observations is taken from concatenating the participant study number with the visit number at which CD4 cell count measurement was done e.g. $557_0, 557_{0.5}, 552_6$ were observations that belonged to participant 557 at baseline visit, 557 at week two and 552 at week six respectively.

It was observed that fitting the linear mixed model and implementing the VSOM at both the observation and subject level improves the model using the AIC and BIC statistics. There was also improvement in residuals after combined VSOM was fit to the data.

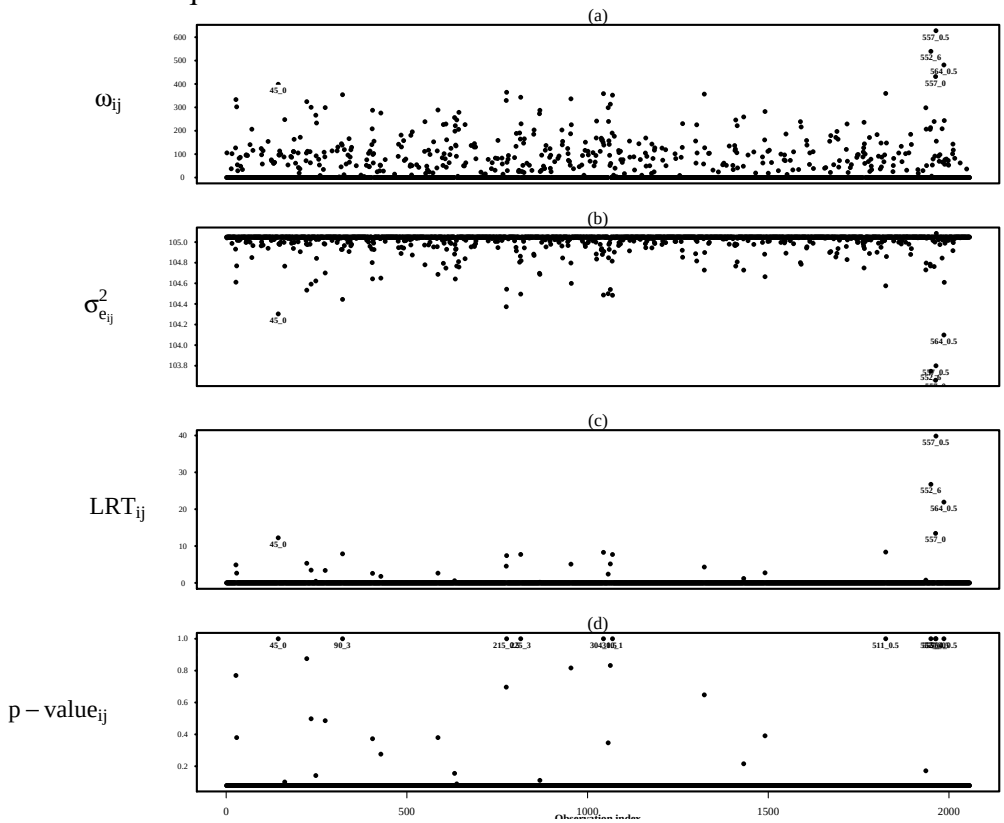
The following subjects ($144, 235, 248, 322, 404, 405, 428,$ and 27) were identified as outlying as tabulated in Table 4.2. A subject level VSOM was implemented with an effect of improving the overall model fit (AIC for M_2 = 11652 vs AIC for M_3 = 11650.02) which compares observation level VSOM against subject level VSOM.

None of the subjects identified as outlying by the first step in VSOM experienced the event of interest, as such, they were all censored administratively at the last study visit. These subjects

were characterised as follows; Subject 27 and 144 had three CD4 cell measurements taken at baseline, three months and six months with (338, 314 and 465 $\mu\ell$) and (140, 306 and 382 $\mu\ell$) values, respectively. Subject 27 was randomised to the placebo arm and was not started on ART, while, subject 144 was randomised to the treatment arm and ART experienced. Subject 235 also had three CD4 cell count measurements taken at two weeks, three and six months (78, 116 and 153 $\mu\ell$), started taking ART at three months' visit and was randomised to the placebo arm. Subject 248 had all CD4 cell count measurements (105, 81, 351 $\mu\ell$) taken at baseline, two weeks and three months respectively. The subject was randomised to the placebo arm, and exposure to ART was reported at month three and six. Subject 322 had three CD4 cell count measurements taken at two weeks, one and three months (106, 367 and 448 $\mu\ell$). The subject was randomised to the placebo arm and was exposed to ART from baseline, also referred to as ART experienced. Subject 404 and 405 had all CD4 cell count measurements taken; at baseline, week two, one month, three months and six months. The following are measurements for subject 404: 226, 324, 261, 210, 350 $\mu\ell$. Subject 405 had the following measurements: 253, 289, 687, 434, 433 $\mu\ell$. They were both ART experienced, with subject 404 reporting exposure at the last visit having been randomised to a placebo arm, while subject 405 reported exposure to ART at the second visit; and was randomised to the treatment arm. Subject 428 had four CD4 measurements taken except the last visit (291, 308, 271, and 315 $\mu\ell$), and was radomised to the placebo arm. Subject 552 had three CD4 cell measurements taken at baseline, two weeks and six months (134, 136, 1020 $\mu\ell$), equally, subject 557 had CD4 cell count measurement taken at baseline, two weeks and three months (453, 1544, and 1125 $\mu\ell$). Both of these subjects were randomised to the treatment arm with subject 557 having no exposure to ART, while, subject 552 was ART experienced. Subject 564 was randomised to the treatment

arm, and had the following CD4 cell count measurements: 368, 1181, 798, and 614 ($\mu\ell$) taken at two weeks, one month, three months and six month, respectively. This subject reported ART exposure at the three months' visit.

The next illustration used the same data to fit a standard Cox model then a Cox model with CD4 cell count as a time-varying covariate. We went further to calculate diagnostics for the two setting using DFBetas and likelihood displacement. We restricted the modelling to treatment and square root CD4 to be consistent with the linear mixed model.

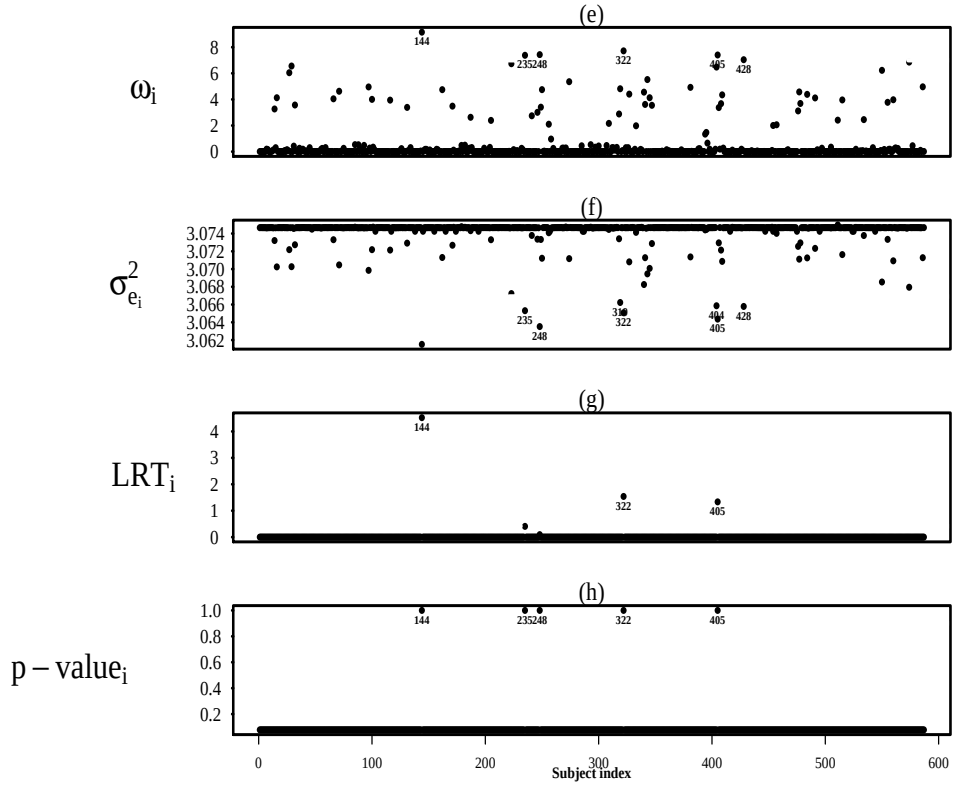


(a) variance shift ω_i, (b) residual σ^2_e, (c) likelihood ratio test (LRT) and (d) BH adjusted p-value

Figure 4.1: Observation level VSOM statistics plotted against observation index.

The p-values in Figure 4.1 and 4.2 were adjusted by using the B-H procedure described in Section 4.1.4, which are results of multiple testing from a $0.68\chi^2_0 + 0.32\chi^2_1$ mixture distribution. The largest p-value, $p < (i/m)Q$, was considered to be significant and all p-values smaller

than it were also significant.



(e) variance shift ω_i, (f) residual σ_e^2, (g) likelihood ratio test (LRT) and (h) BH adjusted p-value

Figure 4.2: Subject level VSOM statistics plotted against subject index.

These p-values included, even, the ones that were more than the respective B-H critical values. The following subjects were considered as having large p-values after using the B-H procedure: 144, 235, 248 and 322. A similar pattern was observed at observation level, where observations belonging to corresponding subjects with large p-values were identified.

The subject level implementation of the variance shift outlier model is shown in Figure 4.2.

Table 4.2: Parameter estimates after fitting VSOM to the linear mixed model using IMPI dataset.

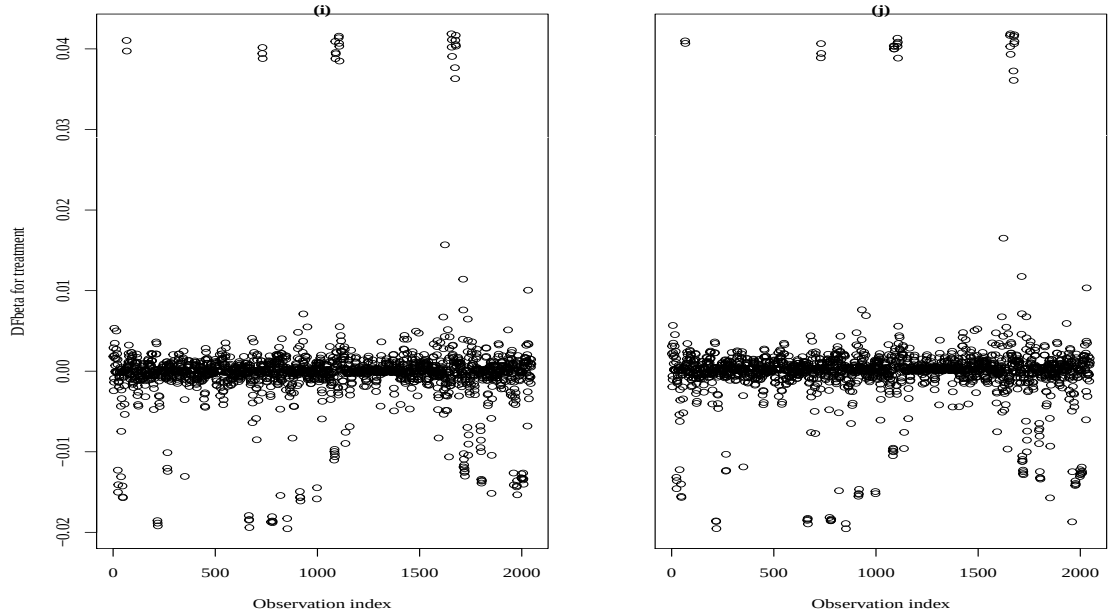
Parameter	LMM[1] Estimate (Std. Error)	VSOM[2] Estimate (Std. Error)	VSOM[3] Estimate (Std. Error)	VSOM[4] Estimate (Std. Error)
Intercept	13.37 (0.28)	39.62 (14.61)	35.42 (15.45)	55.07 (20.16)
Treatment	-0.73 (0.40)	-0.78 (0.40)	-0.73(0.39)	-0.77 (0.40)
Time in months	0.63 (0.03)	0.62 (0.03)	0.62 (0.03)	0.62 (0.03)
σ_e^2	9.45	9.11	8.99	8.73
σ_a^2	20.47	20.17	20.57	20.16
$\omega_{45_{-0}}$	-	84.40	-	83.97
$\omega_{552_{-6}}$	-	129.77	-	131.22
$\omega_{557_{-0}}$	-	4.01	-	5.53
$\omega_{557_{-.5}}$	-	97.39	-	93.09
$\omega_{564_{-.5}}$	-	82.75	-	81.84
ω_{144}	-	-	83.87	-
ω_{235}	-	-	55.19	55.60
ω_{248}	-	-	58.88	59.42
ω_{322}	-	-	63.22	63.37
ω_{404}	-	-	26.35	26.54
ω_{405}	-	-	41.99	42.05
ω_{428}	-	-	50.50	50.87
ω_{27}	-	-	51.64	51.83
-2 loglikelihood	11705.32	11652.36	11650.02	11606.39
AIC	11715.32	11672.36	11676.02	11640.39
BIC	11743.47	11728.65	11749.20	11736.09

Models: 1: Linear mixed model; 2: VSOM at observation level; 3: VSOM at subject level; 4: Combined VSOM
[a] variance shift for subject or observation i (ω_i), random effects (σ_a^2) and residual variance(σ_e^2)

4.4 Illustration using two-stage joint model and VSOM

There were 27 constriction cases during the six months' evaluation period. The variable selection for the event constriction had the following risk factors evaluated at 10% level: Heart rate, sex, NYHA class, peripheral oedema, definite tuberculosis and haemoglobin status.

The extended Cox model was fitted using square root CD4 as a time-varying covariate, then went further to fit a two-stage joint model. Thereafter, we fitted the two-stage joint model having implemented VSOM at observation, subject and combined VSOM. In all the implementations, we obtained DFbetas and likelihood displacement to assess influence of observations and subjects as in Figure 4.3. We noted from plot **a** and **b** that after implementing a combined VSOM using the two-stage approach there were fewer observations that appeared to be influential compared to the extended Cox model implementation.



(i) extended Cox model (j) combined two-stage VSOM

Figure 4.3: An index plot of DFBetas.

The same pattern was evident and consistent when a different statistic was implemented to reveal influential observations. The statistic used on the overall fit of the model to assess influence of observations is the likelihood displacement, which is similar to equation 4.12. This is based on score residuals and evaluated using Cook's distance techniques. It is clear from the two-stage combined VSOM (Figure 4.4 **d**) that, some influential observations have been down-weighted. The implementation of VSOM to the two-stage joint model has been consistent with the rest except. When we compared treatment effect after fitting extended Cox model with CD4 as a time-varying covariate, then a standard two-stage joint model, and further, the observation and subject level VSOM, there were no significant changes observed in the treatment effect. There was on the other hand, a better overall fit of the model using a two-stage joint model with a combined VSOM based on AIC and BIC (Table 4.3).

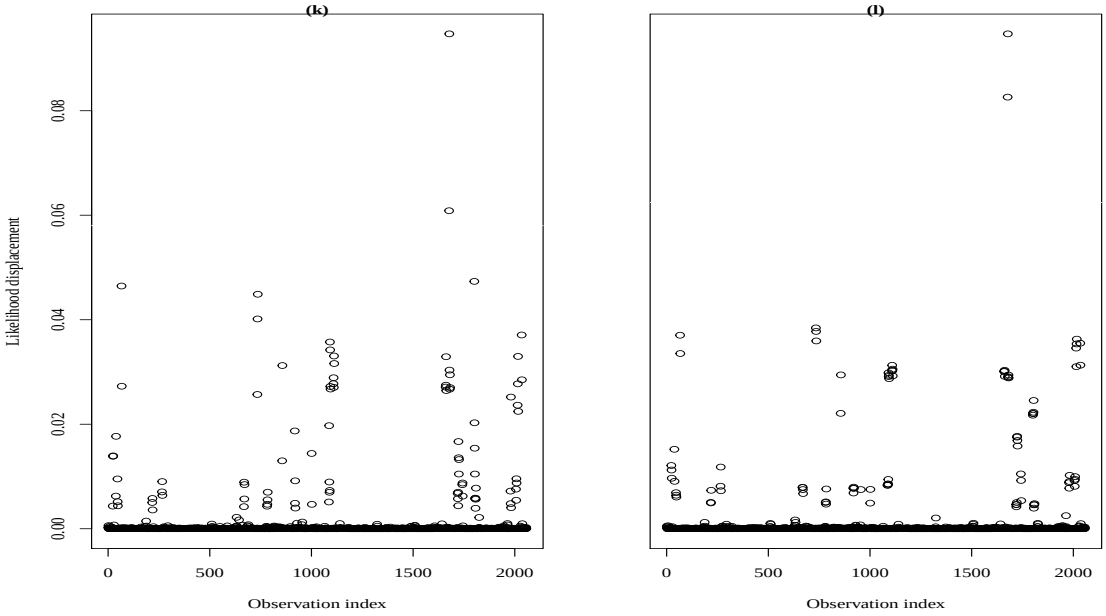


Figure 4.4: An index plot of Likelihood displacement.

Table 4.3: Parameter estimates for model comparison using IMPI dataset.

	Model estimates				
Characteristic	Extended Cox	two-stage JM	two-stage JM with $VSOM_1$	two-stage JM with$VSOM_2$	two-stage JM with $VSOM_3$
	H.R. (95 % C.I.)	H.R. (95 % C.I.)	H.R. (95 % C.I.)	H.R. (95 % C.I.)	H.R. (95 % C.I.)
Prednisolone	0.41 (0.26 , 0.65)	0.42 (0.27 , 0.66)	0.42 (0.27, 0.66)	0.42 (0.27, 0.66)	0.42 (0.27, 0.66)
$\sqrt{\text{CD4}}$ cell count	1.04 (1.00, 1.08)	1.05 (1.01, 1.11)	1.05 (1.01, 1.10)	1.05 (1.01, 1.11)	1.05 (1.01, 1.10)
NYHA class: III-IV	1.57 (1.04, 2.38)	1.58 (1.05, 2.39)	1.58 (1.05, 2.39)	1.58 (1.05, 2.40)	1.58 (1.05, 2.39)
Sex: Male	1.63 (1.02, 2.62)	1.68 (1.05, 2.69)	1.67 (1.04, 2.68)	1.68 (1.05, 2.69)	1.67 (1.04, 2.68)
Definite TB	2.23 (1.46, 3.41)	2.28 (1.49, 3.48)	2.28 (1.49, 2.48)	2.28 (1.49, 3.48)	2.27 (1.49, 3.48)
Peripheral Oedema	1.97 (1.30, 2.97)	1.97 (1.30, 2.97)	1.97 (1.30, 2.97)	1.97 (1.31, 2.97)	1.97 (1.06, 2.97)
Haemoglobim status	1.18 (1.07, 1.31)	1.18 (1.06, 1.30)	1.18 (1.06, 1.30)	1.18 (1.06, 1.30)	1.18 (1.06, 1.30)
AIC	1113.59	1112.38	1112.51	1112.41	1112.53
BIC	1131.32	1130.10	1130.24	1130.14	1130.26

Baseline covariates:- Prednisolone: Treatment vs Placebo; NYHA class: NYHA functional class, stage I-II vs III-IV; Sex: Female vs Male; Definite TB: No vs Yes; Peripheral oedema: NO vs Yes; Haemoglobin status ≤ 10 vs > 10

Models:- two-stage JM with $VSOM_1$ is the two-stage joint model on which vsom is implemented at observation level, two-stage JM with$VSOM_2$ is the two-stage joint model on which vsom is implemented at subject level, and, two-stage JM with $VSOM_3$ is the two-stage joint model on which vsom is implemented combining observations and subjects

The goodness of fit performance for the two-stage joint survival model based on Cox-Snell residuals was better than the standard extended Cox survival model (Figure 4.5). A similar trend was observed when the VSOM was implemented at observation and subject level with subtle differences between the two.

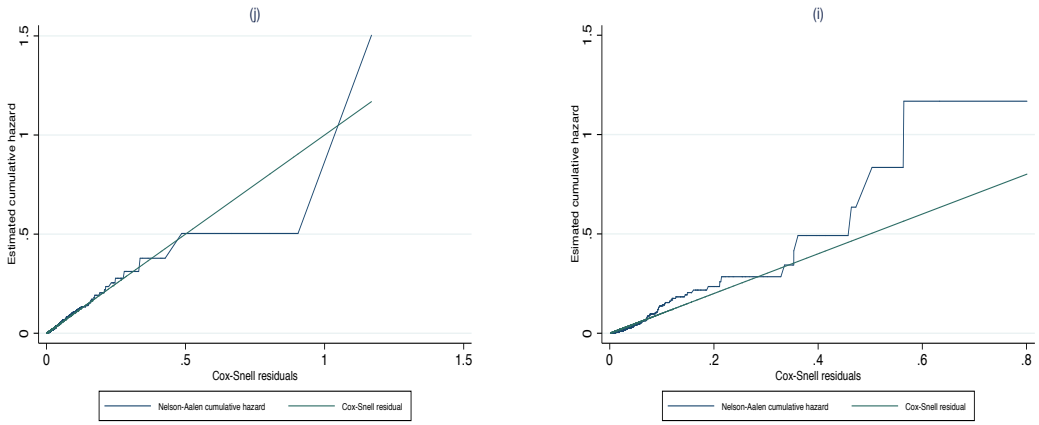


Figure 4.5: Goodness of fit using Cox-Snell residuals

4.5 Summary

We have proposed and implemented a diagnostic approach for the two-stage joint model by fitting the variance shift outlier model on the linear mixed model, which is a sub-model in the two-stage joint model. This approach has been implemented on the multi-center trial on TB pericaditis. We note that, the VSOM identifies and down-weights outlying subjects as well as observations. When the VSOM is implemented at both the subject and observation level, then later combined as one VSOM whilst taking appropriate action to avoid over fitting; fitted values from the combined VSOM are used in the second stage of the two-stage joint model as time-varying covariate. In the linear model there is both improvement in the overall model fit contrasting AICs and BICs statistics after a combined VSOM as well as better estimates of parameters after VSOM.

There is no significant difference in parameter estimates when the two-stage joint model is implemented with the combined VSOM. There is, on the other hand, a better overall fit of the model if the combined VSOM has provided predicted values to be used as time-varying

covariate in the second stage of the joint model. This is evaluated using AICs and BICs from the model fitted to the data.

The possible reason for AICs and BICs to be similar in Table 4.3, could be due to the fact that there is no significant association between square root CD4 cell count and the risk for constriction.

Further work on the proposed approaches is recommended on the implementation of the two-level VSOM in competing risk setting. There is also a potential research area to assess and evaluate the approach in non-Gaussian data as a time-varying covariate.

Chapter 5

Joint models for longitudinal and survival data

Introduction

In this Chapter, we developed case deletion diagnostics for shared parameter joint model for longitudinal and survival data. The developed diagnostics are based on Cook's statistics (Cook and Weisberg 1982). The case deletion diagnostics that we developed are motivated by Zeng, Cai, et al. (2005)'s asymptotic properties of the likelihood from the joint conditional distribution of the joint model. We evaluated the performance the developed case deletion diagnostics in simulations studies and illustrated them using real data from a multi-centre clinical trial (IMPI dataset).

We used a shared parameter joint model for longitudinal and survival data from the types of joint models, *viz* shared parameter and latent class joint model. This class of joint model shares the random effects between the longitudinal and survival processes as follows (Tsiatis and Davidian 2004):

$$p(\mathbf{Y}_i, \mathbf{T}_i, \boldsymbol{\delta}_i) = \int p(\mathbf{Y}_i|\mathbf{b}_i)\{H(\mathbf{T}_i|\mathbf{b}_i)^{\boldsymbol{\delta}_i} S(\mathbf{T}_i|\mathbf{b}_i)\}p(\mathbf{b}_i)d\mathbf{b}_i, \tag{5.1}$$

where $\mathbf{b}_i$ is a vector of random effects that explains the interdependencies between the longitudinal density $p(.)$ and the survival density $S(.)$.

5.1 Parameter estimation in joint models for longitudinal and survival data

Parameter estimation is based on maximizing the log-likelihood for the joint distribution of survival and longitudinal outcomes $\{T_i, \delta_i, y_i\}$. This is based on the following three assumptions, namely; the vector of time-dependent random effects b_i underlies both the survival and longitudinal process, there is conditional independence based on the random effects and that parameter estimates are consistent and asymptotically normal as shown by Rizopoulos (2012) and Tsiatis and Davidian (2001).

The primary estimation method proposed for joint models for longitudinal and survival data are semi-parametric maximum likelihood (Henderson, Diggle, and Dobson 2000; Hsieh, Tseng, and Wang 2006; Tsiatis and Davidian 2001). The estimation process for the shared parameter joint model proceeds with a semi-parametric conditional score estimator, which treats random effects as nuisance parameters (Tsiatis and Davidian 2001). In this particular case estimating equations are deduced based on the derived sufficient statistic and parameter standard errors are subsequently estimated using sandwich matrix estimator.

The maximum likelihood estimates are derived as the modes of the log-likelihood function relating to the joint distribution of the observed outcomes $\{\mathbf{T}_i, \boldsymbol{\delta}_i, \mathbf{Y}_i\}$. Assuming that the vector of time-dependent random effects $\mathbf{b}_i$ are shared by both the survival and longitudinal processes, the joint conditional distribution is as follows (Rizopoulos 2012; Tsiatis and Davidian 2004):

$$p(\mathbf{Y}_i, \mathbf{T}_i, \boldsymbol{\delta}_i | \mathbf{b}_i; \boldsymbol{\theta}) = p(\mathbf{Y}_i, \mathbf{T}_i, \boldsymbol{\delta}_i |_i; \boldsymbol{\theta}) p(\mathbf{Y}_i | \mathbf{b}_i; \boldsymbol{\theta}) \tag{5.2}$$

and

$$p(\mathbf{Y}_i, \mathbf{T}_i, \boldsymbol{\delta}_i | \mathbf{b}_i; \boldsymbol{\theta}) = \prod_{j=1}^{n_i} p\{\mathbf{Y}_i(T_{ij}) | \mathbf{b}_i; \boldsymbol{\theta}\}, \tag{5.3}$$

where $\boldsymbol{\theta} = (\theta_T', \theta_Y', \theta_b')$ denotes the full parameter vector as follows, $\boldsymbol{\theta}_T$ denotes parameters for the survival outcome, $\boldsymbol{\theta}_Y$ denotes the parameters for the longitudinal outcome and $\boldsymbol{\theta}_b$ denotes the parameters for the random effects covariance matrix. Given that the observed history, the censoring mechanism and the visiting process are independent of the true event times and future longitudinal measurements, the marginal log-likelihood contribution for the i^{th} subject can be formulated as follows:

$$\begin{aligned} \log p(\mathbf{T}_i, \mathbf{Y}_i, \boldsymbol{\delta}_i; \boldsymbol{\theta}) &= \log \int p(\mathbf{T}_i, \boldsymbol{\delta}_i, \mathbf{Y}_i, \mathbf{b}_i; \boldsymbol{\theta}) d\mathbf{b}_i, \\ &= \log \int p(\mathbf{T}_i, \boldsymbol{\delta}_i, | \mathbf{b}_i; \boldsymbol{\theta}, \boldsymbol{\beta}) \left[\prod_{j=1}^{n_i} p\{\mathbf{Y}_i(T_{ij}) | \mathbf{b}_i; \boldsymbol{\theta}_b\} \right] p(\mathbf{b}_i; \boldsymbol{\theta}_b) d\mathbf{b}_i, \end{aligned} \tag{5.4}$$

There is no closed form solution for the above integral since the random effects have to be treated as nuisance parameters, therefore, maximising the log-likelihood function $\ell(\boldsymbol{\theta}) = \sum_i \log p(\mathbf{T}_i, \mathbf{Y}_i, \boldsymbol{\delta}_i; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ can be achieved using standard optimisation algorithms such as the Expectation-Maximisation (EM) (Dempster, Laird, and Rubin 1977) or the Newton-Raphson algorithm by Lange and Hunter (2004). The EM algorithm is mostly used in joint modeling (Rizopoulos 2012). The main approach for assessing the joint model has been to use separate analysis of residuals for the sub-models. This has been so, due to the fact that there is informative censoring as a result of the occurrence of an event. This, therefore, does not yield residuals that follow standard properties of randomness. The other approach proposed by Rizopoulos, Verbeke, and Molenberghs (2010) is the multiple imputation of the missing part of the longitudinal sub-model due to the occurrence of an event, thereby, creating a complete dataset.

5.1.1 Estimation of based maximum likelihood

The primary estimation method proposed for joint models for longitudinal and survival data and for brevity joint models are semi-parametric maximum likelihood (Henderson, Diggle, and Dobson 2000; Hsieh, Tseng, and Wang 2006; Tsiatis and Davidian 2001). There have been investigations done on the asymptotic properties of the semi-parametric maximum likelihood estimators under unspecified baseline risk function (Zeng, Cai, et al. 2005). Bayesian estimation methods for the joint models have also been explored using MCMC techniques (Chi and Ibrahim 2006; Lin and Wu 2006; Wang and Taylor 2001; Xu and Zeger 2001). The estimation process for the shared parameter joint model proceeds with a semi-parametric conditional score estimator, which treats random effects as nuisance parameters (Tsiatis and Davidian 2001). Estimating equations have been developed that are deduced based on the derived sufficient statistic and parameter standard errors are, subsequently, estimated using sandwich matrix estimator.

The maximum likelihood estimates are derived as the modes of the log-likelihood function relating to the joint distribution of the observed outcomes $\{T_i, \delta_i, y_i\}$. Maximising the log-likelihood function with respect to the parameter can be achieved using standard optimisation algorithms, such as the Expectation-Maximisation (EM): (Dempster, Laird, and Rubin 1977)) algorithm or the Newton-Raphson algorithm (Lange and Hunter 2004). The EM algorithm has been the mostly used in the joint modeling (Rizopoulos 2012).

Residuals are then derived from this dataset which are then plotted to assess the separate sub-models. We will consider case-deletion diagnostics based on the Cook's statistic for all parameters, fixed effects and variance component parameters in the joint model.

5.1.2 Joint model formulation

The joint model is formulated in three steps, proceeding with some notation as follows; T_i^* is true event time for subject i, T_i is the observed event time for subject i, δ_i is the event indicator, which equals one for true events and zero otherwise, and finally Y_i are the observed longitudinal responses. The first step is to assume that $\mathrm{M}_i(t)$ is known and this is the true and unobserved value of a longitudinal outcome at time point t. The second step is to reconstruct the covariate history for each subject from the observed longitudinal responses $\mathbf{Y}_i(t)$ and, thus, a mixed effects model. The two steps outlined above share common random effects and hence define a model for the joint conditional distribution due to Tsiatis and Davidian (2004).

5.1.3 Parameter estimation in joint models for longitudinal and survival data

Parameter estimation is based on maximizing the log-likelihood for the joint distribution of survival and longitudinal outcomes $\{T_i, \delta_i, y_i\}$ assuming that i.) vector of time-dependent random effects b_i underlies both the survival and longitudinal process, ii.) there is conditional independence based on the random effects, iii.) parameter estimates are consistent and asymptotically normal.

As with all likelihood estimation joint models estimates are also obtained by maximising the log-likelihood corresponding to the joint conditional distribution of the longitudinal and time-to-event outcome. In this setting, the matrix of the time-dependent random effects is shared between both the survival and longitudinal processes and thus these random effects adjust for association between the survival and longitudinal outcomes including the subsequent correlation between repeated measurements in the longitudinal process, which amounts to conditional

independence (Rizopoulos 2012).

5.2 Parameterisation of the joint model

In this Section we cover different parameterisations of the joint model that has been introduced in the previous Section. The parametrisations are for the association structure between the event and longitudinal process involving exogenous time dependent covariates. We limit our presentation of these parameterisations to those available in R, which is the main software used for analysis in this thesis.

The standard joint model we have introduced assumes that the risk for an event at time point t depends on the true value of the longitudinal bio-marker at the same time point, captured by the parameter α. This standard parameterisation is widely used due to the fact that, interpretation of α is straight forward and clear. The parameterisation assuming the true value of the bio-marker is commonly known as the current value structure (Rizopoulos 2012). It is not always practical to expect this parameterisation to suit all situations in medical research and other disciplines, in expressing the relationship between a longitudinal bio-marker and risk for an event. There are four other association structures that may be considered in linking the longitudinal and event processes, and this is without loss of other extensions.

The first parameterisation under consideration after the current value structure is the interaction effects parameterisation. This type of structure relaxes the strong assumption in the current value structure, that says, that the effect of the level of the bio-marker is the same in all the sub groups. The interaction effects structure includes in the linear predictor of the relative

risk model, interactions terms of the bio-marker with baseline covariates as follows:

$$h_i(t) = h_0(t)\exp\{\mathbf{X}'\boldsymbol{\beta}_1 + \alpha'[\boldsymbol{\beta}_2 \times m_i(t)]\}, \tag{5.5}$$

where $\boldsymbol{\beta}_1$ accommodates the direct effect of baseline covariates, to the risk for an event, and $\boldsymbol{\beta}_2$ contains the interaction terms that link the association of $m_i(t)$ in the different sub groups in the data (Rizopoulos 2012).

The second parameterisation is the lagged effects structure, which uses, time lagged covariates with the following formulation of the relative risk model:

$$h_i(t) = h_0(t)\exp\{\mathbf{X}'\boldsymbol{\beta} + \alpha m_i[t-c, 0]\}, \tag{5.6}$$

which claims that the risk at time t depends on the true value of the longitudinal bio-marker at time $t-c$, where c is the time lag of interest.

The third extension that could be of interest is the time dependent slopes parameterisation, which does not depend on the current or previous value of the longitudinal bio-marker. For example, it is intuitively known that each patient has a bio-marker that follows a unique trajectory in time, and therefore, it could be prudent to consider a structure that allows the risk for an event to also depend on other features of this trajectory (Rizopoulos 2012). In this parameterisation, the joint model uses both the current value and slope trajectories at time t, on which the risk for an event depends and has the following formulation of the relative risk model:

$$h_i(t) = h_0(t)\exp\{\mathbf{X}'\boldsymbol{\beta} + \alpha_1 m_i(t) + \alpha_2 m_i'(t)\}, \tag{5.7}$$

where $m_i'(t) = \dfrac{d}{dt}m_i(t)$.

The interpretation of α_1 is the same as previously stated in the standard joint model with current value parameterisation, where as, α_2 measures the strength of association between the

slope of the longitudinal trajectory at time t, with risk for an event at the same time point, while holding $m_i(t)$ constant. This type of parameterisation is particularly useful in instances where interest is in capturing differences in two patients that show similar true bio-marker levels at a specific time point, but may differ at the rate of change of the bio-marker (Rizopoulos 2012).

Lastly but not least, the other parameterisation worth considering would be the one that allows the whole history of the bio-marker to be associated with the hazard for an event. This structure includes in the linear predictor of the relative risk model the sum of the longitudinal trajectory, representing, the cumulative effect of the longidudinal response up to time point t as follows:

$$h_i(t) = h_0(t) \exp\{\mathbf{X}\boldsymbol{\beta} + \alpha \int_0^t m_i(s)ds\}, \tag{5.8}$$

where for any particular time point, α measures the strength of association between the risk for an event at time point t and the area under the longitudinal trajectory up to the same time point. In this case, the area under the longitudinal trajectory is regarded as a suitable summary of the whole trajectory.

Finally, an extension to the standard joint model that could be under consideration is the random effects structure. This structure includes, in the linear predictor of the relative risk model only random effects of the longitudinal sub-model as follows (Rizopoulos 2012):

$$h_i(t) = h_0(t) \exp\{\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\alpha} b_i\}, \tag{5.9}$$

where α represents a vector of association parameters, each one measuring the strength of association between the corresponding random effect and risk for an event. This structure is more useful when a simple random intercept and random slope is assumed for the longitudinal sub-model. In this setting random effects express subject specific departures from the average intercept and average slope (Rizopoulos 2012).

In this thesis we focus on the current value parameterisation of the association structure based

on the dataset (IMPI) that we used, which links square root CD4 cell count and risk of constrictive pericarditis. We consider this parameterisation based on two reasons, firstly it is known that the current level of CD4 cell count is directly related to the health status of an individual i.e. a measure of immunity (Post, Wood, and Maartens 1996; Rabeneck et al. 1993; Weinberg and Kovarik 2010), and secondly, the Cook's (Cook and Weisberg 1982) based diagnostics that we have developed apply in all paremeteristaions without making major alterations to the covariance matrix.

5.3 Case-deletion diagnostics for joint models for longitudinal and survival data

Introduction

Joint modeling for longitudinal and survival data is a technique for analysing data that has two main outcomes; these are the longitudinal outcome, often measured with error and the time-to-event outcome, which causes non-random drop out in the longitudinal outcome due to the occurrence of an event and this is not the focus of this thesis. In this thesis, the focus is on the strength of association between a time varying covariate and the risk of an event. This modeling strategy accounts for the effect of a time varying covariate measured with error as well as accounting for non-random drop out. The type of joint modeling explored in this case deletion diagnostics is the shared parameter model. In this approach the sub-model share a common random effect in the joint likelihood.

The estimation of the joint log likelihood is done using numerical integration methods since the log-likelihood for the joint distribution of the survival and longitudinal probability den-

sities through common shared random effects has no closed form. The maximisation of this log-likelihood is done using quasi Newton-Raphson Lange (2004) algorithm, which requires similar computations to the Expectation Maximization algorithm. This is a hybrid approach which proceeds with both the EM and later with direct maximisation.

5.3.1 Case-deletion diagnostics

Case-deletion diagnostics for the joint model using the Cook's distance approach is motivated by the fact that in both the longitudinal and survival outcomes there may be outliers or influential subjects or even observations that exert influence on model estimates. We will stick to the definition of an outlier and influential subject or observation as provided by Cook (1977). In its simplest expression, an outlier is an observation or subject that is not consistent with the rest of the data under study, where as, an influential observation or subject is one that has noticeable change in parameters when excluded in estimation.

A common approach used in model diagnostics in statistics is case-deletion and this assumes $\boldsymbol{\theta}$ to be a parameter vector of interest in the model under consideration and $\hat{\boldsymbol{\theta}}$ an estimate of $\boldsymbol{\theta}$ using appropriate techniques usually maximum likelihood in the frequentist paradigm Cook and Weisberg (1982). If one excludes the i^{th} observation from the full set of observations, $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ say; refitting the model under the same principle produces $\hat{\boldsymbol{\theta}}_{[-i]}$ of $\boldsymbol{\theta}$ which gives a measure of the difference between $\hat{\boldsymbol{\theta}}_{[-i]}$ and $\hat{\boldsymbol{\theta}}$, which is used to evaluate how the i^{th} observation impacts the estimate $\hat{\boldsymbol{\theta}}$ and subsequent inferences drawn from the model. This measure is called the Cook's distance due to Cook and Weisberg (1982) and has the following

formulation:

$$C_i(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}})' \mathbf{M}^{-1} (\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}}), \tag{5.10}$$

where $\mathbf{M}$ is the covariance matrix for all parameters $\boldsymbol{\theta}$ and it is positive definite (Cook and Weisberg 1982). Model (5.10) evaluates the Cook's distance C_i after excluding the i^{th} subject or observation from the estimation process, thereby, quantifying the impact of the i^{th} subject or observation on estimates of model parameters.

There are extensions of case-deletion diagnostics based on the Cook's statistics to mixed effects model and this has been an active research stream (Banerjee and Frees 1997; Demidenko and Stukel 2005; Lesaffre and Verbeke 1998; Pan and Fang 2002; Zewotir and Galpin 2005). Because of the iterative nature of estimation a simple analytical expression for the change in parameter estimates when a subject is deleted cannot be obtained. However, analytic expressions for the change in $\boldsymbol{\theta}$ if we use $\hat{\boldsymbol{\theta}}$ as a starting value and perform a one step iteration with the i^{th} subject omitted can be derived. Without loss of generality, such one step diagnostics adequately estimate the fully iterated change in the parameter with the i^{th} subject deleted, for models with non linear parameters, see (Pregibon 1981), who pioneered the idea.

We adopt the same approach and extend it in the joint model by decomposing the Cook's statistics and evaluate these diagnostics in simulation studies. We further illustrate this on a multi-center clinical trial on TB pericardiatis data and all this is done using the available JM package in R.

Cook's statistics for overall parameters are given as in equation (5.11)

$$C_i(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}})' \hat{\mathbf{M}}^{-1} (\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}}), \tag{5.11}$$

where $\hat{\mathbf{M}}$ is the covariance matrix for all parameters $\hat{\boldsymbol{\theta}}$ from both longitudinal and survival sub-models. For the sub-models the parameter set is decomposed as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1$ is

the parameter set from a longitudinal sub-model and θ_2 is the parameter set from a survival sub-model. Cook's statistics for all parameters in the longitudinal sub-model are given by:

$$C_i(\boldsymbol{\theta}_1) = (\hat{\boldsymbol{\theta}}_{1[-i]} - \hat{\boldsymbol{\theta}}_1)' \hat{\mathbf{M}}_1^{-1} (\hat{\boldsymbol{\theta}}_{1[-i]} - \hat{\boldsymbol{\theta}}_1), \tag{5.12}$$

where $\hat{\mathbf{M}}_1$ is the covariance matrix for all parameters in the longitudinal sub-model, which includes fixed effects and variance components. Similarly, for the survival sub-model, $\boldsymbol{\theta}_2$ is the parameter set and $\hat{\mathbf{M}}_2$ is the covariance matrix for parameters. In both sub-model C_i has the same interpretation as in model (5.10).

Analogous to the linear regression leverage matrix, Demidenko (2013) defined the $n_i \times n_i$ leverage matrix for the linear mixed model for the i^{th} subject as follows;

$$\mathbf{H}_i = \frac{\partial \hat{\mathbf{Y}}_i}{\partial \mathbf{Y}_i},$$

where $\hat{\mathbf{Y}} = \mathbf{X}_i \hat{\boldsymbol{\beta}} + \boldsymbol{Z}_i \hat{\mathbf{b}_i}$ which is the predicted outcome conditional on the estimated random effects, $\hat{\mathbf{b}}_i$. If we define $\mathbf{V}$ as the covariance matrix of the vector of observations, commonly, known as the generalised variance $\mathrm{Var}(\mathbf{Y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \boldsymbol{\Sigma}_i$; $\boldsymbol{\Sigma}_i$ is the covariance matrix of residuals for each subject and assuming that $\mathbf{V}_i$ is fixed and after differentiating; the leverage matrix can be presented as the sum of three components as follows:

$$\mathbf{H}_i = \mathbf{H}_{i1} + \mathbf{H}_{i2} + \mathbf{H}_{i3}, \tag{5.13}$$

where

$$\begin{cases} \mathbf{H}_{i1} = \mathbf{X}_i \mathbf{M}^{-1} \mathbf{X}_i' \mathbf{V}_i^{-1}, \\ \mathbf{H}_{i2} = \mathbf{Z}_i \mathbf{D} \mathbf{Z}'_i \mathbf{V}_i^{-1} (I - \mathbf{H}_{i1}) \qquad \text{and} \\ \mathbf{H}_{i3} = 2\mathbf{X}_i \mathbf{Z}_i \mathbf{D} \mathbf{Z}'_i \mathbf{V}_i^{-1} (I - \mathbf{H}_{i1}) \end{cases} \tag{5.14}$$

In this framework $\mathbf{H}_{i1}$ is the leverage for the fixed effects, $\mathbf{H}_{i2}$ is the leverage associated with the random effects and $\mathbf{H}_{i3}$ is the random effects variance, which is a measure of covariation between the a change in the average profile and a change in position of subject-specific pro-

files relative to the average profiles and this converges to zero.

We extend this by adopting a decomposition of the Cook's distance for the joint model by Zhang et al. (2014) using model (5.2). In this thesis we adopt a quasi-decomposition of Cook's statistics, which means that, the decomposition of the Cook's statistics similar to what, Zhang et al. (2014) used, but not analytically the same. We use a quasi-decomposition because the shared parameter joint model is estimated iteratively using optimisation algorithms, such as, Expectation Maximisation (EM) or Newton Raphson algorithm. This, renders, formulation of analytical expressions for quantities of the Cook's statistics difficult, however, software that fits shared parameter joint models enables the calculation of Cook's distance through the covariance matrix.

As a matter of emphasis, estimation of the joint model is done using optimisation algorithms, such as, Newton-Raphson (Lange 2004) and the EM algorithm. The iterative nature of estimating the joint model makes it difficult to formulate analytical expressions for a Cook's distance as is the case in linear regression, where, a residual, mean square error and other quantities are calculated. The available software that fit the joint model are able to calculate Cook's statistics based on the covariance matrix. We, thus formulate empirical Cooks statistics for the joint model which are our influence diagnostics. The Cook's statistics evaluated in the shared parameter joint model can be composite or subsets of the parameter space in the joint model.

Since the longitudinal and survival components of the joint model are conditionally indepen-

dent of random effects we can express the joint expectation as;

$$\begin{aligned} p(\mathbf{Y}_i, \mathbf{T}_i, \delta_i) &= \exp\Big[p(\mathbf{T}_i|\mathbf{b}_i) \times p(\mathbf{Y}_i|\mathbf{b}_i)\Big] \\ &= \exp\Big[p(\mathbf{T}_i)|\mathbf{b}_i \times p(\mathbf{Y}_i)|\mathbf{b}_i\Big] \\ &= \exp\Big[p(\mathbf{T}_i)|\mathbf{b}_i\Big] \times \exp\Big[p(\mathbf{Y}_i)|\mathbf{b}_i\Big] \end{aligned} \tag{5.15}$$

Using monte carlo methods after condition on random effects, the expectation is updated after calculating parameters until it converges (Zewotir and Galpin 2005), this process therefore enables formulation of an analogous of the hat matrix, $\mathbf{H}$, which is used in calculation Cook's distance for the i^{th} subject or observation.

The implementation of these case-deletion diagnostics has been based on the following empirical Cook's statistic:

$$C(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}})' \mathbf{M}^{-1} (\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}}), \tag{5.16}$$

where $\mathbf{M}$ is the covariance matrix for $\boldsymbol{\theta}$ from the estimated joint likelihood between the survival and the longitudinal sub-model under the joint model framework. This is derived after iteratively deleting the i^{th} subject from the longitudinal sub-model and re-estimating the joint likelihood each time.

The variance covariance matrix of all parameters, which is estimated using the joint likelihood between the two sub-models, is used to identify influential subjects based on the Cook's distance approach.

5.4 Simulation studies

Introduction

We conducted simulation studies for the joint model dataset, using an algorithm provided

for in the JM package in R by Rizopoulos (2017). The main objective of conducting these studies was to evaluate the performance of the developed case-deletion diagnostics. This simulation starts by setting a number (n) for subjects to be included in the study, number of planned repeated measurements per subject per outcome and the maximum time the follow up will be done. This is followed by setting up parameters for both the longitudinal sub-model and survival sub-model. In the longitudinal sub-model the parameters that are set up are coefficients (β) and measurement error variance (σ^2). The parameters to be set up in the survival sub-model, on the other hand, are the intercept as well as coefficients for the baseline covariates γ, the association parameter α, the shape parameter for the Weibull baseline hazard and means for the exponential distribution for the censoring mechanism.

The assumed longitudinal sub-model from which data is simulated is;

$$Y_i(t) = \beta_0 + (\text{Group})\beta_1 + (\text{Time})\beta_2 + \epsilon; \tag{5.17}$$

which is a random intercept model. To investigate the impact of subjects or observations that are influential on estimates of model parameters, we vary the variance components for the model using different scenarios as tabulated in Table 5.1 below. This allows us to calculate Cook's statistics directly based on case-deletion techniques.

The following survival sub-model is used to simulate relative risk data

$$h_i(t) = h_0(t)\exp\{(\text{Group}) + \alpha \text{m}_i(t)\}, \tag{5.18}$$

where α is the association parameter that links the longitudinal and survival processes through common rand effects and m_i is a representation of the repeated measurements.

To investigate subjects or observation that exert influence on the estimation of parameters in the model; survival times are sorted in ascending order. The association parameter is varied

as in Table 5.1 below, to investigate the effective of association and simultaneously insert an influential subject in the survival sub-model. The simulations based on the above, then, start by simulating random effects, the longitudinal responses, survival times and finally censoring times from the exponential distribution. In addition, calculation of event times is done i.e. $\min(T_i < C_i)$, after which an event indicator is also calculated, $\delta_i = I(T_i^* \leq C_i)$, which takes the value one when an event occurs and zero otherwise. This exercise is followed by some housekeeping, which include but not limited to, keeping the non-missing cases, i.e., drop the longitudinal measurements that were taken after the observed event time for each subject and deleting all unused objects. This process generates two datasets one for the longitudinal sub-model and the other for the survival sub-model.

5.4.1 Simulated dataset

In the longitudinal dataset, random effects and measurement error are generated from a normal distributions from model (5.17), which assumes that $b \sim N(0, \sigma_a^2)$, and $\epsilon \sim N(0, \sigma_e^2)$, respectively. In the event-time dataset we assume the failure time T_{ij} follows a Weibull distribution and the censoring times assumed to be non-informative following an exponential distribution.

5.4.2 Outliers from the Weibull distribution

The Weibull distribution is widely used in survival analysis in many disciplines in science, including medical research, to a large extent, because of its ease of application and versatility.

The general formulation of the Weibull probability density is

$$f(T) = \begin{cases} \frac{\phi}{\eta}\left(\frac{t-\gamma}{\eta}\right)^{\phi-1} e^{-\left(\frac{t-\gamma}{\eta}\right)^{\phi}}, & t \geq 0 \text{ or } \gamma, \eta > 0, \phi > 0, -\infty < \gamma < \infty \\ 0, & \text{otherwise}, \end{cases}$$

where ϕ is the shape parameter also known as the Weibull slope, η is the scale parameter and γ is the location parameter. In most survival analysis the location parameter is seldom used and, thus, it is set to zero, thereby, reducing the Weibull probability density to a two parameter density (ϕ, η). The density can reduce further to a one parameter density of η which is known before hand.

The values of the shape and scale parameter play an important role in the distributional characteristics of the Weibull density with the shape parameter having marked effects on the behaviour of the distribution. If the shape parameter takes some values, the density converges to some known and widely used distributions, e.g. if $\phi = 1$ the Weibull density reduces to an exponential probability distribution with two parameters where $\lambda = \frac{1}{\eta}$ the failure rate.

As can be seen from Figure 5.1 the distribution can take different shapes depending on the value and sign of the shape parameter. The same information can be looked at from the probability plot which gives a clear understanding as to why the Weibull shape parameter has an effect on the density.

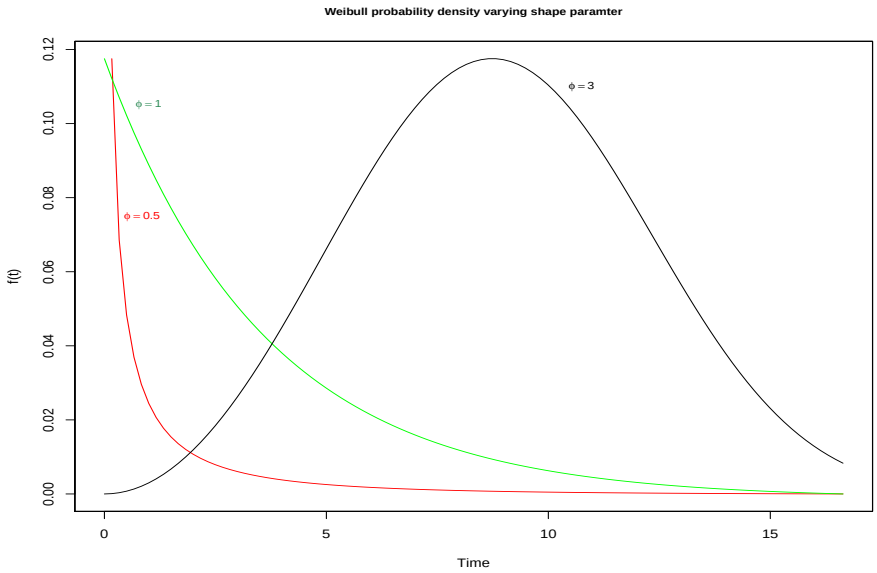


Figure 5.1: Weibull probability density function with varying shape parameter.

In survival data analysis in medical research a characteristic of the Weibull distribution, where the shape parameter has an obvious effect, is on the failure rate. This is important because it defines the rate at which events occur.

The cumulative density function of the Weibull distribution is given by

$$F(t) = 1 - e^{-\left(\frac{t-\gamma}{\eta}\right)^{\phi}},$$

which is known as the unreliability function and the reliability function is one minus the cumulative distribution function and, therefore, the reliability for the Weibull is

$$R(t) = e^{-\left(\frac{t-\gamma}{\eta}\right)^{\phi}}$$

The failure rate function is, therefore, the quotient of the probability density function and the reliability function

$$\lambda(t) = \frac{f(t)}{R(t)} = \frac{\phi}{\eta}\left(\frac{t-\gamma}{\eta}\right)^{\phi}.$$

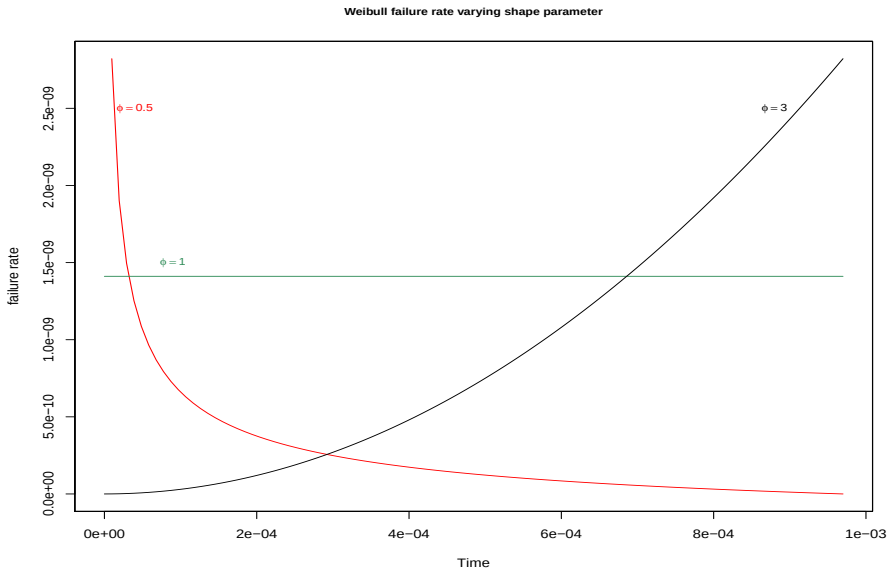


Figure 5.2: Weibull failure rate function with varying shape parameter.

In Figure 5.2, it can be seen that if the shape parameter is less that one ($\phi < 1$) the failure rate decreases monotonically with time, which is also known as early life failure, if, on the other hand, the shape parameter is set to one, there is a fairly constant failure rate also known as useful life or random failures. It is seen from the same figure, that when the shape parameter is greater than one, then the failure rate increases monotonically with time, a situation also known as wear-out failure. Setting the shape parameter as a mixture of the illustrated three distinct subpopulation produces the classic Weibull "bathtub curve." The curve has all the three settings when the shape parameter is less than one, equal to one and greater than one, which is what is seen in accelerated failure time models. All three life stages of the "bathtub curve" can be modeled with the Weibull distribution and varying values of ϕ.

Using the "bathtub curve" from the mixture of the three settings above for the shape parameter, it is expected that outliers or influential data points will be those that belong to either early life ($\phi < 1$) or worn-out life ($\phi > 1$).

The scale parameter in the Weibull probability density function has the same effect as changing the abscissa scale once it's values are changed. When the value of the scale parameter η is increased while holding the shape parameter constant, it stretches out the density. Using the axiom of probability that the area under a probability density function is constant and integrates to one; it, therefore, follows that, "the peak" of the curve for probability density function decreases with an increase in η. In summary, once the scale parameter is increased while holding the shape and location parameter constant, the distribution gets stretched out to the right and its height decreases while maintaining the shape and location. Conversely, decreasing the scale parameter has the effect of pushing the distribution towards the left (zero) and its height increases.

For simplicity, we arbitrarily set the following parameters in Table 5.1 below to simulate the joint model dataset. The simulated datasets with different parameter specifications are used to evaluate case-deletion diagnostics for the joint model. The case-deletion diagnostics are based on the Cook's statistics.

Table 5.1: Parameter value for the simulated joint model dataset.

	Scenarios values			
Parameter	One	Two	Three	Four
Global				
Number of subjects	50	100	50	100
Number planned repeated measurements	5	5	10	10
Maximum follow up time	12	12	18	18
Longitudinal sub-model				
Group	3.40	3.40	3.40	3.40
Variance components	0.90	0.90	0.41	0.41
Survival sub-model				
Intercept	-5.73	-5.73	-5.73	-5.73
Group	0.41	0.41	0.41	0.41
Association parameter	0.80	0.80	0.90	0.90

The four datasets generated by the parameters in the Table 5.1 are, firstly, used to insert outliers in the linear mixed model by varying the variance parameter. This was done by changing the measurement error (σ^2) in multiples of ten for a random selection of ten percent of the sample size in each scenario. The maximum of the simulated value is added to each of the ten percent candidate data points to create influential subjects.

Simulating outliers in time needs deciding the distribution of $T|\mathbf{X};\theta$, in this case a Weibull distribution is assumed. An outlier in time is, therefore, defined in two ways, the first being subjects that have large, herein after called "Too large", survival probabilities than is expected and, secondly, being those that have short, herein after called "Too early", survival probabilities than is expected. This is formally presented as

$$T_{ij} = \begin{cases} \text{"Too large"} & \text{if} \quad \delta_{ij} = 1, \\ \text{Weibull} & \text{if} \quad \delta_{ij} = 0 \end{cases} \tag{5.19}$$

For a given indicator δ denoting whether survival probabilities are "too large" or not. A similar set up can be done for survival probabilities that are "too early." A function of survival times is then defined with an indicator as follows:

$$f_T(t) = I_{\Delta=1}\Pi + I_{\Delta=0}\Pi$$

Simulating with 'cointoss' given $D = 1$ at critical points with respect to π, it then follows that $D \sim Bin(1, \pi)$, which means that if $D = 1$, then $T =$ "Too large" , for instance, time greater than 99^{th} percentile. This, therefore, implies that T should not be changed and, subsequently, if $D = 0$, then continue with standard simulations. Using the same approach for $E = 1$ at critical points with respect to ξ implies that $E \sim Bin(1, \xi)$, which means that if $E = 1$ then $T =$ "Too early". T is thus, distributed as Weibull, with mean $1/10$ of the "normal" and uniform on the short interval.

From the above simulation scheme, a ten percent random sample is picked and identified, whose survival times are contaminated as either being “too large” or “too early.”

5.4.3 Case-deletion for simulated longitudinal model

The simulated dataset from the previous Section has two components, thus a longitudinal as well as a survival component. The distribution of the longitudinal dataset for scenario one is as in Figure 5.3, the rest of the scenarios had similar distributions from those commonly seen in longitudinal studies, especially, clinical trials. In the Figure 5.3, Y is the longitudinal outcome and Time is the time at which longitudinal measurements are taken.

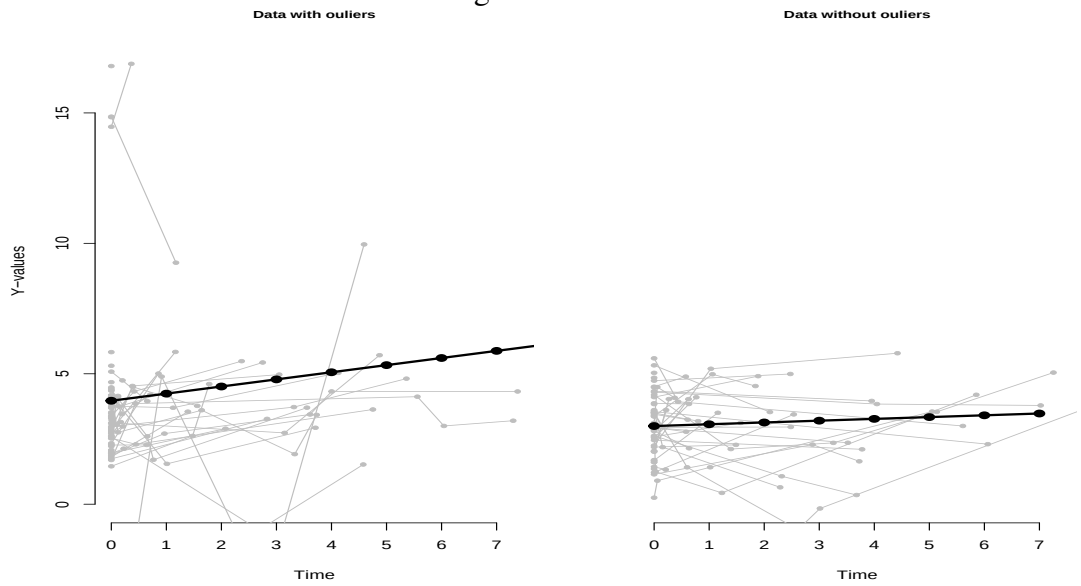


Figure 5.3: Scatter plot of simulated data point with mean profile.

We implemented case-deletion diagnostics to the simulated dataset using the linear mixed model alone as a starting point for the four sets of the variance that were inserted. In this implementation, we used the Cook’s distance statistic to evaluate and identify influential subjects

as in model (5.20).

$$C_i^*(\theta) = (\hat{\theta}_{[-i]} - \hat{\theta})' \hat{\mathbf{M}}^{-1} (\hat{\theta}_{[-i]} - \hat{\theta}), \tag{5.20}$$

In this model all parameters are held in θ and $\mathbf{M}$ is the variance covariance matrix.

In the following Section a simulation schema has been proposed that will generate four scenarios on which the diagnostics will be evaluated. The scenarios are intended to mimic the clinical trial setting, which is the primary focus of this thesis.

5.4.4 Case-deletion for simulated joint model for longitudinal and survival data

The first scenario, as in Table 5.1 showed that there were no differences in the two groups with respect to survival probabilities. The log-rank test statistics comparing the two groups had a p-value of 0.86.

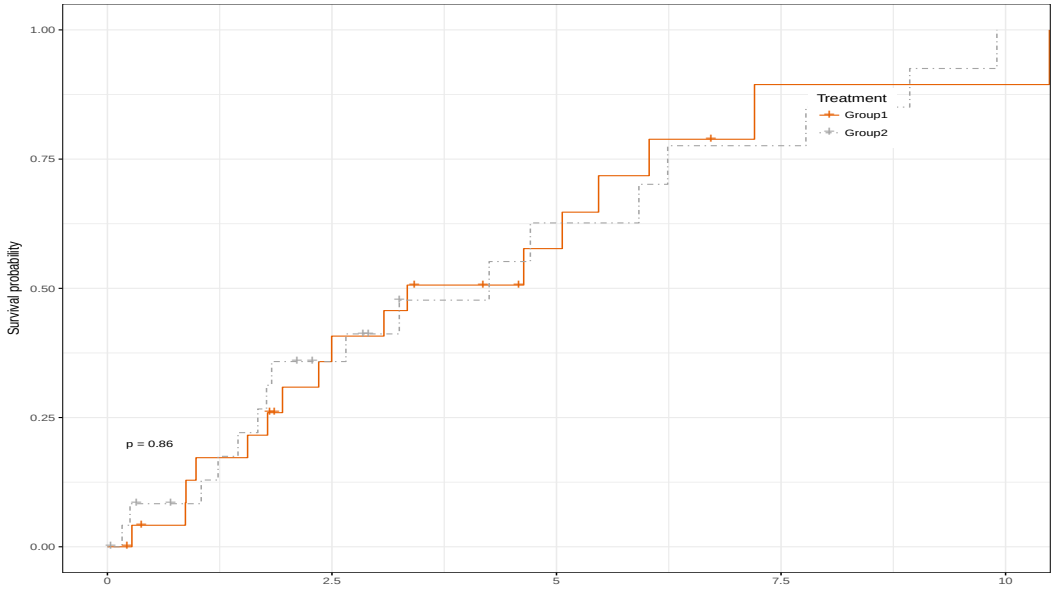


Figure 5.4: Survival probabilities of simulated data point.

The simulated event time dataset will be jointly modeled with the longitudinal dataset to evaluate the proposed diagnostics which are based on the Cook's statistics.

5.4.5 Scenario One

In this scenario, a total of 50 subjects were simulated, with five pseudo repeated measurements planned and followed up for a maximum of 12 time units, mimicking a clinical trial. In this simulation ten percent of the last simulated data points have been contaminated so that they are influential. The longitudinal and event-time datasets that were simulated are modeled jointly, conditional on random effects as shared parameters.

There are 88 % (44) events from the simulated event-time data, with the median survival time of 3.26 time units, 95 % Confidence Limits (2.77, 5.08). After jointly modeling the two processes, the group and time effects were significant in the two sub-models and the association parameter, which is a measure of strength linking the longitudinal and survival processes was also statistically significant (p-value < 0.001).

We went ahead to implement case-deletion diagnostics to the joint model based on the simulated dataset. The case-deletion diagnostics was implemented on the full parameter space, which includes fixed effects and variance components from the longitudinal sub-model and fixed effects, variance components and association parameter from the survival sub-model. The case-deletion, as previously described in Chapter 2 Section 2.2, evaluates the contribution of the i^{th} subject in the full model after exclusion in the next parameter estimation based on Cook's distance (5.20) which is derived from the hat matrix.

The following subjects: 8, 46, 47, 48, 49 and 50, were identified as influential as can been seen from Figure 5.5 below, which is consistent with the values that are contaminated (10%).

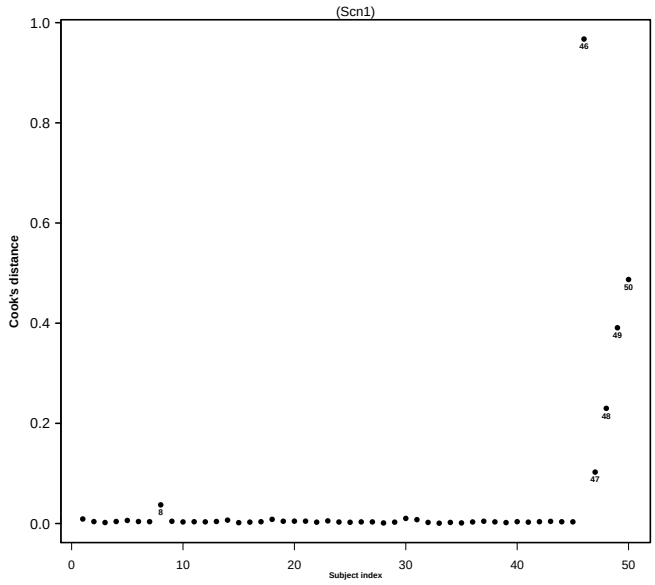


Figure 5.5: Cook's distance statistic for simulated data.

5.4.6 Scenario Two

In this scenario, a total of 100 subjects were simulated, with five pseudo repeated measurements planned and followed up for a maximum of 12 time units, mimicking a clinical trial. Just like in scenario one this simulation had ten percent of the last simulated data points contaminated so that they are influential. The longitudinal and event-time datasets that were simulated are again modeled jointly, conditional on random effects as shared parameters.

There were 80 % (80) events from the simulated event-time data, with the median survival time of 3.94 time units, 95 % Confidence Limits (3.36, 4.76). After jointly modeling the two processes ,the group and time effects were not statistically significant at 5% significance level in the two sub-models and the association parameter, which is a measure of strength linking the longitudinal and survival processes was also not statistically significant at the same level of significance.

We went ahead to implement case-deletion diagnostics to the joint model based on the simulated dataset. The case-deletion diagnostics was implemented on the full parameter space similar to that in scenario one and identified the following as influential subjects based on Cook's distance (5.20).

The following subjects: 91, 92, 93, 94, 95, 96, 97, 98, 99, and 100, were identified as influential as can been seen from Figure 5.6 below which is consistent with the values that were contaminated (10%).

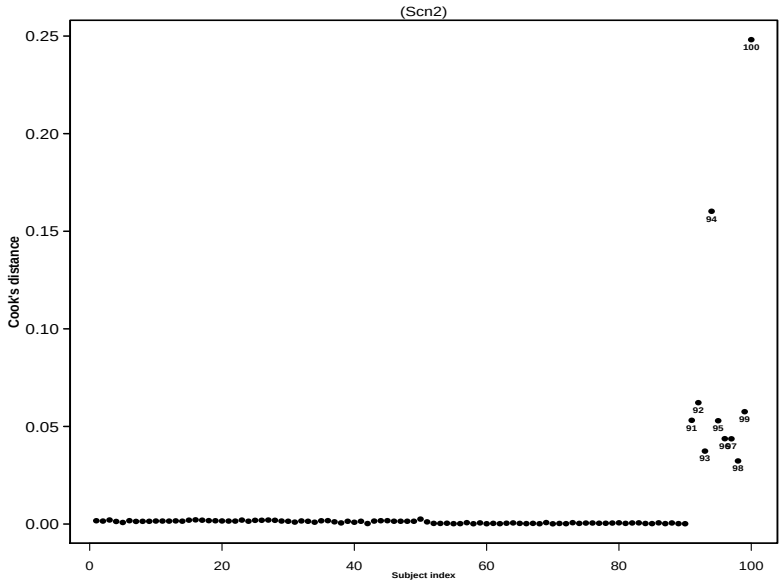


Figure 5.6: Cook's distance statistic for simulated data

5.4.7 Scenario Three

In this scenario, a total of 50 subjects were simulated, with ten pseudo repeated measurements planned and followed up for a maximum of 18 time units as a model for a clinical trial. Just like in scenarios one and two, this simulation had ten percent of the last simulated data points contaminated so that they are influential. The longitudinal and event-time datasets that were simulated were again modeled jointly, conditional on random effects as shared parameters just

like in the previous scenarios.

There were 76 % (38) events from the simulated event-time data, with the median survival time of 2.92 time units, 95 % Confidence Limits (2.74, 4.26). After jointly modeling the two processes, the group and time effects were statistically significant at 5% significance level in both of the sub-models, the association parameter, which is a measure of strength linking the longitudinal and survival processes, on the other hand, was not statistically significant (p-value = 0.11).

We went ahead to implement case-deletion diagnostics to the joint model based on the simulated dataset. The case-deletion diagnostics was implemented on the full parameter space similar to that in scenario one and two. The subjects identified by the case-deletion diagnostics are as follows; 46, 47, 48, 49 and 50 as shown by Figure 5.7 which is consistent with the values that are contaminated (10%).

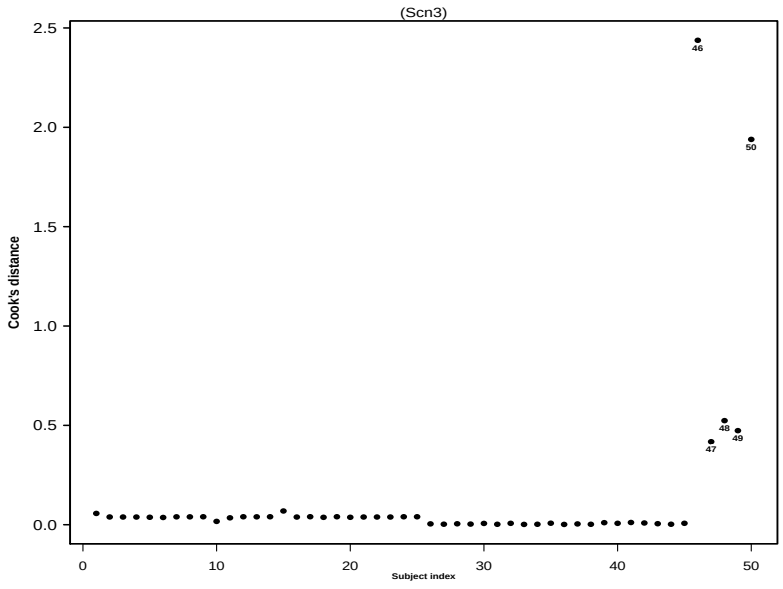


Figure 5.7: Cook's distance statistic for simulated data.

5.4.8 Scenario Four

In this scenario, a total of 100 subjects were simulated, with ten pseudo repeated measurements planned and followed up for a maximum of 18 time units as a model for a clinical trial. Just like in previous three scenarios, this simulation had ten percent of the last simulated data points contaminated so that they are influential. The longitudinal and event-time datasets that were simulated were again modeled jointly, conditional on random effects as shared parameters just like in the previous scenarios.

There were 88 % (88) events from the simulated event-time data, with the median survival time of 3.18 time units, 95 % Confidence Limits (2.77, 4.07). After jointly modeling the two processes, the group and time effects were statistically significant at 5% significance level in both of the sub-models, the association parameter, which is a measure of strength linking the longitudinal and survival processes on the other hand was not statistically significant (p-value = 0.14).

We went ahead to implement case-deletion diagnostics to the joint model based on the simulated dataset. The case-deletion diagnostics was implemented on the full parameter space similar to that in scenario one and two. The subjects identified by the case-deletion diagnostics are as follows; 42, 91, 92, 93, 94, 95, 96, 97, 98, 99, and 100, as shown by Figure 5.8 which is consistent with the values that are contaminated (10%).

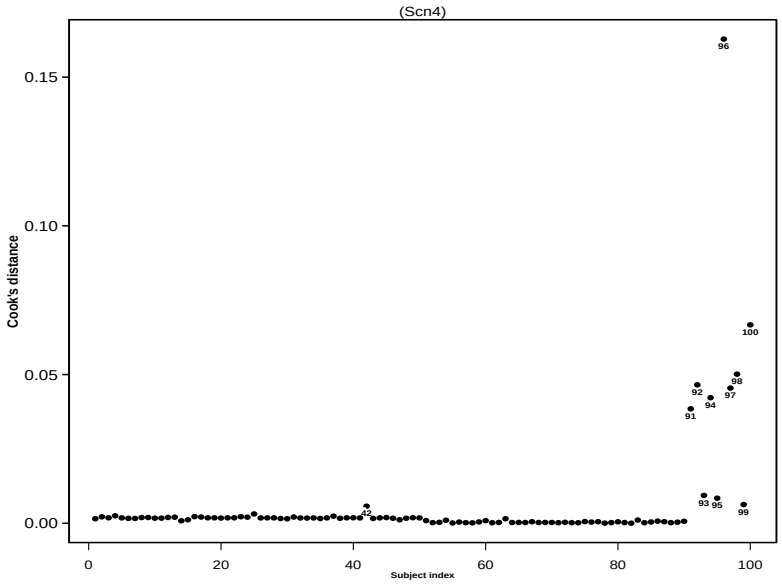


Figure 5.8: Cook's distance statistic for simulated data.

5.5 Summary

We have simulated joint model datasets for four different scenarios, each scenario had the parameters set based mainly on sample size, number of repeated measurements, length of follow up, measurement error in the longitudinal setting and association parameter in the survival setting. There has been a scheme to contaminate ten percent of the last data points in each scenario. This is what has been used as a bench mark to evaluate the case-deletion diagnostics based on Cook's statistics. In all the scenarios, the case-deletion diagnostics have been able to identify the ten percent contaminated data points. In scenarios one and four, besides identifying the ten percent contaminated data points the case-deletion diagnostics also identified subject 7 and 42, respectively.

The case-deletion diagnostics implemented in the IMPI dataset, therefore, perform just as expected even in simulated conditions.

5.6 Joint modeling CD4 count data and death using IMPI dataset

The linear mixed model that has been implemented is the one in Chapter 2.1 model (2.36). This model posits that square root CD4 cell count of all subjects in the study have the same trajectory in time but only differ at baseline, which means that each subject has his/her own intercept and hence the random intercept only model. We fitted the random intercept and random slope model to the data which was not better than the random slope only model comparing AICs as seen from the table 2.1.

We implemented Cox proportional hazards model with the following covariates; treatment, New York Heart Association (NYHA) classification, which basically categorises cardiac failure in an ordinal manner, Creatinine, peripheral oedema, which are categorical covariates and baseline weight (/10kg) and age are continuous covariates. The model is formulated as follows:

$$\begin{aligned} h_i(t) = h_0(t)\exp\{&\beta_1\text{pred} + \beta_2\text{nyclass} + \beta_3\text{creatn} \\ &+ \beta_4\text{weight} + \beta_5\text{oedema} + \beta_6\text{age} + \alpha Y_i(t)\}, \end{aligned} \tag{5.21}$$

where pred, is treatment which, is placebo or prednisolone, nyclass is the New York Heart Association (NYHA) classification, which basically categorises cardiac failure in an ordinal manner, creat is creatinin clearance, weight10 is the baseline weight (/10kg), oedema is peripheral oedema and age is the baseline age. The parameter that measures the strength of association between the longitudinal outcome and risk of death is denoted by α. This parameter has the interpretation that, for a unit increase in the longitudinal outcome the hazard of constriction is β. Model (5.21)has been chosen using the step-wise forward selection approach having set the threshold for covariate inclusion in the model at 10%.

Table 5.2: Estimates from joint modeling square root CD4 cell count and death.

Parameter	Estimate	Standard Error	p-value
Longitudinal submodel			
Intercept	14.19	1.10	<0.01
Treatment	-0.74	0.39	0.06
Time	0.63	0.03	<0.01
Survival submodel			
Prednisolone	0.51	0.37	0.17
NYHA class (III-IV)	0.52	0.36	0.15
Creatinine(>105)	1.37	0.40	<0.01
Peripheral Oedema (Yes)	0.31	0.37	0.41
Weight(/105)	-0.33	0.17	0.05
Age	-0.01	0.02	0.87
Association parameter	0.01	0.01	0.31

5.6.1 Identification of influential subjects

The case deletion technique that has been implemented using the IMPI data set is based on the Cook's distance approach. At the subject level, all parameters estimated from the joint likelihood have been used to evaluate the Cook's distance after excluding the i^{th} subject from the model as in model (5.10). Subjects that have been identified as influential on all parameters excluding the baseline risk function values within the intervals specified by the knots are: 140, 174, 277, 424 and 564 Figure 5.9. Those subjects that exert influence on the fixed effects using the longitudinal sub-model are: 253, 277, 379 and 564, while those that exert influence on the variance component in the longitudinal sub-model are: 140, 174, 277, 424, 557 and 564 Figure (5.10). In the survival sub-model, there seemed to be no influential subjects both on the fixed effects and the association parameter Figure 5.11. It is, thus seen, that subject

277 is influential in all the parameters using the Cook's distance approach, on the other hand, subjects: 174, 277, 424 and 564 are the most common influential subjects. Almost all of them were not receiving or taking ART and had had low CD4 cell count $(< 100\mu l)$ at end of the study by definition of this analysis.

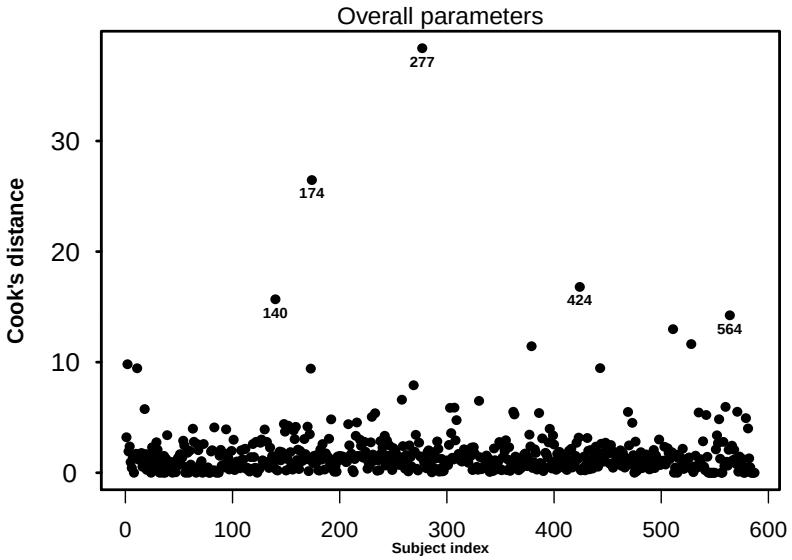


Figure 5.9: Cook's distance for all parameters in the joint model.

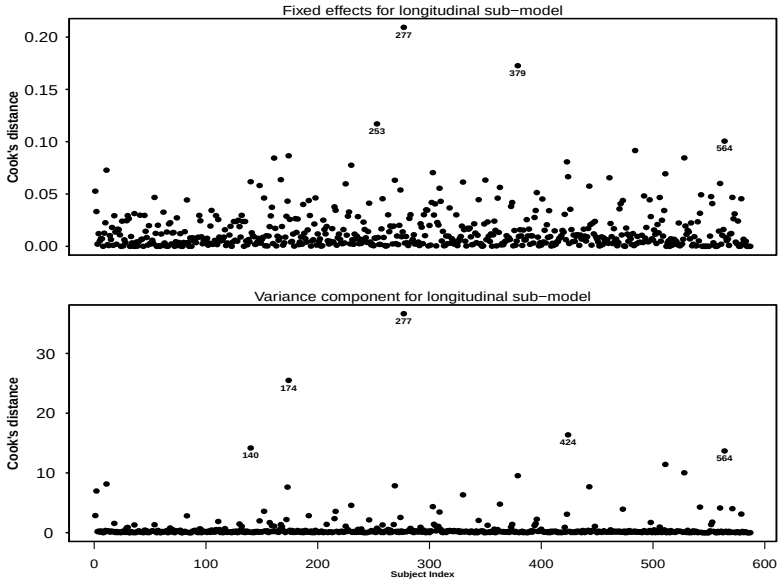


Figure 5.10: Cook's distance for fixed effects and variance components.

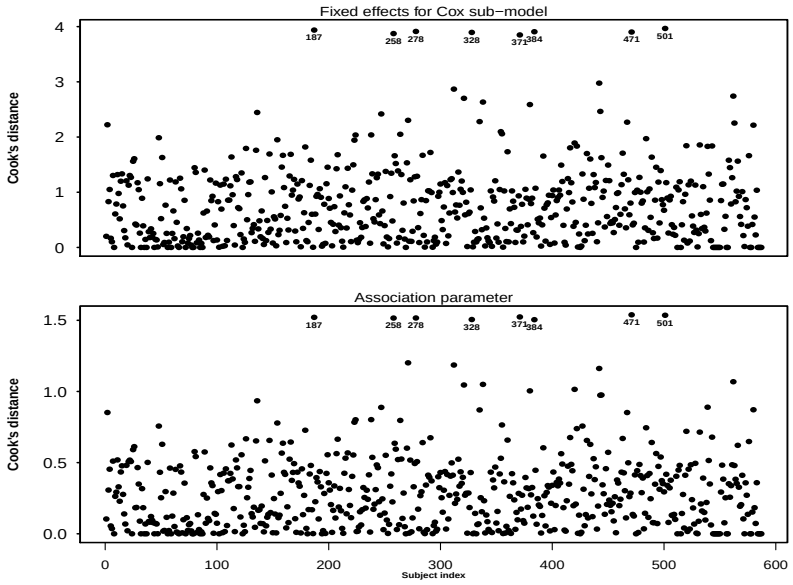


Figure 5.11: Cook's distance for fixed effects and association parameter.

The association parameter and fixed effects identify the following subjects as influential: 187, 258, 278, 328, 371, 384, 471 and 501. All of these have not been picked by the Two-stage joint model case deletion diagnostics (Chapter 4 above) as well as the case deletion diagnostics implemented in Section 5.6.1.

5.6.2 Identification of influential observations

At the observation level, the Cook's distance formulation that has been implemented is as follow:

$$C^*_{ij}(\theta) = (\hat{\theta}_{[-ij]} - \hat{\theta})' \hat{\mathbf{M}}^{-1} (\hat{\theta}_{[-ij]} - \hat{\theta}), \tag{5.22}$$

where $\hat{\theta}$ and $\mathbf{M}$ are as in equation (5.22). The ij^{th} notation denotes an observation taken at time point j for subject i. In this particular implementation it means a CD4 cell count collected at time point j for subject i.

The CD4 cell count observation 215_{-0} denotes that this is for subject number 215 taken at

baseline and $557_{-0.5}$ denotes that this is for subject number 557 with CD4 cell count taken at week two, the rest of the underscore whole numbers denotes month at which CD4 cell count was taken. The list of influential observations is as follows for all parameters: $215_{-0}, 246_{-0.5}$ and 379_{-0} Figure 5.12.

These observations have influence on the overall parameter space estimated by the joint model using the above Cook's distance formulation in equation (5.22).

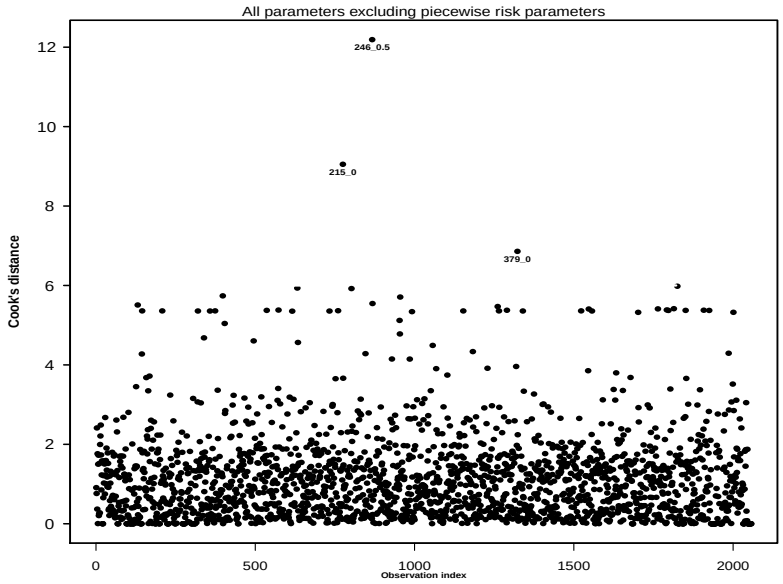


Figure 5.12: Cook's distance for all parameters in the joint model.

Those observations that exert influence on the fixed effects using the longitudinal sub-model are: $246_{-0.5}, 246_{-6}$ and 552_{-6}, while those that exert influence on the variance component in the longitudinal sub-model are: $215_{-0}, 246_{-0.5}, 557_{-0}$ and $557_{-0.5}$ (Figure 5.13). There seem to be little to no influence on the CD4 cell count observed on fixed effects of the survival sub-model and association parameter (Figure 5.14).

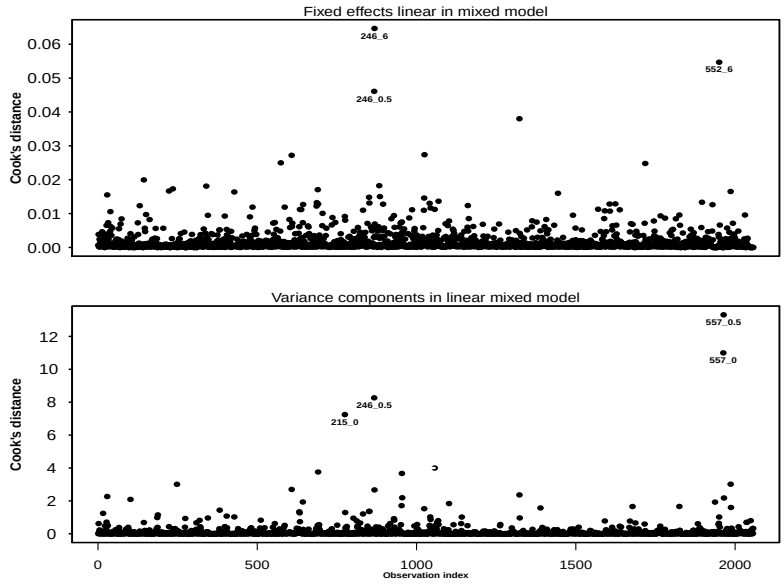


Figure 5.13: Cook's distance for Fixed effects and variance components.

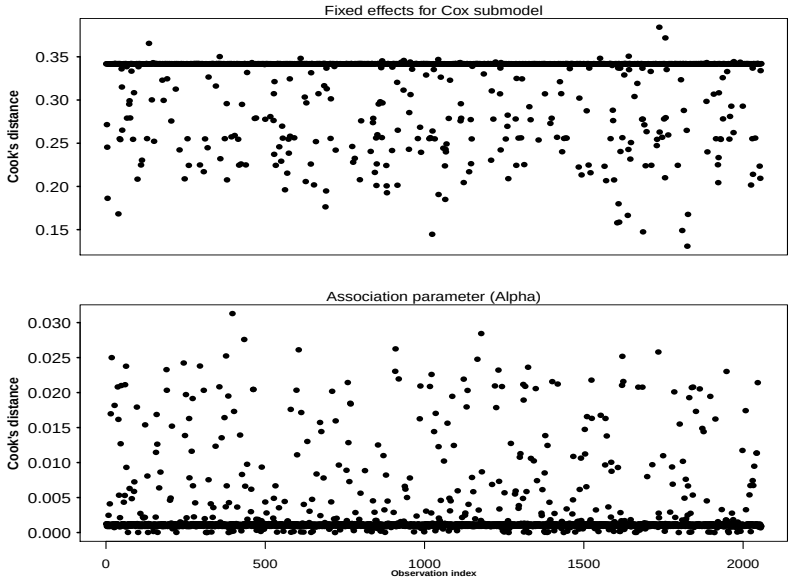


Figure 5.14: Cook's distance for fixed effects and association parameter.

5.7 Joint modeling CD4 count data and constriction

The linear mixed model is, as in Chapter 2.1 model (2.36) while the standard Cox proportional hazards model, that has been implemented posits, that the risk of constriction depends on

CD4 cell count, treatment arm, New York Heart Association (NYHA) classification which is as stated in Section 5.6, gender, peripheral oedema, haemoglobin status and TB status at baseline. The model is formulated as follows:

$$\begin{aligned} h_i(t) = h_0(t)\exp\{&\beta_1 \text{pred} + \beta_2 \text{nyclass} + \beta_3 \text{asex} \\ &+ \beta_4 \text{oedema} + \beta_5 \text{haemoglobin status} + \beta_5 \text{definitetbp} + \alpha Y_i(t)\}, \end{aligned} \tag{5.23}$$

where asex is gender (male or female), definitetb is the TB status i.e. whether one had probable or confirmed TB, haemoglobin status indicates whether the subject was anaemic or not at baseline and the rest of the variable are as defined in Section 5.6 above.

5.7.1 Identification of influential subjects

Using the same formulation as in equation (5.23) above, subjects identified as influential for constriction on the overall parameter space excluding the baseline risk function values within the intervals specified by the knots. These are 174, 277, 424, 511 and 557 as in Figure 5.15. Those subjects that exert influence on the fixed effects using the longitudinal sub-model are: 277, 303, 511 and 557, while those that exert influence on the variance component in the longitudinal sub-model are: 174, 277, 424, 511, and 557 Figure 5.16. In the survival sub-model, subjects that are influential on the fixed effects were quite a number of them, as follows; 9, 14, 17, 23, 64, 76, 184, 205, 217, 241, 266, 291, 315, 316, 320,468, 473, 474, 485, 486, 492, 506, 507, 562, 569, 571, and 578. As for the association parameter influential subjects are: 9, 76, 266, 316 and 578 (Figure 5.17).

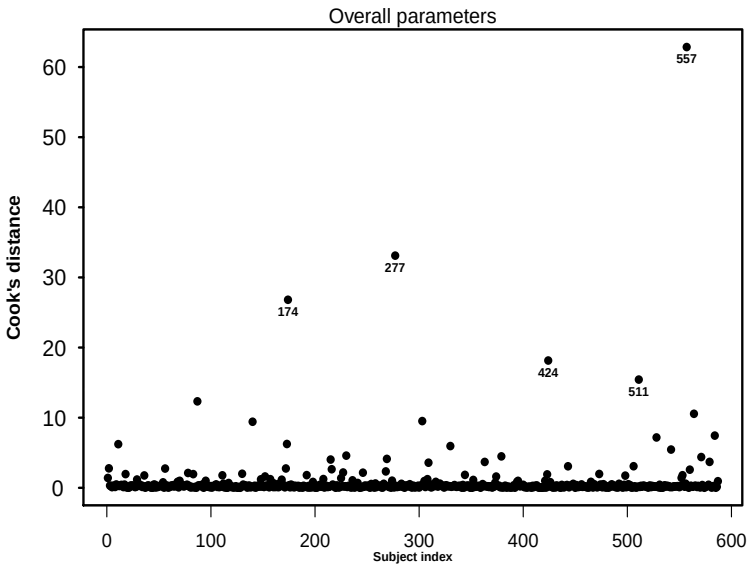


Figure 5.15: Cook's distance for all parameters.

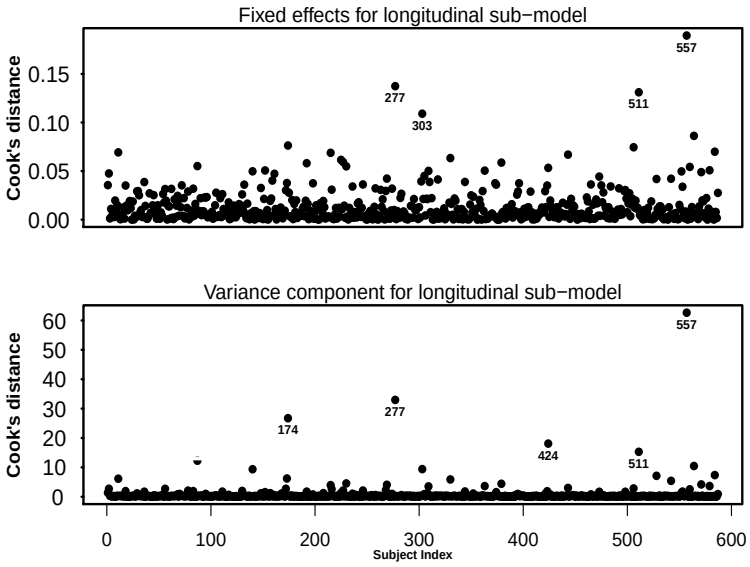


Figure 5.16: Cook's distance for fixed effects and variance component.

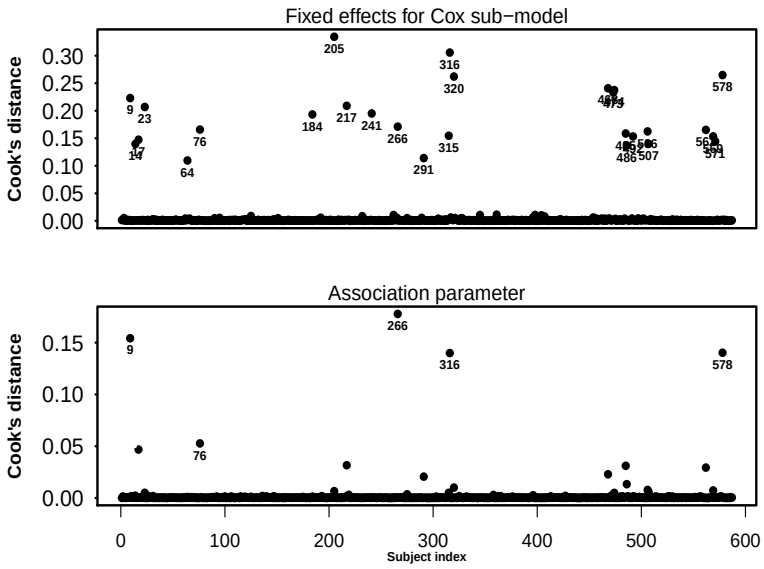


Figure 5.17: Cook's distance for fixed effects and association parameter.

5.7.2 Identification of influential observations

Using a similar approach as in Equation (5.22) above; followed by the notation for CD4 cell count observations in the preceding Section. Observation identified as influential on the overall parameter space for the full constriction model are:

$215_{-0}, 246_{-0.5}, 277_{-0.5}, 557_{-0}$ and $557_{-0.5}$ (Figure 5.18).

Those observations that are considered influential on the fixed effects in the longitudinal sub-model are: $246_{-0.5}, 246_{-6}$ and 552_{-6} while those that are influential on the variance component are: $215_{-0}, 246_{-0.5}, 277_{-0}, 557_{-0}$ and $557_{-0.5}$ (Figure 5.19).

On the other hand there is one observation considered influential on fixed effects in the survival sub-model 567_{-0}, while the association parameter does not seem to have influential observations (Figure 5.20).

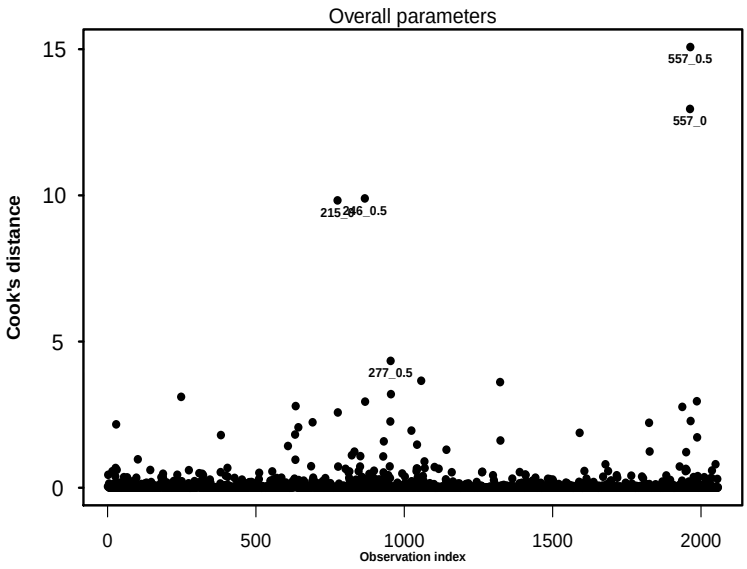


Figure 5.18: Cook's distance for all parameters.

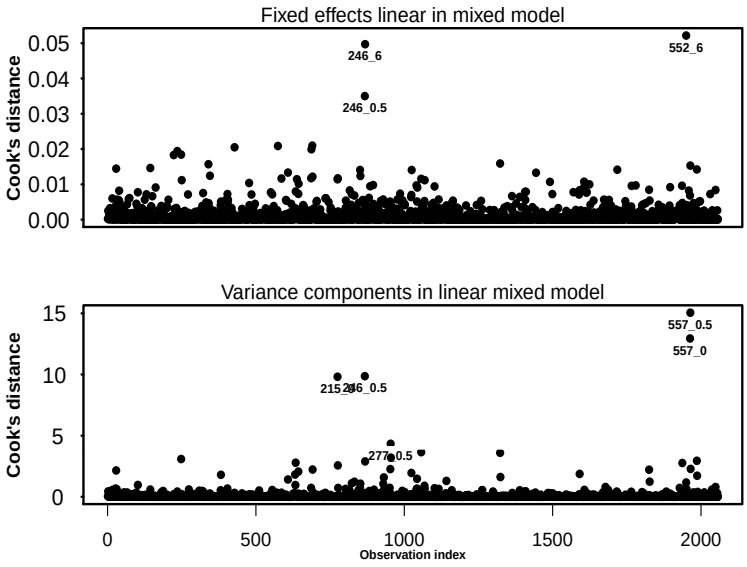


Figure 5.19: Cook's distance for fixed effects and variance components.

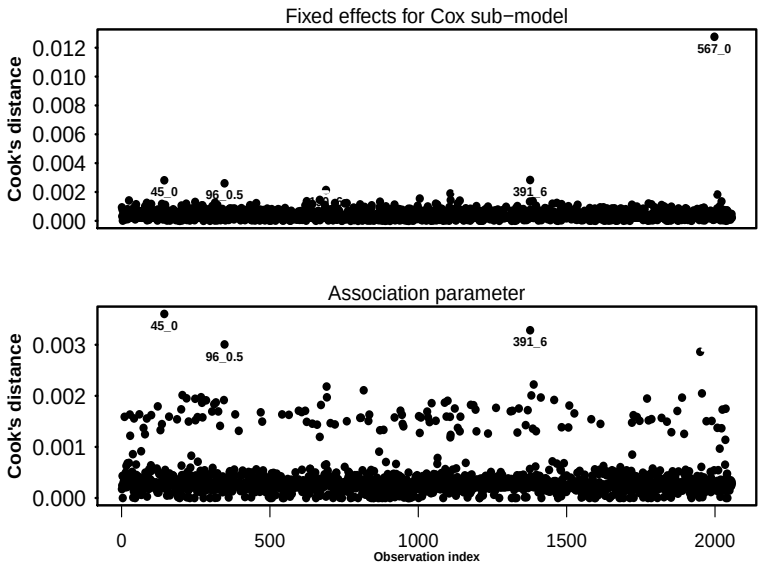


Figure 5.20: Cook's distance for fixed effects and association parameter.

Subjects that have been identified as influential are 7:8, 779, 733, 499, 468, and 396 which are common to almost all parameters. There was a 52% reduction in the estimate of the hazard ratio for the CD4 cell count and risk of death, after deleting the influential subjects.

Table 5.3: Joint model estimates for square root CD4 cell count and death before and after deleting influential subjects.

	M_7		M_8	
Parameter	Estimate (Std. Error)	p-value	Estimate (Std. Error)	p-value
Longitudinal sub-model				
Intercept	14.19 (1.10)	<0.001	14.12 (1.07)	<0.001
Prednisolone	-0.74 (0.39)	0.06	-0.63 (0.38)	0.10
Time	0.63 (0.03)	<0.001	0.64 (0.03)	<0.001
Survival sub-model				
Prednisolone	0.51 (0.37)	0.17	0.51 (0.37)	0.17
NYHA class (III-IV)	0.52 (0.36)	0.15	0.51 (0.36)	0.16
Creatinin(>105)	1.37 (0.40)	<0.001	1.38 (0.40)	<0.001
Weight(>105)	-0.33 (0.17)	0.05	-0.34 (0.17)	0.05
Peripheral Oedema	0.31 (0.37)	0.41	0.30 (0.37)	0.42
Age	-0.003 (0.020)	0.87	-0.003 (0.020)	0.89
Association parameter	0.006 (0.006)	0.31	0.006 (0.006)	0.31
-2 loglikelihood	11561.30	-	11398.62	-

M_7 is the model with full data while M_8 is the model with influential subjects deleted.

The joint model for CD4 cell count and constriction does not change parameter estimates after deleting influential subjects (Tables tables 5.3 and 5.4).The p-values for joint modeling CD4 cell count and death as well as constriction increased, mostly statistically insignificant, when influential subjects are deleted from the model. The estimate for the effect of treatment does not change, whether, there are influential subjects or not when jointly modeling CD4 cell count and constriction.

Table 5.4: Joint model estimates for square root CD4 cell count and constriction before and after deleting influential subjects.

	M_9		M_{10}	
Parameter	Estimate (Std. Error)	p-value	Estimate (Std. Error)	p-value
Longitudinal sub-model				
Intercept	14.46 (0.43)	<0.001	14.34 (0.41)	<0.001
Prednisolone	-0.75 (0.39)	0.06	-0.64 (0.38)	0.09
Time	0.63 (0.03)	<0.001	0.64 (0.03)	<0.001
Survival sub-model				
Prednisolone	-0.85 (0.42)	0.05	-0.84 (0.42)	0.05
NYHA class (III-IV)	0.35 (0.39)	0.37	0.35 (0.39)	0.35
Sex	0.61 (0.43)	0.16	0.61 (0.43)	0.16
Peripheral Oedema	0.50 (0.39)	0.20	0.51 (0.39)	0.20
TB Status (Yes)	0.81 (0.39)	0.04	0.81 (0.39)	0.04
Association parameter	0.009 (0.006)	0.16	0.008 (0.006)	0.17
-2 loglikelihood	12169.89	-	11956.40	-

M_9 is the model with full data while, M_{10} is the model with influential subjects deleted.

5.8 Description of influential subjects

The subjects that are identified as influential from the Cook's distance calculation have the following characteristics; Subject 586 had only two CD4 cell count measurements taken at baseline and one month, which had the same value of $134\mu\ell$. This subject was randomised to the prednisoline arm, was never on ART and died 49 days after randomisation. Subject 557 also had two CD4 cell measurements taken at baseline and one month, which had also

the same CD4 value at both visits of 164$\mu\ell$. This particular subject was randomised to the prednisolone arm, reported ART use at visit on third month and month six. The subject did not experience an event during the study period. Subject 511 had all the CD4 cell count measurements with a peak (1086 $\mu\ell$) at week two visit and a dip (589 $\mu\ell$) at the last visit. This subject was randomised to the placebo arm, was not on ART and was censored at the last visit. Subject 277 had two CD4 cell count measurements at baseline 1000 $\mu\ell$ and at two week visit 565 $\mu\ell$. The subject was randomised to the placebo arm, was not on ART, had confirmed and was censored at the last visit. Subject 246 had two CD4 measurements taken at week two (425 $\mu\ell$) and at the last visit (15 $\mu\ell$). This subject was randomised to the prednisolone arm and was censored at the last visit. Subject 174 had all CD4 cell counts taken, he was randomised to the placebo arm, had relatively high CD4 cell count up to month three and went down by month six and was censored after six months. Subject 303 had the first reading of CD4 cell count after three months into enrollment (654 $\mu\ell$) and the next CD4 cell count was at the six month visit (812 $\mu\ell$). The subject was 29.9 years old and randomised to the active arm, he was never on ART. He did not experience constriction and was censored at the end of study. Subject 424 had all CD4 cell count measurements taken (690, 757, 772, 780, 920 $\mu\ell$) for study visits, which were within normal range. The subject was aged 22.1 years, randomised to the placebo and had not experienced any event at the end of the study as in Figure A.5.

5.9 Summary

We had implemented influence diagnostics on the linear mixed model using the motivating dataset and identified influential subjects. The implementation had mainly focused on the

Cook's statistic applied to the entire vector space θ. We noted that the Cook's statistic was able to isolate about the same influential subjects in fixed effects as well as random effects.

The influence diagnostics implemented on the Cox model has its origins similar to the Cook's statistic i.e. DFBetas and likelihood displacement. Both of these were able to identify influential subjects using the motivating data set. Implementation of these diagnostics was been at subject level.

Standard joint model residuals diagnostics were implemented using the IMPI data set as shown in the above Section. Case deletion diagnostics were implemented on the joint model on the entire parameter space's variance covariance matrix, which includes: fixed effects, variance components and the association parameter. Using the Cook's distance approach, all the subject identified as influential in the linear mixed model have also been identified as influential in the joint model.

The next step was to use a similar approach at using the balanced survival data, further implementation would be using the variance shift outlier model within the joint model as suggested in Chapter 8 Section 8.6. All the above diagnostic tools based on Cook's statistics were implemented in simulation studies and extensions to multivariate joint model for longitudinal and survival data settings is presented in the following Chapter (6).

Chapter 6

Diagnostics for multivariate joint model for longitudinal and survival models

Introduction

In this Chapter, we introduced the multivariate linear mixed model starting with a setting where interest is on multiple Gaussian outcomes. This setting is, then, extended to environments where longitudinal outcomes are from a mix of exponential family of distributions using generalised linear mixed effects models (GLMM). We also wrote the associated likelihood functions for these types of models. We further motivated, why these models should be implemented when analysing data that falls in this category.

The main contribution of this Chapter is the development of case deletion diagnostics for the multivariate shared parameter joint model. We illustrated, the developed diagnostics using IMPI dataset (Mayosi et al. 2014).

6.1 Multivariate linear mixed model

We assumed that the number of participants $N, i = 1, 2, \ldots, N$ and that there are n_i repeated measures for each participant. We also assumed that there are K Gaussian outcomes, which could be extended to any exponential family using generalised linear model formulation for each participant $k = 1, 2, \ldots, K$. The responses for each participant i were collected in matrix

$\mathbf{Y}_i = [Y_{ik}]$. We denoted Y_{ik} as the measurement of the i^{th} participant, of the k^{th} outcome at some time point. We further assumed that $\mathbf{Y}_i = (y_{i1}, y_{i2}, \ldots, y_{ik})$ is the $n_i \times K$ response matrix for the i^{th} participant, where each of the $\mathbf{Y}_{ik}(k = 1, 2, \ldots, K)$ is the k^{th} response in the $n_i -$ vector of participant i. The multivariate longitudinal data can be modeled using the linear mixed model for the i^{th} participant as in model (4.4)of Section 4.1.1 rewritten as follows:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i,$$

where $\mathbf{X}_i$ is the $n_i \times p$ fixed between-subject design matrix, $\boldsymbol{\beta}$ is the matrix of fixed effects assumed to be common for all subjects, $\mathbf{Z}_i$ is the $n_i \times q$ random within-subject design matrix, b_i is the $q \times m$ matrix of random effects and, $\boldsymbol{\epsilon}_i$ is the $n_i \times m$ matrix of the measurement errors associated with the response matrix $\mathbf{Y}_i$

The distributional assumptions usually related to multivariate linear mixed models (MLMM) are as follows (Gad and EL-Zayat 2018; Schafer and Yucel 2002):

- $\tilde{\epsilon}_i = \text{vec}(\epsilon_i) \sim N_{n_ik}(\mathbf{0}_{n_ik}, \boldsymbol{\Sigma}_i) = \boldsymbol{\Sigma} \otimes \mathbf{I}_{n_i}$, the rows of the ϵ_i are normally distributed with mean zero and $k \times k$ covariance matrix $\boldsymbol{\Sigma}$. The vec is the vectorised operator that stacks all columns of the matrix ϵ_i vertically.
- The random effects $\tilde{\mathbf{b}}_i = \text{vec}(\mathbf{b}_i) \sim N_{qm}(\mathbf{0}_{qm}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is the $qm \times qm$ unstructured matrix and furthermore $b_i \quad \perp\!\!\!\perp \quad \epsilon_i$.
- The marginal distribution of $\mathbf{Y}_i$ has normal distribution with expectation $\text{E}(Y_i) - \mathbf{X}_i\beta$ dispersion matrix $\mathbf{V}_i$ of dimension $n_im \times n_im$;

$$\mathbf{V}_i = \Big(I_m \otimes \mathbf{Z}_i\Big)\boldsymbol{\Psi}\Big(L_m \otimes \mathbf{Z}_i\Big)' + \Big(\boldsymbol{\Sigma} \otimes \mathbf{I}_{n_i}\Big). \tag{6.1}$$

The responses $\mathbf{Y}_i$ can be written as a vector $\tilde{\mathbf{Y}}_i = \text{vec}(\mathbf{Y}_i)$, where $\tilde{\mathbf{y}}_i \sim N_{n_i m}(\mu_i = (I_m \otimes \mathbf{X}_i)\tilde{\boldsymbol{\beta}}, \mathbf{V}_i$ and can be modeled as:

$$\tilde{\mathbf{Y}}_i = \Big(I_m \otimes \mathbf{X}_i\Big)\tilde{\boldsymbol{\beta}} + \Big(I_m \otimes \mathbf{Z}_i\Big)\tilde{\mathbf{b}}_i + \epsilon_i, \tag{6.2}$$

where $i = 1, 2, \ldots, N$ and $\tilde{\boldsymbol{\beta}} = \text{vec}(\boldsymbol{\beta})$. The probability density function of $\tilde{\mathbf{Y}}_i$ is given by

$$\begin{aligned} f_{\tilde{\mathbf{Y}}_i}(\tilde{\mathbf{Y}}_i) &= (2\pi)^{-\frac{n_i m}{2}} |\mathbf{V}_i|^{-\frac{1}{2}} e^{-\frac{1}{2}([\tilde{Y}_i - \mu_i]' |\mathbf{V}_i^{-1}| [\tilde{Y}_i - \mu_i])} \\ &= (2\pi)^{-\frac{n_i m}{2}} |\mathbf{V}_i|^{-\frac{1}{2}} e^{\text{tr}\Big\{-\frac{1}{2}[\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}]\mathbf{V}_i^{-1}[\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}]'\Big\}}. \end{aligned} \tag{6.3}$$

6.1.1 The likelihood function for the observed data

Given that the k outcomes are fully observed, the likelihood function for the observed data for all parameters; $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$ can be obtained using maximum likelihood techniques as follows:

$$\begin{aligned} L_1(\theta) &= \prod_{i=1}^{n_i} f(\mathbf{Y}_i | \mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \times f(\mathbf{b}_i | \boldsymbol{\Psi}) \\ &= \prod_{i=1}^{n_i} f(\mathbf{Y}_i | \boldsymbol{\beta}, \boldsymbol{\Sigma}) \times f(\mathbf{b}_i | \boldsymbol{\Psi}) \\ &= \prod_{i=1}^{n_i} f(\mathbf{Y}_i | \boldsymbol{\theta}). \end{aligned} \tag{6.4}$$

The maximum likelihood estimate of θ can be obtained by maximising the likelihood function in equation (6.4) above, which is based on the marginal distribution of $\mathbf{Y}_i$ in the absence of random effects. If the main focus is in the variance structure introduced by $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$, random effects are considered nuisance parameters.

The log likelihood function based on equation (6.3) for the sample can be expressed as:

$$\begin{aligned} \ell(\theta) &= \log\left[\prod_{i=1}^{n_i} f_{\tilde{\mathbf{Y}}_i}(\tilde{\mathbf{Y}}_i | \theta)\right] = \sum_{i=1}^{N} \log f_{\tilde{\mathbf{Y}}_i}(\tilde{\mathbf{Y}}_i | \theta) \\ &= -\frac{mM}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N} \log|\mathbf{V}_i| - \frac{1}{2}\sum_{i=1}^{N} \mathbf{r}_i' \mathbf{V}_i^{-1} \mathbf{r}, \end{aligned} \tag{6.5}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{\Sigma})$ is considered the parameter matrix of interest, $M = \sum_{i=1}^{N} n_i$ and

$$\mathbf{r}_i = \text{vec}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) = \tilde{\mathbf{Y}}_i - \Big(I_m \otimes \mathbf{X}_i\Big)\tilde{\boldsymbol{\beta}}.$$

There are simpler alternatives to the above log-likelihood function that have been derived (Gad and EL-Zayat 2018; Schafer and Yucel 2002). In the next Section 6.2, we applied the theory shown in this Section and extended it to joint models for longitudinal and survival data using the generalised linear mixed model framework.

6.2 Multivariate Joint Models - Theory

6.2.1 Multivariate Joint Model Specification

Joint models for longitudinal and survival data have gained a lot of attention in recent years. These models have been extended to handle among others, multivariate longitudinal data, competing risks and recurrent events. In recent times, there also exist several freely available software packages for their implementation. From the aforementioned extensions, the one that is most practically relevant is in the multivariate longitudinal data setting. Even though this extension is mathematically straightforward, joint models with multiple longitudinal outcomes remain difficult to fit in practice due to the high number of random effects, which is computer intensive. This difficulty has also hampered to a degree, their practical application. In this thesis the estimation parameters has been done using the Bayesian techniques based on Dirichlet priors. The current version of JMbayes implements a novel approach that enables fitting such joint models in realistic computing times.

We start with a general definition of the framework of multivariate joint models for multiple longitudinal outcomes and an event time. Let $D_n = \{T_i, T_i^U, \delta_i, Y_i; i = 1, \ldots, n\}$ denote a

sample from the target population, where we let T_i^* denote the true event time for the i^{th} subject, T_i and T_i^U are times at which participants are observed, and $\delta_i \in 0, 1, 2, 3$ denotes the event indicator, with 0 corresponding to right censoring ($T_i^* > T_i$), 1 to a true event ($T_i^* = T_i$), 2 to left censoring ($T_i^* < T_i$), and 3 to interval censoring ($T_i < T_i^* < T_i^U$). Assuming $\mathbf{K}$ longitudinal outcomes we let y_{ik} denote the $n_{ik} \times 1$ longitudinal response vector for the k^{th} outcome ($k = 1, \ldots, K$) and the i^{th} subject, with elements Y_{ijk} denoting the value of the k^{th} longitudinal outcome taken at time point $t_{ijk}, j = 1, \ldots, n_{ik}$.

To accommodate multivariate longitudinal responses of different types in a unified framework, we postulate a generalized linear mixed effects model. In particular, the conditional distribution of Y_{ik} given a vector of random effects b_{ik} is assumed to be a member of the exponential family, with linear predictor given by:

$$g_k[\mathrm{E}\{Y_{ik}(t)|b_{ik}\}] = \eta_{ik}(t) = X_{ik}^T(t)\beta_k + Z_{ik}^T(t)b_{ik}, \tag{6.6}$$

where $g_k(.)$ denotes a known one-to-one monotonic link function, and $Y_{ik}(t)$ denotes the value of the k^{th} longitudinal outcome for the i^{th} subject at time point t, $X_{ik}(t)$ and $Z_{ik}(t)$ denote the design vectors for the fixed-effects β_k and for the random effects b_{ik}, respectively. The dimensionality and composition of these design vectors is allowed to differ between the multiple outcomes, and they may also contain a mix of baseline and time-varying covariates. To account for the association between the multiple longitudinal outcomes we link their corresponding random effects. More specifically, the complete vector of random effects $b_i = (b_{1i}, b_{2i} \ldots b_{iK})^T$ is assumed to follow a multivariate normal distribution with mean zero and variance-covariance matrix $\mathbf{D}$. For the survival process, we assume that the risk for an event depends on a function of the subject-specific linear predictor $\eta_i(t)$ and/or the random

effects. More specifically, we have

$$\begin{aligned} h_i(t|H_i(t), X_i(t)) &= \lim_{\Delta t \to 0} \Pr\{t \leq T_i^* < t + \Delta t | T_i^* \geq t,\, H_i(t), X_i(t)\}/\Delta t,\, t > 0 \\ &= h_0(t) \exp\left[\gamma^T X_i(t) + \sum_{k=1}^{K} \sum_{l=1}^{L_k} f_{lk}\{H_{ik}(t),\, X_i(t), b_{ik},\, \alpha_{lk}\}\right] \end{aligned} \tag{6.7}$$

where $H_{ik}(t) = \{\eta_{ik}(s),\, 0 \leq s < t\}$ denotes the history of the underlying longitudinal process up to t, $h_0(.)$ denotes the baseline hazard function, $X_i(t)$ is a vector of exogenous, possibly time-varying, covariates with corresponding regression coefficients γ. Functions $f_{lk}(.)$, parameterized by vector α_{lk}, specifies which components/features of each longitudinal outcome are included in the linear predictor of the relative risk model. Some examples motivated by the literature are (subscripts lk have been dropped in the following expressions but are assumed):

$$f\{H_i(t), X_i(t), b_i, \alpha\} = \alpha \eta_i(t) \tag{6.8a}$$

$$f\{H_i(t), X_i(t), b_i, \alpha\} = \alpha_1 \eta_i(t) + \alpha_2 \eta_i'(t), \text{ with } \eta_i'(t) = \frac{d\eta_i(t)}{dt} \tag{6.8b}$$

$$f\{H_i(t), X_i(t), b_i, \alpha\} = \alpha \int_0^t \eta_i(s) ds \tag{6.8c}$$

These formulations of $f(.)$ postulate that the hazard of an event at time t may be associated with the underlying level of the biomarker at the same time point, the slope of the longitudinal profile at t or the accumulated longitudinal process up to t. In addition, the specified terms from the longitudinal outcomes may also interact with some covariates in the $X_i(t)$. Furthermore, note that we allow a combination of L_k functional forms per longitudinal outcome. In this thesis we have not considered extensions or parameterisations for the following Equations (6.8b) and (6.8c). We have provided a description of these parameterisations in section 5.2 and potential research areas in section 8.6.

Finally, the baseline hazard function $h_0(.)$ is modeled flexibly using a B-splines approach, i.e.,

$$\log h_0(t) = \sum_{q=1}^{Q} \gamma_{h_0,q} B_q(t, v), \tag{6.9}$$

where $B_q(t, v)$ denotes the q^{th} basis function of a B-spline with knots $v_1, \ldots, v_Q$ and γ_{h_0} the vector of spline coefficients. To avoid the task of choosing the appropriate number and position of the knots, we include, a relatively, high number of knots (e.g., 15 to 20) and appropriately penalize the B-spline regression coefficients γ_{h_0} for smoothness using the differences in penalty. It has to be noted that using the B-splines approach can be computationally challenging, and, therefore a balance has to be struck between flexibility and computational time.

6.2.2 Likelihood and Priors

For the estimation of joint model's parameters we will use a Bayesian approach. The posterior distribution of the model parameters given the observed data is derived under the assumptions, that given the random effects, the longitudinal outcomes are independent from the event times, the multiple longitudinal outcomes are independent of each other, and the longitudinal responses of each subject in each outcome is independent. Under these assumptions the posterior distribution is analogous to:

$$p(\theta, b|Y_{ki}) \propto \prod_{i=1}^{n} \prod_{k=1}^{K} \prod_{j=1}^{n_{ki}} p(y_{kij}|b_{ki}, \theta) p(T_i, , T_i^U, \delta_i|b_{ki}, \theta) p(b_{ki}|\theta) p(\theta),$$

where θ denotes the full parameter vector, which included fixed effects from both sub-models, variance components from the GLMM and association parameters from the survival sub-model, further

$$p(y_{kij}|b_{ki}, \theta) = \exp\left\{ \Big[y_{kij}\psi_{kij}(b_{ki}) - c_k\{\psi_{kij}(b_{ki})\} \Big] \Big/ a_k(\varphi) - d_k(y_{kij}, \varphi) \right\}, \tag{6.10}$$

with $\psi_{kij}(b_{ki})$ and φ denoting the natural and dispersion parameters in the exponential family, $c_k(.), a_k(.)$ and $d_k(.)$ are known functions specifying the member of the exponential family. The survival part of the model has the following expression:

$$
\begin{aligned}
p(T_i, T_i^U, \delta_i | b_i, \theta) = & \Big\{ h_i(T_i | H_i(T_i), X_i(T_i)) \Big\}^{I(\delta_i=1)} \exp\left\{ - \int_0^{T_i} h_i(s | H_i(s), X_i(s)) ds \right\}^{I(\delta_i \in \{0,1\})} \\
& \times \left\{ 1 - \exp\left\{ - \int_0^{T_i} h_i(s | H_i(s), X_i(s)) ds \right\} \right\}^{I(\delta_i=2)} \\
& \times \left\{ \exp\left\{ - \int_0^{T_i} h_i(s | H_i(s), X_i(s)) ds \right\} - \exp\left\{ - \int_0^{T_i^U} h_i(s | H_i(s), X_i(s)) ds \right\} \right\}^{I(\delta_i=3)},
\end{aligned}
\tag{6.11}
$$

where $I(.)$ denotes the indicator function. The integral in the definition of the cumulative hazard function does not have a closed form solution, and thus a numerical method must be employed for its evaluation. Standard options are the Gauss-Kronrod and Gauss-Legendre quadrature rules.

For the parameters of the longitudinal outcomes we use standard default priors, more specifically, independent normal priors with zero mean and variance 1000 for the fixed effects and inverse Gamma priors for scale parameters. The covariance matrix of the random effects is parameterized in terms of a correlation matrix Ω and a vector of δ_d. For the correlation matrix Ω we use LKJ-Correlation prior as defined by Rizopoulos (2012) with parameter $\zeta = 1.5$. For each element of δ_d we use a half-Student's t prior with 3 degrees of freedom. For the regression coefficients γ of the relative risk model we assume independent normal priors with zero mean and variance 1000. The same prior is also assumed for the vector of association parameters α. However, when α becomes high dimensional (e.g., when several functional forms are considered per longitudinal outcome), we opted for a global-local ridge-type shrinkage prior,

i.e., for the s-th element of α we assume

$$\alpha_s \sim \mathcal{N}(0, \tau\psi_s), \tau^{-1} \sim \text{Gamma}(0.1, 0.1), \psi^{-1} \sim \text{Gamma}(1, 0.01). \tag{6.12}$$

The global smoothing parameter τ has, sufficiently, mass near zero to ensure shrinkage, while the local smoothing parameter ψ_s allows individual coefficients to attain large values. Other options of shrinkage or variable-selection priors could be used as well. Finally, the penalized version of the B-spline approximation to the baseline hazard is specified using the following hierarchical prior for γ_{h_0}:

$$p(\gamma_{h_0}|\tau_h) \propto \tau_h^{\rho(K)/2} \exp\left(-\frac{\tau_h}{2} \gamma_{h_0}^T K \gamma_{h_0} \right), \tag{6.13}$$

where τ_h is the smoothing parameter that takes a Gamma(1,$\tau_{h\delta}$) prior distribution, with a hyper-prior $\tau_{h\delta} \sim \text{Gamma}(10^{-3}, 10^{-3})$, which ensures a proper posterior distribution for γ_{h_0}, $K = \Delta_r^T \Delta_r + 10^{-6} I$, with Δ_r denoting the r^{th} difference penalty matrix, and $\rho(K)$ denotes the rank of K.

6.3 Application on IMPI dataset at subject level

The marginal distribution of CD4 cell count looks linear over the period of six months with an increasing trend. There is a similar trend observed with ART uptake, at baseline about 20% of participants reported to be on ART, which steadily increased to about 70% during the six

months' time period (Figure 6.1).

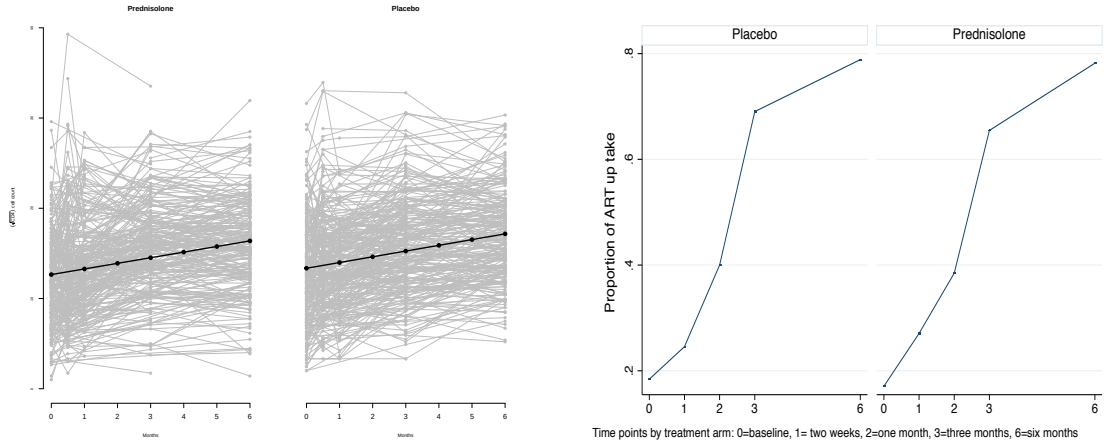


Figure 6.1: Distribution of CD4 and ART uptake over time.

The proportion of ART uptake by the participants included in this analysis looks linear in time as seen in Figure 6.1, it is, therefore, reasonable to model the data using a linear model.

In this analysis, interest is on individual CD4 cell counts profiles over time and subject specific response to ART uptake over the same time period. We therefore, postulated a generalised multivariate linear mixed effects model for the two outcome as formulated in equation (6.6). The main advantage of using multivariate linear mixed effects models is that, given the random effects, these models account for correlation between the set of longitudinal outcomes under study. In this particular case, it is assumed that given the random effects, the generalised linear mixed model (6.6) accounts for correlation between CD4 cell count and ART uptake by conditioning on random effects. In this analysis regression coefficients have subject-specific interpretation in terms of changes in the mean response for a specific subject, which is assumed to have come from a specific distribution usually normal with mean zero and some variance covariance matrix.

In the next section we implemented case deletion diagnostics on the multivariate joint model

(6.7) using the IMPI dataset.

6.3.1 Identification of influential subjects

We implemented case deletion diagnostics on the multivariate joint survival model based on Cook's statistics (Cook and Weisberg 1982). We evaluated Cook's statistics using model (5.10) of Section 5.3.1 as presented below in order to elucidate the different components of the multivariate joint model.

$$\mathbf{C}^{6.14}(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}})' \mathrm{M}^{-1} (\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}}), \tag{6.14}$$

where M is the variance covariance matrix for all parameters $\boldsymbol{\theta}$ derived from the Markov Chain Monte Carlo (MCMC) samples based on model (6.10)using JAGS in R package. We used the last parameters from the chain, after *burn-in*, under the assumption that, the model will have fully converged by the time the last parameters are estimated. The parameter set $\boldsymbol{\theta}$ is decomposed as follows: $\boldsymbol{\theta} = \boldsymbol{\theta}_{k1}, \boldsymbol{\theta}_2$ and $\mathrm{M} = (\mathrm{M}_{k1}, \mathrm{M}_2)$ for the multivariate GLMM and the survival sub-models, respectively. The case of $\boldsymbol{\theta}_{k1}$: $k = 2$ is the number of longitudinal outcomes, which are CD4 cell count and ART uptake.

The Cook's statistics evaluated using variants of Equations (5.10) and (6.14) identified the following subjects as influential: 424, 442, 475, 478, 570 and 584 on all parameters in both sub-models as shown in Figure 6.2. The subjects identified by the developed case-deletion diagnostics are characterised as follows;

Subject 424 had CD4 cell measurements taken at all the visit times from baseline CD4 values (690, 757, 772, 780 and 920 $\mu\ell$). The CD4 cell count values steadily increased from baseline up to the time of administrative censoring, which were within normal range. This subject was randomised to the placebo arm and reported ART use from baseline to month six. This subject

was also identified as influential in Section 5.7.1 above.

Subject 442 had three CD4 cell count measurements, 566, 348 and 124 $\mu\ell$ taken at baseline, week two and month three. The subject consistently reported to be on ART from baseline to the last visit. The subject was randomised to the placebo arm.

Subject 475 had all CD4 cell count measurements taken except month number one as follows: 8, 63, 90, and 71 $\mu\ell$. This subject, despite having a steady increase in the CD4 cell count values from baseline, the measurements were way below the normal range (500 - 1,500 $\mu\ell$). The subject reported ART use at the last two visits only and was randomised to the placebo arm.

Subject 478 had no recording of CD4 cell count at month one and month six. The three measurements taken were at baseline, week two and month three as follows: 34, 96, and 302 $\mu\ell$, which were also relatively below normal range of CD4 cell count values. This subject reported ART use only at month three and was randomised to the active arm.

Subjects 570 and 584 had no CD4 cell count measurements at month one with their respective recordings as follows: 493, 174, 560, and 418 $\mu\ell$ for subject 570 and 9, 15, 25, and 2 $\mu\ell$) for subject 584. These two subjects reported consistent use of ART and were randomised to the active arm. Notably, subject 584 had very low CD4 cell count readings compared to the rest of the influential subjects. All of the influential subjects did not experience constriction and,

thus, were either censored by death or administratively.

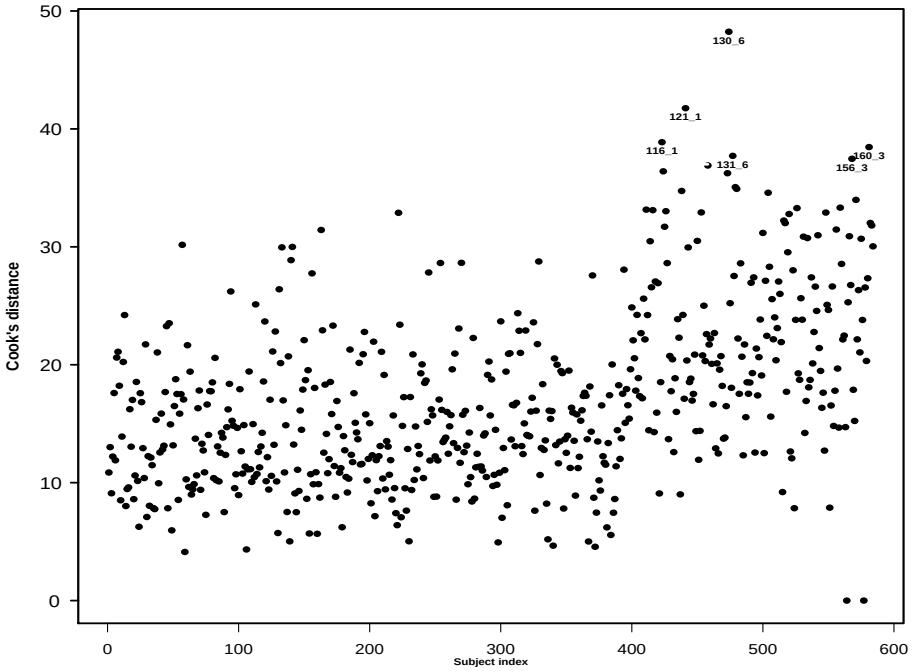


Figure 6.2: Influential subjects for all parameters in the multivariate joint model.

We fitted the multivariate joint model excluding subjects that were identified as influential and observed that there was a slight increase in posterior means for the treatment in the effect, but not significantly different since the 95 % credible intervals overlap. It is noteworthy that there is a significant association between CD4 cell count and risk of constriction in either scenario based on the 95 % credible interval. By contrast, there is no association between ART up take and risk of constriction based on the same metric. It is, further, observed that there is a subject-specific effect of time for both CD4 cell count and ART uptake (Table 6.1)

Table 6.1: Parameter estimates for multivariate joint survival model at subject level.

	M_1		M_2	
Parameter	Posterior Mean(S.D)	95 % C.I.	Posterior Mean (S.D)	95 % C.I.
Longitudinal sub-model				
Intercept	12.80 (0.22)	12.35, 13.23	12.78 (0.23)	12.34 , 13.24
$\sqrt{CD4}$ Time point	0.65 (0.04)	0.58, 0.72	0.66 (0.04)	0.59 , 0.73
Residual σ_e^2	2.80 (0.06)	2.68, 2.93	2.79 (0.07)	2.67 , 2.92
ART Uptake Intercept	-4.73 (0.46)	-5.69, -3.87	-4.78 (0.45)	-5.67 , -3.91
ART Uptake Time point	1.70 (0.14)	1.43, 1.96	1.68 (0.14)	1.41, 1.95
Survival sub-model				
Prednisolone	-0.94 (0.41)	-1.79, -0.15	-0.92 (0.44)	-1.80 , -0.06
Gender (Male)	0.65 (0.45)	-0.18, 1.56	0.67 (0.45)	-0.19, 1.67
NYHA class (III-IV)	0.42 (0.39)	-0.35, 1.21	0.41 (0.40)	-0.40, 1.17
Peripheral Oedema	0.48 (0.39)	-0.29, 1.19	0.50 (0.38)	-0.31, 1.21
Definite TB (Yes)	0.83 (0.43)	-0.04, 1.63	0.82 (0.39)	0.02, 1.57
Haemoglobin status	0.18 (0.09)	-0.01, 0.36	0.19 (0.09)	0.01, 0.37
Association parameter - $\sqrt{CD4}$	-0.02 (0.01)	-0.04, -0.01	-0.02 (0.01)	-0.04 , -0.01
Association parameter - ART Uptake	0.01 (0.01)	-0.01, 0.02	0.01 (0.01)	-0.01 , 0.02

Models: M_1 is the initial model and M_2 is the model after deleting influential subjects
Influential subjects IDs: $424, 442, 475, 478, 570, 584$

6.4 Application on IMPI dataset at observation level

We implemented case deletion diagnostics at observation level for the multivariate joint model survival model using the Cook and Weisberg (1982) formulation as follows:

$$\mathbf{C}^{6.15}(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{[-ij]} - \hat{\boldsymbol{\theta}})' \mathrm{M}^{-1} (\hat{\boldsymbol{\theta}}_{[-ij]} - \hat{\boldsymbol{\theta}}), \tag{6.15}$$

where i denotes the subject and j denotes the time point at which the subject came for a scheduled visit as tabulated in Table 4.1 of Section 4.3.

The following observations were identified as influential after implementing the Cook's based case deletion diagnostics in the multivariate setting as is shown in Figure 6.3:

$27_6, 59_1, 84_6, 149_1, 221_1, 314_1, 366_0, 456_6, 459_0$ and 489_6

The observations had the following characteristics; Observation 27_6 had a CD4 cell count of

465 $\mu\ell$, the subject did not report any ART use at this visit and was randomised in the placebo arm.

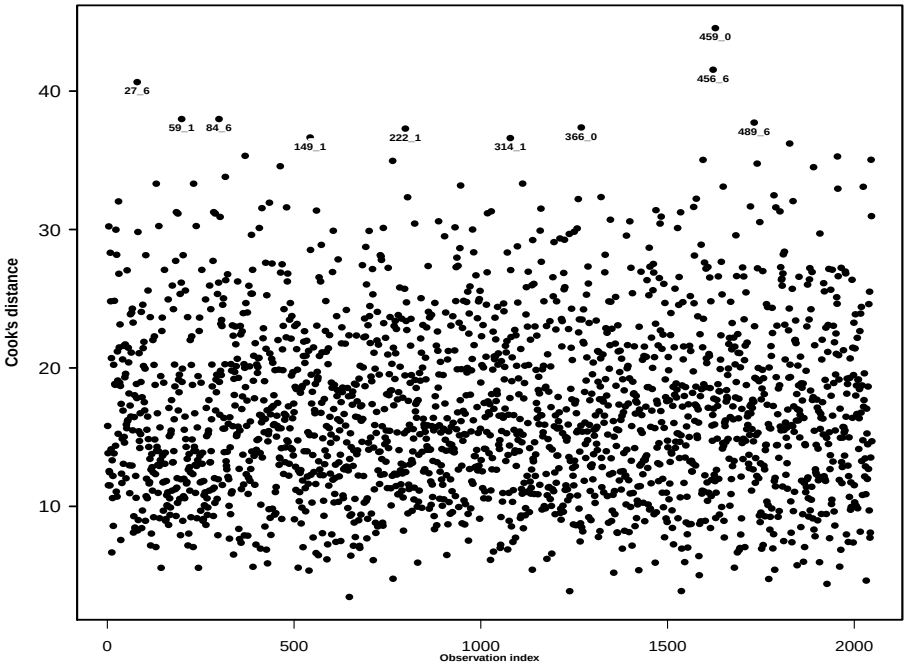


Figure 6.3: Observation level index plot for Cook's statistics in the multivariate joint survival model.

Observation 59_1 had CD4 cell count of 409 $\mu\ell$, and was not on ART at this visit time. This subject was randomised in the active arm. The next observation was, 84_6 with a CD4 cell count reading of 164 $\mu\ell$, was not on ART at the time of the visit and randomised to the active arm. Observation 149_1 and 221_1, reported to be on ART on this visit, had their respective CD4 cell count of 123 and 149 $\mu\ell$; both of these subjects were randomised in the active arm. Observations 314_1 and 459_0 had CD4 cell counts of 223 and 50 $\mu\ell$ respectively, subjects of these observations reported no ART uptake at the time of their visits. These subjects were both randomised in the active arm. Subject 366, whose observation was 366_0, had CD4 cell count of 225 $\mu\ell$ and missing data for ART uptake at this visit. However, on all the subsequent visits the subject reported no use of ART. This subject was randomised in the placebo arm.

Observation 456_6 and 489_6 had the following CD4 cell counts: 338 and 523 $\mu\ell$, respectively, and both of these subjects reported being on ART at the time of the visit. The observation 456_6 came from a subject randomised to active arm, while 489_6 came from a subject randomised to the placebo arm. Both of these observations came from subjects who did not experience any event and thus were censored administratively.

Table 6.2: Observation level parameter estimates for multivariate joint survival model.

	M_3		M_4	
Parameter	Posterior Mean(S.D)	95 % C.I.	Posterior Mean (S.D)	95 % C.I.
Longitudinal sub-model				
Intercept	12.80 (0.22)	12.35, 13.23	12.79 (0.22)	12.36, 13.23
$\sqrt{CD4}$ Time point	0.65 (0.04)	0.58, 0.72	0.65 (0.04)	0.58, 0.73
Residual σ_e^2	2.80 (0.06)	2.68, 2.93	2.81 (0.07)	2.68, 2.94
ART Uptake Intercept	-4.73 (0.46)	-5.69, -3.87	-4.72 (0.46)	-5.69, -3.83
ART Uptake Time point	1.70 (0.14)	1.43, 1.96	1.69 (0.15)	1.40, 1.99
Survival sub-model				
Prednisolone	-0.94 (0.41)	-1.79, -0.15	-0.93 (0.41)	-1.82, -0.16
Gender (Male)	0.65 (0.45)	-0.18, 1.56	0.63 (0.43)	-0.20, 1.51
NYHA class (III-IV)	0.42 (0.39)	-0.35, 1.21	0.42 (0.39)	-0.35, 1.17
Peripheral Oedema	0.48 (0.39)	-0.29, 1.19	0.49 (0.38)	-0.28, 1.25
Definite TB (Yes)	0.83 (0.43)	-0.04, 1.63	0.83 (0.40)	0.06, 1.64
Haemoglobin status	0.18 (0.09)	-0.01, 0.36	0.19 (0.09)	0.03, 0.36
Association parameter - $\sqrt{CD4}$	-0.02 (0.01)	-0.04, -0.01	-0.02 (0.01)	-0.04, -0.01
Association parameter - ART Uptake	0.01 (0.01)	-0.01, 0.02	0.01 (0.01)	-0.01, 0.02

Models: M_3 is the initial model and M_4 is the model after deleting influential observations
Influential observations IDs: $27_6, 59_1, 84_6, 149_1, 222_1, 314_1, 366_0, 456_6, 459_0, 489_6$

We went ahead to carry out a sensitivity analysis by fitting the multivariate joint survival model excluding influential observations (Table 6.2). There were subtle differences in parameter estimates comparing the full model M_3 to the reduced model M_4, which had influential observations excluded in the multivariate joint survival model. It was observed that almost all of the 95% credible intervals for the posterior means overlap for these models, rendering the differences in parameter estimates to be insignificant.

6.5 Summary

We implemented multivariate joint survival model for our case deletion diagnostics; the multivariate joint model had two longitudinal outcomes, which can accommodate more repeated measures outcome. The longitudinal outcomes that were used are, CD4 cell count, which measured on a continuous scale and ART uptake, which was a binary response. The CD4 cell count was square root transformed and modeled as under the assumption of normality in the linear model while ART was modeled under the binomial assumption using the generalised linear model. The two longitudinal outcomes were thus jointly modeled using the generalised linear mixed effects modeling framework given the random effects on the basis that they come from exponential family of distributions.

The estimation of multivariate joint model's parameters was based on Bayesian techniques using Markov Chain Monte Carlo (MCMC); this is an extension of the Gibbs Sampler known as JAGS, which stands for "Just Another Gibbs Sampler." The derivation of posterior distribution of model parameters given the data assumes independence for the longitudinal outcomes as well as event times. In JAGS, the default number of iterations for the generalised linear mixed model is set at 28000 and in the multivariate joint model it is set at 1000 iterations. The MCMC samples used for evaluating Cook's statistics for parameters were extracted from the last row of the matrix that were estimated. In either case of MCMC, sample estimates *burn-in* was at about ten percent of iterations, and therefore, the last row of the estimates are believed to have been generated by a fully converged model.

The case deletion diagnostics approach that was used in the multivariate joint survival model was the same as the standard one. We obtained estimates of all parameters, $\boldsymbol{\theta}$, from the model

that was believed to have converged from the MCMC samples and extracted the variance covariance matrix $\mathbf{M}$. These estimates were for fixed effects from the multivariate generalised linear mixed sub-model for the two outcomes, i.e. CD4 cell count and ART uptake. These estimates also included variance components from the GLMM in addition to fixed effects and association parameters from the survival sub-model. Evaluation of the Cook's statistics was by excluding the i^{th} subject and/or the ij^{th} observation from the modeling process to get a new set of parameter estimates for $\hat{\boldsymbol{\theta}}$ and their corresponding variance covariance matrix $\hat{\mathbf{M}}$.

We, then, computed the Cook's distances for each subject and/or observation using Equations (6.14)and (6.15), respectively. Influential subjects and observations were those that had relatively large values of Cook's distance and the Cook's distances were plotted in an index plot for each subject and observation.

Using the above case deletion diagnostics, the following subjects were identified as influential 424, 442, 475, 478, 570 and 584, while the following observations were identified; 27_6, 59_1, 84_6, 149_1, 222_1, 314_1, 366_0, 456_6, 459_0 and 489_6. We re-fitted the multivariate joint model excluding influential subjects at first and later influential observations. We observed subtle changes in treatment effect in either case Tables 6.1 and 6.2. The changes in treatment effect as well as other parameters were not statistically significant as shown by overlapping 95 % credible intervals for posterior means both at subject and observation level.

It was noted that the association parameter for CD4 cell count, which quantifies how the longitudinal CD4 cell counts values are related to the risk of constriction was statistically significant while that for ART uptake was not Tables 6.1 and 6.2.

The case deletion diagnostics that we implemented can be extended to evaluate Cook's statistics for fixed effects and variance components of both sub-models. These can further be ex-

tended on association parameters quantifying CD4 cell count and ART uptake repeated measures to the risk of constriction.

The estimation of joint models becomes complete by employing numerical integration algorithms, because the integral is intractable given shared random effects. The numerical intergeneration in multivariate joint models is computationally intensive due to the fact that each longitudinal outcome is assumed to have its own set of random effects over which integration takes place. As a result, the dimensions of random effects vector gets large and increases computational burden.

This analysis included 587 subjects from the IMPI dataset, which means fitting 587 multivariate joint survival models to re-estimate parameters at subject level and about 2050 multivariate joint survival models to re-estimate parameters at observation level. It took about 20 minutes to fit one multivariate joint model on a Core i7 personal computer installed with Windows 10. This, therefore, meant that we would have to wait several days to run full case deletion diagnostics on this dataset ($\approx$ 8 days at subject level and $\approx$ 28 days at observation level).

We were forced to fit and run the multivariate joint models on University of Cape Town's High Performance Computing (HPC) facilities and further used parallel computing on these facilities. This reduced computing time by over 80 % (20 minutes to 4 minute) for one multivariate joint model. This limitation led us to evaluating Cook's statistics at subject and observation level for all parameters. However, extensions to parameter specific diagnostic is straight forward.

Chapter 7

Diagnostics for joint models for longitudinal and survival data with multiple failure types

7.1 Introduction

In the following Chapter we extended joint model diagnostics implemented in Section 5.3 of Chapter 5 to multiple failure types. We begin with an introduction of competing risk in the survival sub-model of the joint model. This is augmented by the longitudinal process modeled as (2.1), which is a sub-model in the joint model application using the IMPI dataset with the two failure types.

This Chapter contributes case-deletion diagnostics based on Cook's statistics implemented in the competing risk setting. We discussed results obtained from these diagnostics and their application.

7.2 Multiple failure times

Multiple failure times or Competing risks concern the scenario where there is a possibility of more than one cause for an event(s). These events could be causes of death in medical research settings, only the first of these to occur is observed. In other settings, observations after the first event may be observable, and may not be of interest. A competing risks model

can be represented graphically with an initial state being generally event-free and a number of different endpoints, as shown in Figure 7.1 below (Putter, Fiocco, and Geskus 2007).

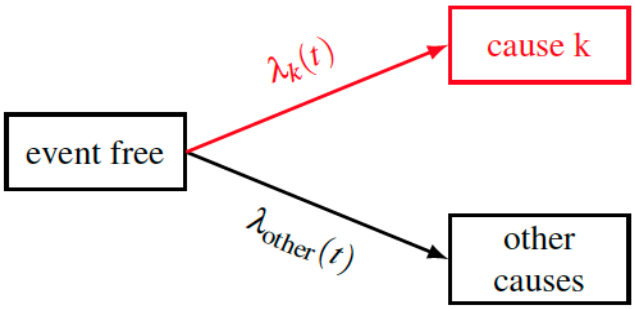


Figure 7.1: A multi-state approach for cause-specific hazard.

In the medical field, there are some examples of multiple failure times for instance in bone marrow transplantation, death from different kinds of infections are possible *viz* bacterial, viral, or fungal and even death due to relapse, let alone other causes. In rear diseases such as cancer, death due to cancer may be of interest but there would be deaths due to other causes such as surgical mortality and old age which are in essence are competing risks. In certain situations, one could be interested in time to relapse, where the competing risk is death due to any cause.

There, are equally, good illustrations in other fields that include failure of different components in a system, in industrial reliability testing or time to part or full-time employment after graduating from university. The subject of competing risks dates as far back as the 18^{th} century at the time Bernoulli (1760) studied the possible consequences of eradication of smallpox on mortality rates. Understandably, the problem of estimation of cause specific failure probabilities after exclusion (or changing) of one of the competing risks has been of great importance and has been the subject of much debate in the 1970s (Gail 1975; Prentice et al. 1978).

7.2.1 Approaches to multiple failure times (Competing risks)

The data observed in a competing risks model is represented by the time to event T, the cause of event is D, and possibly a covariate vector $\mathbf{X}$. We made inferences based on the joint distribution of T and D, possibly given $\mathbf{X}$. The principle concept in competing risk models is the cause-specific hazard function, which is the hazard of an event from a given cause in the presence of other competing events for participant i presented as follows:

$$\lambda_{ik}(t) = \lim_{\delta t \to 0} \frac{\text{Prob}(t \leq T^* < t + \delta t, \delta_i = k | T^* \geq t)}{\delta t}, \quad t > 0 \tag{7.1}$$

The cause-specific hazard can be estimated from the data as illustrated in the expression (7.2) below, which constitutes all important information that can be observed from the data. It is also noted that anything uniquely derived from the cause-specific hazard can be estimated. In the past, techniques implemented for competing risks models treated these models as a multivariate event time model, where each individual is assumed to have a potential event time for each type of event. The earliest of these events is actually observed and others are latent. Let $\tilde{T}_k$ denote event time due to cause k then under this assumption we only observe $T = \min\{\tilde{T}_k\}$ and D. In this case D is an index variable, which specifies which event happened first. If a participant is censored for all events by end of study or loss to follow-up, $D = 0$ where an extra censoring distribution $C \sim G$ is introduced and this distribution is assumed to be independent of all the other events.

Anything that can be uniquely determined by the cause-specific hazards is estimable (Putter, Fiocco, and Geskus 2007) with the cumulative cause-specific hazard defined by

$$\Lambda_k(t) = \int_0^t \lambda_k(s) ds \tag{7.2}$$

and the survival function given by

$$S_k(t) = \exp\Big(-\Lambda_k(t)\Big).$$

It has to be noted that, although $S_k(t)$ can be estimated, it should not be interpreted as a marginal survival function; this interpretation is correct if the competing event time distributions and the censoring distribution are independent. This, therefore, means that, the marginal distribution describes the event time distribution in the situation that the competing events do not occur. If we define

$$S(t) = \exp\left(-\sum_{k=1}^{K}\Lambda_k(t)\right) \tag{7.3}$$

then gives the probability of not having failed from any cause at time t. The cumulative incidence function of cause k, $I_k(t)$, is defined by the probability Prob$(T \leq t, D = k)$ of failing from cause k before time t. This can be expressed in terms of the cause-specific hazards as follows:

$$I_k(t) = \int_0^t \lambda_k(s)S(s)ds \tag{7.4}$$

There are several alternative names for this function, namely 'crude cumulative incidence function' or 'sub-distribution function.' The latter name has its origin in the fact that the cumulative probability for an event from cause k remains below one, $I_k(\infty) =$ Prob$(D = k)$, hence it is not a proper probability distribution.

As noted by Putter, Fiocco, and Geskus (2007), as events from causes other than k are treated as censored, the naive Kaplan-Meier estimate of the probability of failing from cause k before or at time t is estimating

$$1 - S_k(t) = \int_0^t \lambda_k(s)S(s)ds$$

The difference with the cumulative incidence function $I_k(t)$ from equation (7.4) is that, $S(s)$ is replaced by $S_k(s)$. Since $S(t) \leq S_k(t)$, we get $I_k(t) \leq 1 - S_k(t)$, with equality at t if there is no competition, meaning that if $\sum_{j=1,j\neq k}^{K}, \Lambda)j(t) = 0$ which shows that there is bias in the Kaplan–Meier estimator.

The estimation of the cumulative incidence functions proceeds as follows:

Let $0 < t_1 < t_2 < \cdots < t_N$ be ordered distinct time points at which events of any cause occur. Let d_{kj} denote the number of participants failing from cause k at t_j, and let $d_j = \sum_{k=1}^{K} d_{kj}$ denote the total number of events from any cause at t_j. If there are no ties, only one of the d_{kj} equals one for a given j, and $d_j = 1$. The expressions are also valid even when there are tied events. Let n_j be the number of participants that is at risk, these are participants that are still in follow-up and have not experienced an event from any cause at time t_j. The overall survival probability $S(t)$ at t can be estimated, without considering the cause of an event, by the Kaplan–Meier estimator;

$$\hat{S}(t) = \prod_{j:t_t \leq 1} \left(1 - \frac{d_j}{n_j}\right). \tag{7.5}$$

As demonstrated by Putter, Fiocco, and Geskus (2007) the discretized version of the cause-specific hazard of equation (7.1) is

$$\lambda_k(t) = \text{Prob}\Big(T = t_j, D = k | T > t_j - 1\Big), \tag{7.6}$$

this quantity would be estimated by

$$\hat{\lambda}_k(t_j) = \frac{d_{kj}}{n_j}, \tag{7.7}$$

which is the proportion of participants at risk that fail from cause k. Expression (7.5) can also be written as follows:

$$\hat{S}(t) = \prod_{j:t_t \leq 1} \left(1 - \sum_{k=1}^{K} \hat{\lambda}_k(t_j)\right). \tag{7.8}$$

The unconditional probability of failing from cause k at t_j , $p_k(t_j) = \text{Prob}(T = t_j, D = k)$ is the product of the hazard and the probability of being event-free at t_j, and is estimated as

$$\hat{p}_k(t_j) = \hat{\lambda}_k(t_j)\hat{S}(t_{j-1}), \tag{7.9}$$

and finally, the cumulative incidence $I_k(t)$ of cause k at t is estimated as the sum of these terms for all time points before t.

7.3 Joint model for longitudinal and survival data with multiple failure times

It is common in longitudinal studies to observe multiple failure types besides collecting information about bio-markers. This simply means that during study follow-up each participant may experience one of K distinct events or may be censored. In this Section, we limit our focus on one longitudinal marker and time to two competing events. The longitudinal marker considered here is square root CD4 cell count and the competing events are constriction and death whichever occurs first.

In the time to event analysis we implemented a cause-specific hazard regression which assumes separate relative risks for each event (Putter, Fiocco, and Geskus 2007).

These events could be analyzed separately using a Cox regression model by treating one event as a censored observation. However, when interest is in assessing the strength of association between a longitudinal time dependent marker on the risk of an event a joint modeling approach is appropriate.

In the case of handling multiple failures types the notation in the survival process is extended as follows. Let the number of different failure types be K; therefore $T_{i1}^*, ..., T_{ik}^*$ denotes the

true failure times for each one of them. The observed data for the i^{th} subject, therefore, comprises of observed event times as follows:

$$T_i = \min(T_{i1}^*, ..., T_{ik}^*, C_i),$$

where C_i is the censoring time and the event indicator is $\delta \in (0, 1, ..., k)$ with zero corresponding to censoring and $1, ..., k$ to competing events.

The Cox model for k^{th} competing event is as follows:

$$h_{ik}(t) = h_{0k}(t) \exp\{\mathbf{X}_i \boldsymbol{\beta}_k + \alpha_k m_i(t)\}, \quad k = 1, 2, \ldots, K \tag{7.10}$$

which has the effect of baseline covariates $\mathbf{X}_i$ and the current value for the longitudinal marker $m_i(t)$. The joint model is completed by including a suitable mixed model for the longitudinal outcome $Y_i(t)$. The estimation of the joint model is similar to the previously described process for estimation of the joint model. The slight difference is in the construction of the likelihood part of the event process.

$$\begin{aligned} p(T_i, \delta_i | b_i, \theta) = \prod_{k=1}^{K} \Big[h_{0k} \exp\{\mathbf{X}_i' \boldsymbol{\beta}_k + \alpha_k m_i(t)\} \Big]^{I\delta_i = k} \\ \times \exp\Big(- \sum_{k=1}^{K} \int_0^{T_i} h_{0k}(s) \exp\{\mathbf{X}_i' \boldsymbol{\beta}_k + \alpha_i m_i(s)\} ds \Big). \end{aligned} \tag{7.11}$$

This part of the likelihood is computed by transforming the data into the competing risk format. In this format, there are k rows, one for each possible failure type. Event times T_i for each subject are repeated k times with two indicator variables one identifying the cause and another indicating whether the corresponding event type is the one that occurred.

We fit the joint model to the IMPI dataset treating constriction and death as competing events. We assumed a random intercept linear mixed model for the square root CD4 cell count for the

longitudinal sub-model as in model (2.36) of Section 2.7.2 as follows;

$$
\begin{aligned}
Y_{ij} &= \beta_0 + t_{ij}\beta_1 + b_i + \epsilon_{ij} \\
b_i &\sim \mathcal{N}(0, \sigma_b^2), \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \\
b_i &\perp\!\!\!\perp \epsilon_{ij}
\end{aligned}
$$

For the survival sub-model a cause-specific Cox model was implemented in the IMPI dataset treating constriction and death as competing events as follows:

$$
\begin{aligned}
h_{i1}(t) &= h_{01}(t)\exp\{\text{Pred}\beta_{11} + \alpha_1 m_i(t)\} \\
h_{i2}(t) &= h_{02}(t)\exp\{\text{Pred}(\beta_{11} + \beta_{21}) + (\alpha_1 + \alpha_2)m_i(t)\},
\end{aligned}
\tag{7.12}
$$

where Pred is treatment, i.e. Predinolone (Active agent) or Placebo.

The formulation of parameters in model (7.12) above have mean that β_{11} and α_1 denote the effect of treatment and square root CD4 cell count on the risk of constriction, where as β_{21} and α_2 denote the additional effect of treatment and square root CD4 cell count on the risk of death. If $\beta_{21} = 0$ it means that the hazard ratio for treatment is the same for both constriction and death. It is also assumed in both of the cause-specific Cox models that the baseline risk function is different and fully unspecified.

7.4 Results of case deletion diagnostics in competing risk

In this Chapter, we have outlined the theory behind analysis of competing events in standard survival models using two approaches, namely cause-specific and sub-distribution hazards. We have adopted the cause-specific approach for reasons specified in Section 7.2.1 evaluated using equation (7.4). We further outlined extensions of analysis of competing survival events in the presence of a bio-marker(s) that is measured repeatedly over time in Section 7.3. In this

analysis, we have focused on one longitudinal outcome, which is CD4 cell count. It is possible to include multiple outcomes in the analysis using similar extensions described in Chapter 6. The main contribution of this Chapter is to implement case deletion diagnostics for joint survival model in a competing risk setting.

7.4.1 Subject level case deletion diagnostics in multiple failure types

The approach taken is that of Singini, Mwambi, and Gumedze (2018) which is based on the Cook's statistics (Cook and Weisberg 1982)

$$\mathbf{C}^{7.1}(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}})' \mathbf{M}^{-1} (\hat{\boldsymbol{\theta}}_{[-i]} - \hat{\boldsymbol{\theta}}), \tag{7.13}$$

where $\mathbf{M}$ is the variance covariance matrix for all parameters estimates $\hat{\boldsymbol{\theta}}$. The vector of parameters $\boldsymbol{\theta}$ are estimated using model (7.12), which includes fixed effects from both sub-models, variance components from the linear mixed model and two association parameters for each event.

Table 7.1: Parameter estimates for joint survival model in competing risk at subject level.

	M_5			M_6		
Parameter	Estimate (S.E)	95 % C.I.	p-value	Estimate (S.E)	95 % C.I.	p-value
Longitudinal sub-model						
Fixed effects						
Intercept	12.66 (0.29)	(12.09, 13.24)	<0.001	12.70 (0.29)	(12.13, 13.26)	<0.001
Prednisolone	-0.20 (0.39)	(-0.96, 0.57)	0.62	-0.21 (0.40)	(-0.93, 0.57)	0.60
$\sqrt{CD4}$ Time point	0.61 (0.07)	(0.53, 0.69)	<0.001	0.62 (0.03)	(0.54, 0.69)	<0.001
Variance components						
$\hat{\sigma}_{b_0}$	4.66			4.66		
$\hat{\sigma}_{b_1}$	0.41			0.40		
$\hat{\rho}$	-0.21			-0.21		
Residual $\hat{\sigma}_e$	2.93			2.95		
Survival sub-model						
Prednisolone	0.51 (0.54)	(0.17, 1.47)	0.21	0.62 (0.49)	(0.23, 1.61)	0.32
Gender (Male)	0.67 (0.58)	(0.21, 2.10)	0.49	0.56 (0.46)	(0.23, 1.37)	0.20
NYHA class (III-IV)	0.49 (0.42)	(0.20, 1.03)	0.06	0.49 (0.35)	(0.24, 0.97)	0.04
Peripheral Oedema	0.47 (0.36)	(0.23, 0.96)	0.04	0.50 (0.35)	(0.25, 1.00)	0.05
Definite TB (Yes)	0.66 (0.37)	(0.31, 1.34)	0.24	0.66 (0.37)	(0.32, 1.35)	0.26
Haemoglobin status	0.84 (0.13)	(0.65, 1.07)	0.16	0.87 (0.11)	(0.71, 1.08)	0.21
Prednisolone $\times$ death	5.32 (0.68)	(1.38, 20.49)	0.02	4.09 (0.65)	(1.15, 14.53)	0.03
Association parameter : constriction	1.09 (0.03)	(1.02, 1.16)	0.01	1.08 (0.02)	(1.04, 1.12)	<0.001
Association parameter : death	0.96 (0.04)	(0.87, 1.02)	0.17	0.95 (0.03)	(0.90, 1.01)	0.08

Models: M_5 is the initial model and M_6 is the model after deleting influential subjects
Influential subjects IDs:- $97, 185, 221, 267, 281, 359,$ and 497

The following seven subjects were identified as influential by using model (7.13)above: 97, 185, 221, 267, 281, 359 and 497 as depicted in the index plot of Figure 7.2. The characteristics of these subjects were:

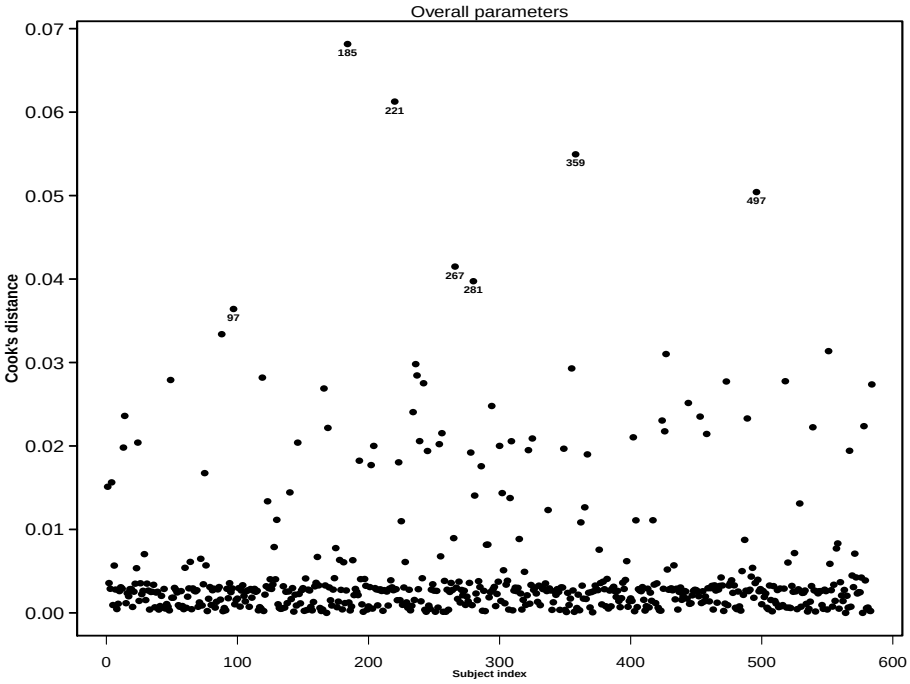


Figure 7.2: Subject level index plot for Cook's statistics in competing risk for joint survival model.

Subject 97 had CD4 cell measurements taken at baseline and week two only with the following values (109 and 119 $\mu\ell$). This subject was randomised to the active treatment arm and reported no ART experience. The subject was recorded to have died after 150 days. Subject 185 had three CD4 cell count measurements taken at baseline visit, week two, month three and the last visit with the following values: 72, 160, 167 and 181 $\mu\ell$. The CD4 cell count values increased at each subsequent visit the measurements were taken. This subject was randomised to the placebo arm and reported ART use from baseline visit up to the last visit, where the subject was censored. Subject 221 also had four CD4 cell count measurements taken except the last visit as follows: 348, 149, 273 and 187 $\mu\ell$. The subject had no observed pattern for the CD4 cell count with time and reported ART use on all the scheduled study visits. This subject was randomised to the active treatment arm and did not experience any event. Subject 267

had CD4 measurements taken starting from month one up to the last scheduled month with these values: 180, 287 and 241 $\mu\ell$. This subject reported no ART experience at all, and was ramdomised to the active treatment arm and had not experience any event of interest. Subject 281 had all CD4 cell count measurement taken as follows: 143, 157, 110, 229 and 281 $\mu\ell$ with no observable trend over time. This subject reported ART experience on all visits except visit month three. The subject was randomised to the active treatment arm and was censored at the last visit. Subject 359 had missing CD4 cell count measurement at month one visit only, the rest of the CD4 cell count values were: 375, 465, 349 and 288 $\mu\ell$. This subject reported no ART experience, was randomised to the active treatment arm and experienced none of the events under observation. The last influential subject was 497 who had low CD4 count on all the scheduled visits (6, 32, 28, 82 and 87). This subject reported ART use starting from month one visit up to the last scheduled visit. The subject was randomised to the placebo arm and did not experience any of the events of interest.

We conducted a sensitivity analysis by re-fitting the joint model by excluding the seven subjects who were identified as influential Table 7.1. There were little effects on the estimates from the linear mixed model for both fixed effects and variance components. It was also observed that there were marginal changes in parameter estimates in the survival sub-model, especially for NYHA class and Definite TB. Estimates for the following risk factors: Treatment, Gender, Peripheral Oedema and Haemoglobin status, were sensitive after re-fitting the model without influential subjects in the survival sub-model (models $\mathbf{M}_5$ and $\mathbf{M}_6$). The 95 % confidence intervals for the estimates, however, overlap despite the observed sensitivity, therefore, not statistically significant.

7.4.2 Observation level case deletion diagnostics in multiple failure types

We went further to illustrate the developed case deletion diagnostics at observation level using the same approach as in model (7.13). At the observation level we evaluated the following Cook's statistic:

$$\mathbf{C}^{7.2}(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}_{[-(ij]} - \hat{\boldsymbol{\theta}})' \mathrm{M}^{-1} (\hat{\boldsymbol{\theta}}_{[-(ij)]} - \hat{\boldsymbol{\theta}}), \tag{7.14}$$

where M is the variance covariance matrix for all parameters estimates $\hat{\boldsymbol{\theta}}$ for the i^{th} observation at the j^{th} time point. The vector of parameters $\boldsymbol{\theta}$ are estimated using model (7.12), which includes fixed effects from both sub-models, variance components from the linear mixed model and two association parameters for each event. We adopt the defining of an observation from table 4.1 on page 81.

The following observations were identified as influential using evaluation from model (7.14) above; $46_6, 61_3, 102_{0.5}, 113_0, 120_0, 156_{0.5}, 216_6, 235_3, 329_{0.5}, 370_{0.5}, 382_3$, and $441_{0.5}$ (Figure 7.3). These observations were from subjects with the following characteristics: Subject 46 had 139 $\mu\ell$ of CD4 cell count at visit six and did not report any exposure to ART on all the six visits. This subject was randomised to the placebo arm and did not experience any event. Subject number 63 had 329 $\mu\ell$ CD4 cell count measurement at visit three, reported ART use at visits three and six. The subject did not experience any of the events and was randomised to the placebo arm. Subject 102 had recorded 162 $\mu\ell$ of CD4 cell count at week two visit and reported to ART exposure on all the six visits but did not experience any event. This subject was also randomised to the placebo arm. Subject 113 was randomised to the active treatment arm, had a CD4 cell count reading of 282 $\mu\ell$ at baseline. This subject reported ART exposure at the last two visits and exited the study without experiencing an event. Subject 120

had 112 $\mu\ell$ CD4 cell count at study entry, reported ART exposure at the last two visits and randomised to the placebo arm. Subject 156 had a reading of 248 $\mu\ell$ of CD4 cell count at week two visit, reported ART exposure at the last two visit and was randomised to the active treatment arm. Subject 216 had 441 $\mu\ell$ of CD4 cell count at visit six, reported ART use only at the last visit and was randomised to the placebo arm. Subject 235 had 116 $\mu\ell$ of CD4 cell count measurement at visit number three, reported ART exposure at the last two visits and was randomised to the placebo arm. Subject 329's CD4 cell count reading was 21 $\mu\ell$ at week two visit, this subject reported ART exposure at week two and month three visits. The subject was randomised to the placebo arm. Subject 370 had a CD4 cell count measurement of 64 $\mu\ell$ at week two visit, reported no exposure to ART at all the visits and was randomised to the active treatment arm. Subject 382 had a 406 $\mu\ell$ reading of CD4 cell count at visit three and this subject reported ART use at all the visits having been randomised to the placebo arm. Lastly subject 441 had a recording of 220 $\mu\ell$ of CD4 cell count at week two visit, reported ART use at the same visit and was randomised to the active treatment arm. All influential observations came from subjects who did not experience the events of interest.

Table 7.2: Parameter estimates for joint survival model in competing risk at observation level.

	M_5			M_7		
Parameter	Estimate (S.E)	95 % C.I.	p-value	Estimate (S.E)	95 % C.I.	p-value
Longitudinal sub-model						
Fixed effects						
Intercept	12.66 (0.29)	(12.09, 13.24)	<0.001	12.75 (0.31)	(12.15, 13.35)	<0.001
Prednisolone	-0.20 (0.39)	(-0.96, 0.57)	0.62	-0.29 (0.40)	(-1.08, 0.50)	0.47
$\sqrt{CD4}$ Time point	0.61 (0.07)	(0.53, 0.69)	<0.001	0.62 (0.04)	(0.54, 0.69)	<0.001
Variance components						
$\hat{\sigma}_{b_0}$	4.66			4.65		
$\hat{\sigma}_{b_1}$	0.41			0.40		
$\hat{\rho}$	-0.21			-0.21		
Residual $\hat{\sigma}_e$	2.93			2.94		
Survival sub-model						
Prednisolone	0.51 (0.54)	(0.17, 1.47)	0.21	0.52 (0.49)	(0.16, 1.65)	0.26
Gender (Male)	0.67 (0.58)	(0.21, 2.10)	0.49	0.66 (0.46)	(0.17, 2.55)	0.54
NYHA class (III-IV)	0.49 (0.42)	(0.20, 1.03)	0.06	0.46 (0.35)	(0.17, 1.18)	0.11
Peripheral Oedema	0.47 (0.36)	(0.23, 0.96)	0.04	0.49 (0.35)	(0.23, 0.96)	0.04
Definite TB (Yes)	0.66 (0.37)	(0.31, 1.34)	0.24	0.66 (0.37)	(0.32, 1.34)	0.24
Haemoglobin status	0.84 (0.13)	(0.65, 1.07)	0.16	0.84 (0.11)	(0.63, 1.12)	0.25
Prednisolone $\times$ death	5.32 (0.68)	(1.38, 20.49)	0.02	5.0 (0.65)	(1.17, 20.51)	0.03
Association parameter : constriction	1.09 (0.03)	(1.02, 1.16)	0.01	1.09 (0.02)	(1.01, 1.18)	0.04
Association parameter : death	0.96 (0.04)	(0.87, 1.02)	0.17	0.94 (0.03)	(0.86, 1.03)	0.20

Models: M_5 is the initial model and M_7 is the model after deleting influential observations
Influential subjects IDs:- $46_6, 61_3, 102_1, 113_0, 120_0, 156_1, 216_6, 235_3, 329_1, 370_1, 382_3,$ and 441_1

Overall the event rate was 9.7 % (57) with 30 deaths, 25 constrictions and 2 experienced constriction and later died.

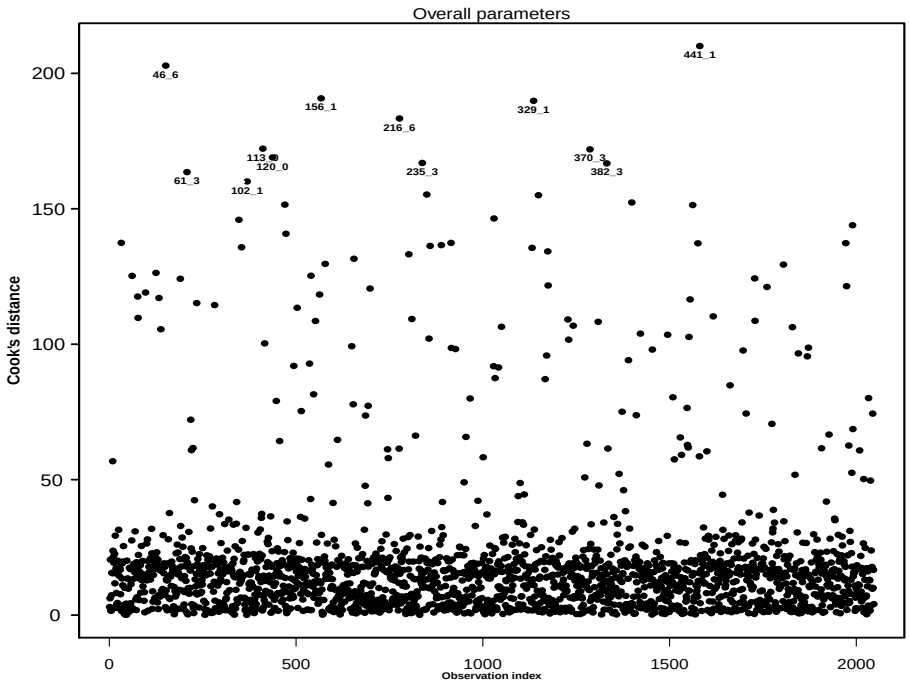


Figure 7.3: Observation level index plot for Cook's statistics in competing risk for joint survival model

Similar to Sub-Section 7.4.1 above, we re-ran the joint model with the two competing events whilst excluding influential observation to investigate the sensitivity of estimates (Table 7.2). The estimates by models M_5 and M_7, just like estimates obtained at subject level, were not sensitive to the change in the linear mixed model. This is for both fixed effects and variance components.

The survival sub-model estimates had similar estimates with subtle differences for NYHA class and Peripheral Oedema but not significant. The association parameter for constriction was significant compared to that for death.

7.5 Summary

We illustrated our case deletion diagnostics for joint models in multiple failure types as follows: there are two competing events that were used in the evaluation of the diagnostics, which are based on Cook's statistics and these were constrictive pericarditis and death. The two competing events were treated as cause-specific in the survival sub-model (Section 7.3 of page 157). We limited the illustration of the diagnostics to one longitudinal outcome; CD4 cell count which was measured on a continuous scale. We square root transformed CD4 cell count to achieve normality, which is one of the assumptions adopted in joint models with a continuous outcome, among others. We then modeled square root CD4 using a linear mixed effects model.

The estimation of the joint model parameters was based on maximum likelihood with a slight modification to the construction of the likelihood on the survival part of the joint model (7.11), which incorporates multiple failure types.

We adopted a similar approach by Singini, Mwambi, and Gumedze (2018) to evaluate Cook's distance for the i^{th} subject and/or the ij^{th} observation (Equations (7.13) and (7.14)). We got estimates of all parameters $\boldsymbol{\theta}$ from the joint model, including their respective variance covariance matrix $\mathbf{M}$. These estimates were for fixed effects and variance components from the linear mixed effects sub-model. The survival sub-model provided estimates for fixed effects and association parameter for the two competing events including the interaction term for the association between death and time. The calculation of Cook's statistics was done using the standard case deletion approach, which iteratively excludes the i^{th} subject and/or the ij^{th} observation from the modeling process to get a new set of parameter $\hat{\boldsymbol{\theta}}$ and their corresponding

variance covariance matrix $\hat{\mathbf{M}}$.

Outliers or influential subjects and/or observations were those that had relatively large values of the calculated Cook's distance. We then plotted these Cook's distances using an index plot for each subject and observation.

Based on the calculated Cook's distance for each subject, the following seven subjects were identified as influential: $97, 185, 221, 267, 281, 359$ and 497, with the following twelve observations, similarly, identified: $46_6, 61_3, 102_1, 113_0, 120_0, 156_1, 216_6, 235_3, 329_1, 370_1, 382_3$, and 441_1.

We went ahead to conduct a sensitivity analysis by re-running the joint model excluding influential subjects at first and later influential observations. We observed some changes in parameter estimates in the survival component of the joint model with multiple failure types. on the other hand, there were negligible changes in the linear mixed effects sub-model at subject level. At the observation level, there were subtle differences in the estimates after re-fitting the model both for the linear and the survival components of the joint model (Tables 6.2 and 7.1). The changes in parameter estimates were not statistically significant as shown by overlapping 95 % confidence intervals at both subject and observation level.

It was observed that the association parameter for CD4 cell count, which quantifies how the longitudinal CD4 cell counts values are related to the risk for constriction and death was statistically significant for constriction only (Tables 6.2 and 7.1). This result is consistent with the original results from the main study (Mayosi et al. 2014).

The case deletion diagnostics that we have illustrated can be implemented for each of the parameters being estimated, i.e. for fixed effects, variance components and association parameter.

Implementing case deletion diagnostics can be resource intensive in terms of computing time

due to the estimation of the joint model which employs numerical integration techniques since the integral is intractable given the random effects. In our case we implemented these diagnostics using parallel computing in order to be efficient.

Chapter 8

Discussion, Conclusions and future research direction

8.1 Summary and diagnostics for joint models for longitudinal and survival data

The development of effective treatment regimens is usually through conducting randomised clinical trials which are regarded as phase III-IV studies. The design of these trials, is normally, longitudinal in nature, where a biomraker or a set of biomarkers are repeatedly measured or observed and recorded at selected time points of the study, which might be for short periods of time, say three months or even extend for years. These biomarkers are what is considered as response variables at the time of statistical analysis. In medical research, it is often of interest to observe simultaneously a survival outcome using the same longitudinal design, e.g. death. Longitudinal study designs in medical research potentially have three main aims under investigation and these are: an assessment of a cross-sectional effect, longitudinal effect and calculation of survival times. The cross-sectional effect investigates how the effect of treatment differs on average at specific time points, while the longitudinal effect investigates if there exists differences in the treatment effect. The calculation of survival times involves estimating the risk of an event while adjusting for censoring, which is usually considered non-informative (Rizopoulos 2012, p.16). In the past three decades or so the two main processes which are, the longitudinal and survival process, were analysed separately.

It is well known that longitudinal measurements from each subject over time are positively correlated. This, therefore, means that standard statistical tools which are based on the independence assumption, such *t*-test and simple linear regression based on ordinary least squares estimation (OLS) can not be used for longitudinal data analysis. In order to provide valid inferences from the analysis of longitudinal outcomes, there is need to account for correlation. In the past decades, linear mixed models (LMM) have been extensively studied and developed (Diggle, Liang, and Zeger 1994; Laird and Ware 1982; Molenberghs and Verbeke 2000) among others. These models account for both the within and between subject correlation structure through random effects and it is useful in medical research including other disciplines of biological sciences.

On the other hand, the most commonly used approach to analyse a time to event outcome is the semi-parametric Cox proportional hazards model (Cox 1972b). This model can be extended to handle time varying covariates which make use of all information collected as opposed to using baseline information only which was the standard practice some four decades ago. The extended Cox model also known as Anderson and Gill is widely used based on the counting process formulation (Andersen and D 1982; Fleming and Harrington 1991). The motivation for this formulation is to imagine that the occurrence of events is a realization of a slow Poisson process (Rizopoulos 2012). This model is ideal when the time-varying covariates are external or exogenous in nature under the following three assumptions: (1) the covariates are predictable (2) measured without error and (3) a complete path for the covariates is fully specified. When modeling internal or exogenous time-varying covariates for example, biomarkers it is unrealistic to make such assumptions because biomarkers do not have the same value between follow-up visits, which could be months apart, at the same time the covariates are mea-

sured with error and their complete path cannot be fully specified. It is therefore imperative that using the extended Cox model for analysis of endogenous covariates such as biomarkers is not optimal, which therefore begs a modeling framework that is specifically designed to correct for special features of endogenous time-varying covariates (Rizopoulos 2012).

It has been outlined that the extended Cox model is ideal for exogenous time-varying covariates, as such, not appropriate for longitudinal biomarkers. In cases where the main objective of a statistical analysis is to assess the strength of association between endogenous time-varying covariates and survival times; the recommended modeling framewrok, that has been an active research area for the past two and a half decades, is joint modeling for longitudinal and survival data (Faucett and Thomas 1996; Henderson, Diggle, and Dobson 2000; Tsiatis, Degruttola, and Wulfsohn 1995; Tsiatis and Davidian 2004; Wulfsohn and Tsiatis 1997). The motivation towards this modeling framework has been to marry the survival model with an appropriately specified measurement model for the endogenous covariates to account for the limitations in the extended Cox model (Rizopoulos 2012).

There are several approaches taken towards jointly modeling a time-to-event outcome and a longitudinal outcome, with the initial one being a Two-stage approach. The Two-stage joint model could be preferred because of its simplicity, since it does not need numerical computation of the likelihood. The Two-stage joint model is done in two main steps (Self and Pawitan 1992), the first is to fit a linear mixed model just as in model (4.4). When this is done, extract the predicted values from the model which are, then, used as time-varying covariate in the extended Cox model. The predicted values have the effect of accounting for measurement error and there is no assumption that the time-varying covariate is constant between two consecutive measurements. The Two-stage joint model is relatively easy to implement us-

ing standard software but it has been shown that it produces biased results (Dafni and Tsiatis 1998; Sweeting and Thompson 2011; Tsiatis and Davidian 2001). This is, particularly, so because the errors generated from the first step are not carried through to the next step of the extended Cox model. It has, however, been shown by Mauff et al. (2018), that the bias that the Two-stage joint model produces is addressed using importance sampling in the Bayesian framework. In this set up Mauff et al. (2020), propose for the correction of estimates derived from the two-stage approach using importance sampling weights. All the parameter estimates obtained from the survival sub-model, the longitudinal sub-model and random effects (variance components) are considered a weighted sample from the full posterior of the joint model (Mauff et al. 2020).

We recognise that diagnostics as well as model assessment tools have not been fully developed for the Two-stage joint model. We were, therefore, motivated by the simplicity of implementing a Two-stage joint model and the recent development that corrects for the bias that we proposed, a diagnostic approach based on the variance shift outlier model (Gumedze et al. 2010). In our approach we fitted the linear mixed model (4.4) to the longitudinal outcome and run diagnostics based on Cook's statistics to identify outliers. Once we have identified outliers and/or influential subjects, we then implement the VSOM to the longitudinal sub-model to down-weight these subjects (model (4.9)). We then extracted the predicted values from the down-weighted model and used them as time-varying covariate in the extended Cox model. We accounted for multiple testing by adjusting $p-$ values using the Benjamini and Hochberg (1995) procedure, while the problem of regularity conditions under the null hypothesis, where the parameter lies in the boundary of the parameter space, was addressed by using a $0.65\chi^2_0 + 0.35\chi^2_1$ mixture distribution on p-values of the likelihood ratio tests.

We implemented this in a univariate joint model, but this approach can be extended in the multivariate joint model, where the computational benefits would be enormous as has been discussed in Chapter 6.

Another approach to joint modeling that has been an active research area is the shared parameter joint model. The estimation of these joint models that have been studied are based on semiparametric maximum likelihood (Henderson, Diggle, and Dobson 2000; Hsieh, Tseng, and Wang 2006; Wulfsohn and Tsiatis 1997). There has been investigations by Zeng, Cai, et al. (2005), who established that estimators from the joint likelihood are not only asymptotic but also efficient. In the Bayesian paradigm, estimation of these joint models through MCMC techniques have been studied by Chi and Ibrahim (2006), R Brown and G Ibrahim (2003), Wang and Taylor (2001), and Xu and Zeger (2001) among other authors. The other type of joint modeling framework is known as latent class joint model, which are commonly used for personalised medicine and prediction modeling (Proust-Lima et al. 2009, 2014). These models are beyond the scope of this thesis, which focuses on joint models that assess the strength of association between a bio-marker and risk of an event (Henderson, Diggle, and Dobson 2000; Rizopoulos 2012; Tsiatis and Davidian 2004).

Despite their adoption and application, joint models in medical research model checking for these shared parameter joint models is done by separate analysis of residuals for the sub-model and joint model diagnostics are not fully developed. The approach towards joint model diagnostics has been using qualitative graphical displays to isolate departures from model assumptions and influential observations by using summary measures (Dobson and Henderson 2003). This had been in part due to lack of computing power to do a full case deletion diagnostics and re-estimation for each subject which was not practical as observed by Dobson and

Henderson (2003).

There has been studies on model assessment tools using multiple imputation of residuals by Rizopoulos, Verbeke, and Molenberghs (2010). This is motivated by the fact that there is attrition in the longitudinal outcome due to occurrence of an event. This non-random drop out makes the use of standard goodness of fit measures from residuals unattainable because there is no reference distribution available and hence residual plots based on observed data alone can be misleading because they do not have standard properties. The main idea in this study is to augment the observed data with randomly imputed data points that would, otherwise, have been provided, had the subject not dropped out. This creates complete data from which residuals can be calculated and plots produced, this is based on Bayesian techniques. Analysis of theses residuals is again through separate sub-models (Rizopoulos, Verbeke, and Molenberghs 2010).

There has also been a Bayesian approach to assessing influence measures for the joint model by Zhu et al. (2012) motivated by a multi-center clinical trial of breast cancer. This work involves, independently, perturbing the joint model, subjects and random effects for both submodels. These perturbations are done to priors and all components of the Bayesian joint model. The three perturbations are combined to priors which creates a sampling distribution from which data is obtained. These data permit an assessment on simultaneous sensitivity of all components of the Bayesian analysis. The proposed methods are able to detect outliers, influential observations and assess the sensitivity of inference to various unverifiable assumptions on the Bayesian framework. Another Bayesian approach that has been studied using Q-functions for case-deletion is by Tang, Tang, and Zhu (2017), with applications to skew-normal distributions. In this work, a modified likelihood displacement diagnostic statistic is

used. This is because it is difficult to evaluate a likelihood displacement of observed data using Q-functions. A quantity that is a Q-function is calculated which evaluates the effect of the i^{th} subject on the maximum Q-function after excluding the i^{th} subject from the estimation of $\boldsymbol{\theta}$.

An assessment of goodness of fit has been developed for joint models using Akaike information criterion (AIC) and the Bayesian information criterion (BIC) (Baghfalaki, Ganjali, and Verbeke 2017; Park and Qiu 2014; Zhang et al. 2014). The work of Baghfalaki, Ganjali, and Verbeke (2017), mainly focuses on distributional assumptions, which has the potential of misleading inferences if there is heterogeneity of random effects in the shared parameter joint model, while Park and Qiu 2014 focus on separate goodness of fit assessment and diagnostics for the two sub-models. The goodness of fit by the two authors is based on both extensions to AIC and BIC, which accounts for the number of parameters and sample size. On the other hand Zhang et al. (2014) developed a quantifiable change in AIC and BIC for fitting survival data with and without the longitudinal data. This is done by decomposing the AIC and BIC into additive components from which the goodness of fit is derived. Using this approach, it is possible to compare longitudinal outcomes in terms of their contribution to the overall fit of the survival data. Model assessment is done in two parts, based on likelihood theory by using conditional density of the random effects given the longitudinal part, which is replicated in the survival part of the model. This motivates an idea to decompose Cook's statistics for diagnostics of the joint model using both survival and longitudinal sub-models given that parameter estimates from the joint likelihood are asymptotic and efficient (Zeng, Cai, et al. 2005).

There has been little work done on developing diagnostics for the joint model parameters obtained from maximum likelihood estimation on the joint distribution conditional on shared random effects. This is true for such settings as univariate, multivariate and multiple failure

types joint models. This, therefore, means that if there are influential subjects and/or observations in the dataset which is practically feasible, there would be misleading inferences based on the estimated obtained. In the case of clinical trials, where the objective is to evaluate treatment effect, it could misinform researchers about the true effect of an active treatment. We propose to develop case deletion diagnostics for joint models based on the Cook's statistics (Cook and Weisberg 1982). We build on the case deletion diagnostics by Singini, Mwambi, and Gumedze (2018) and extend this to multivariate joint models as well as joint models in competing risk. We begin by evaluating our case deletion diagnostics in simulation studies as outlined in Chapter 5 Section 5.4. We then illustrate these case deletion diagnostics on a dataset on multi-centre trial on tuberculaosis pericarditis (IMPI) using univariate joint model; and further extend to multivariate as well as joint models in competing risk setting. The longitudinal outcomes in the IMPI dataset are square root CD4 cell count (continuous) and ART uptake (binary), while the competing events are constrictive pericarditis and death.

8.2 Results

The analysis of the square root CD4 cell count dataset from the IMPI trial (Mayosi et al. 2014) revealed that the time and prednisolone effects are statistically significant (see table 2.1). There is, also, an observed increase in the level of subjects' CD4 cell count over time and subjects who received ART at each schedule visit have higher CD4 cell count levels relative to those who are not on ART treatment at each schedule visit (see Figures 2.8 and 6.1). The main treatment, prednisolone time interactions effect is not significant. This means that, time does not modify the effect on prednisolone on the subjects.

We conducted simulation studies to evaluated the case deletion diagnostics using four different scenarios. In each scenario we, artificially, inserted outliers to be identified by the diagnostics based on Cook's statistics. Each scenario had the parameters set based mainly on sample size, number of repeated measurements, length of follow-up, measurement error in the longitudinal setting and association parameter in the survival setting (see table 5.1).

The simulation for the joint model dataset was successful after getting guidance and a template of R code from Professor Dimitris Rizopolous in 2018 (see customized version in appendix A.1). The overall setting was meant to mimic a clinical trial and thus, included such variables as sample size, number of planned repeated measurements and maximum follow-up time. The longitudinal data had the following to be set up for simulation; $\boldsymbol{\beta}$ for group effects and σ while the survival part had $\boldsymbol{\gamma}$ for baseline covariate effect, α the association parameter, ϕ the shape parameter from the Weibull distribution and, lastly, the means for the censoring mechanism from the exponential distribution.

We used a scheme that contaminated ten percent of the last data points in each scenario and this is what had been used as a bench mark to evaluate the case-deletion diagnostics. We inserted an artificial outlier in the longitudinal dataset by inflating the variance i.e. σ^2 for the last ten percent of the data points. In the survival data set we changed values of the shape parameter to anything less than one, for early life (survival) and anything greater than one, for late life since the Weibull distribution has "bath tab" properties (see Figure 5.2). In all the scenarios, the case-deletion diagnostics have been able to identify the ten percent contaminated data points, at least, more than 95% of the time (see Figures 5.5 to 5.8), using model (5.10). In scenarios one and four, besides identifying the ten percent contaminated data points the case-deletion diagnostics also identified data point 7 and 42, respectively. This is expected as

the data generation process is stochastic and thus, would generate some influential data points, which could be identified by the case deletion diagnostics.

The focus of analysis for illustration of case deletion diagnostics using the IMPI dataset was on HIV positive subjects, who were presumed to be sicker and were in majority 939 (67%) subjects. There were 713 (76%) subjects who reported no exposure to ART. In the shared parameter joint model we considered 587 (42%) subjects who had, at least, two CD4 cell count measurements, which is a requirement to fit the joint model using the JM Package in R. There were a total of 57 (9.7%) subjects, who experienced events; 27 (4.6%) had constrictive pericarditis, which is the main event of interest in this analysis, 32 (5.5%) died which is a competing event and two (0.3%) experienced both events (PIDs: 64 and 473).

We implemented the developed case deletion diagnostics at subject and observation levels (see Equations (5.11)and (5.22), respectively. This was done for all parameters and further decomposed for fixed effects, variance components and association parameter in the case of the survival sub-model (see Figures 5.9 to 5.20). In all these illustrations of case deletion diagnostics on IMPI data set, be it, at subject or observation level; the overlap for the identified subjects was mostly for all parameters in the joint model and all parameters in the linear mixed effects sub-model (see Figures 5.9 and 5.10 and Section 5.7.1). There was no overlap in the identified subjects in all parameter in the joint model and those identified in the survival sub-model for both fixed effects and association parameter (see Figures 5.9 and 5.11 and Section 5.7.1). We then proceeded to conduct a sensitivity analysis by excluded influential subjects and/or observations to assess the degree of change in parameter estimates. There was about 15% upward change on the effect of prednisolone on the longitudinal sub-model after excluding influential subjects (see table tables 5.3 and 5.4) and this was in both constriction and

death. On the other hand, there were subtle difference in parameter estimates in the survival sub-model, when influential subjects were excluded in the joint model. This, therefore, suggests that, if the analysis can be treated as intent to treat, inferences about the treatment effect would not be different whether there are influential subjects or not. This pattern is expected to be similar if the sensitivity analysis was conducted at observation level.

We have proposed and implemented a diagnostic approach for the two-stage joint model by fitting the variance shift outlier model on the linear mixed model which is a sub-model in the two-stage joint model. This approach has been implemented on the multi-center trial on TB pericaditis. We implemented the VSOM (Gumedze and Chatora 2014; Gumedze et al. 2010), that identifies outlying subjects as well as observations and at the discretion of the researcher, down-weights these outliers. Our implementation of the VSOM was at both the subject and observation level, then later combined as one VSOM whilst taking appropriate action to avoid over fitting, which involved excluding observations that had a subject already identified as outlying. The predicted values from the combined VSOM are used in the second stage of the two-stage joint model as time-varying covariate. In the linear model, there is both improvement in the overall model fit contrasting AICs and BICs statistics after a combined VSOM as well as better estimates of parameters after VSOM (see Table 4.2).

It was observed that, there was no significant difference in parameter estimates when the two-stage joint model is implemented with the combined VSOM. On the other hand, a better overall fit of the model was noted when the combined VSOM had provided fitted values to be used as time-varying covariate in the second stage of the joint model. This is evaluated using AICs and BICs from the model fitted to the data (see Table 4.3 and Figure 4.5).

We ran a multivariate joint survival model for our case deletion diagnostics; the multivariate

joint model had two longitudinal outcomes, but can accommodate more repeated measures outcome. The longitudinal outcomes that were used were CD4 cell count which measured on a continuous scale and ART uptake, which was a binary response. We transformed the CD4 cell count to square root and modeled it under the assumption of normality in the linear model (2.1), while ART was modeled under the binomial assumption using the generalised linear model (6.6).

The estimation of multivariate joint model's parameters was based on Bayesian techniques using Markov Chain Monte Carlo (MCMC); this is an extension of the Gibbs Sampler known as JAGS (see Section 6.2 of Chapter 6). The case deletion diagnostics approach that was used in the multivariate joint survival model was extended from the standard one, which is based on Cook's statistics (model 5.10). We obtained estimates of all parameters $\boldsymbol{\theta}$ from the model that was believed to have converged from the MCMC samples. These MCMC samples were extracted from the last row of the chain. Parameter estimates that were obtained, include variance components from the GLMM in addition to fixed effects and association parameters from the survival sub-model. Evaluation of the Cook's statistics was by excluding the i^{th} subject and/or the ij^{th} observation from the modeling process to get a new set of parameter estimates $\hat{\boldsymbol{\theta}}$ their corresponding variance covariance matrix $\hat{\mathbf{M}}$. We, then, computed the Cook's distances for each subject and/or observation using (6.14) and (6.15), respectively. Influential subjects and observations were those that had relatively large values of Cook's distance, which were plotted in an index plot for each subject and observation (see Figures 6.2 and 6.3). We re-fitted the multivariate joint model, excluding influential subjects at first and later influential observations. We observed subtle changes in treatment effect in either case Tables 6.1 and 6.2. The changes in treatment effect as well as other parameters were not statistically significant

as shown by overlapping 95 % credible intervals for posterior means both at subject and observation level.

It was noted that the association parameter for CD4 cell count, which quantifies how the longitudinal CD4 cell counts values are related to the risk of constriction was statistically significant, while that for ART uptake was not significant (Tables 6.1 and 6.2).

We used the IMPI dataset to illustrate our case deletion diagnostics for joint models in multiple failure types. The two competing events that were used in the evaluation of the diagnostics, which are based on Cook's statistics, were constrictive pericarditis and death. The two competing events were treated as cause-specific in the survival sub-model (Section 7.3 of page 157). We limited the illustration of the diagnostics to CD4 cell count only in the measurement model. We then modeled square root CD4 cell count using a linear mixed effects model. The estimation of the joint model parameters was based on maximum likelihood with a slight modification to the construction of the likelihood on the survival part of the joint model (7.11), which incorporates multiple failure types. We adopted a similar approach by Singini, Mwambi, and Gumedze (2018) to evaluate Cook's distance for the i^{th} subject and/or the ij^{th} observation as in models (7.13) and (7.14). We got estimates of all parameters $\boldsymbol{\theta}$ from the model estimates, including the respective variance covariance matrix $\mathbf{M}$. The calculation of Cook's statistics was done using the standard case deletion approach, which, iteratively, excludes the i^{th} subject and/or the ij^{th} observation from the modeling process to get a new set of parameter estimates $\hat{\boldsymbol{\theta}}$ and their corresponding variance covariance matrix $\hat{\mathbf{M}}$. Outliers or Influential subjects and/or observations were those that had relatively large values of the calculated Cook's distance. We, then, plotted these Cook's distances using an index plot for each subject and observation (see Figures 7.2 and 7.3). We, then, went ahead to conduct a sen-

sitivity analysis by re-running the joint model excluding influential subjects at first and later influential observations. We observed some changes in parameter estimates in the survival component of the joint model with multiple failure types, by contrast, there was negligible sensitivity in the linear mixed effects sub-model at subject level. At the observation level, there were subtle differences in the estimates after re-fitting the model both for the linear and the survival components of the joint model (Tables 6.2 and 7.1). The changes in parameter estimates were not statistically significant as shown by overlapping 95 % confidence intervals at both subject and observation levels. We observed that the association parameter for CD4 cell count, which quantifies how the longitudinal CD4 cell counts values are related to the risk for constriction and death was statistically significant for constriction only (Tables 6.2 and 7.1). This result is consistent with the original results from the main study (Mayosi et al. 2014).

8.3 Remarks on case deletion diagnostics

We observe that case deletion diagnostics based on Cook's statistics (Cook and Weisberg 1982) does not in any way mean that the study design, data collection or primary analysis methods are questionable. In any case, this is one way by which more confidence is placed on the primary analysis conclusions. When inferences drawn from an analysis which has influential subjects and/or observations are comparable to those drawn from an analysis without influential subjects and/or observation, for example, treatment effect in a clinical trial, one would place more confidence in the conclusions as opposed to a case where this was not done (Jansen et al. 2006).

In medical research studies, biomarkers are repeatedly collected over the follow-up time of the

study, while observing the occurrence of and event (survival analysis) such as death, disease relapse. In this case a shared parameter modeling framework called joint models (JM) would be a natural choice for performing a statistical analysis. The advantage of shared parameter joint models is that, if interest is on the longitudinal outcome, they account for non-random drop out dues to the occurrence of an event. If, on the other hand, interest of analysis is on a survival outcome; they account for measurement error. No assumption is made that the biomarker remains constant between two successive measurement and an assessment of surrogacy with its quantification is made between a biomarker and risk of an event (Baghfalaki, Ganjali, and Verbeke 2017; Henderson, Diggle, and Dobson 2000; Hsieh, Tseng, and Wang 2006; Tsiatis and Davidian 2004; Wulfsohn and Tsiatis 1997).

We note that the estimation of shared parameter joint models has been an active research area in the past two and a half decades. However, model assessment for these joint models is done using separate analysis of residuals from the sub-models, which could be sub-optimal. There has been inadequate development in model diagnostic tools to assess influence on parameter estimates or to perform sensitivity analysis (Baghfalaki, Ganjali, and Verbeke 2017; Dobson and Henderson 2003; Rizopoulos, Verbeke, and Molenberghs 2010; Singini, Mwambi, and Gumedze 2018).

There has been low uptake on the use of two-stage joint model, possibly, because of the biased results that the model produces (Dafni and Tsiatis 1998; Sweeting and Thompson 2011; Tsiatis and Davidian 2001). We note that recent methodological developments to correct the bias (Mauff et al. 2018, 2020) would render the two-stage joint model useful and popular. This is the case because the two-stage joint model is relatively easy to implement in standard software and it is not computationally intensive compared to a shared parameter joint model

which requires numerical integration techniques conditional on random effects (Mauff et al. 2018; Self and Pawitan 1992). It, therefore, motivates the idea to identify influential subjects and/or observations, then implement a combined VSOM at the discretion of the researcher at the second stage of the two-stage joint model. This has been seen to improve the model fit based on AIC and BIC. The possible reason for AICs and BICs to be similar in Table 4.3, could be due to the fact that there is no significant association between square root CD4 cell count and the risk for constriction.

We extended the case deletion diagnostics from the univariate joint model to multivariate joint model. In the multivariate setting, there were more than one longitudinal outcome, which could be from any exponential family of distributions. This, therefore, means that, one can formulate the measurement model using the generalised linear mixed model (**??**) given appropriate assumptions. The advantage of multivariate joint model is that, the account for correlation between longitudinal outcomes conditional on random effects (Rizopoulos 2012). We used the same IMPI dataset to implement case deletion diagnostics on the multivariate joint model based on Cook's statistics for all parameters (6.14) (see Figures 6.2 and 6.3). The case deletion diagnostics that we implemented can be extended to evaluate Cook's statistics for fixed effects and variance components of both sub-models. These can, further, be extended on association parameters quantifying CD4 cell count and ART uptake repeated measures to the risk of constriction. We were forced to fit and run the multivariate joint models on University of Cape Town's High Performance Computing (HPC) facilities and further used parallel computing on these facilities due to computational challenges. This reduced computing time by over 80 % (20 minutes to 4 minute) for one multivariate joint model. This limitation led us to evaluating Cook's statistics at subject and observation level for all parameters, however,

extensions to parameter specific diagnostic is straight forward.

Lastly but not least, we illustrated the case deletion diagnostics in competing risk setting, as it is common in time-to-event studies to observe more than one event. We focused on the cause specific competing risk and not the sub-distribution hazard. Our implementation of the case deletion diagnostics in multiple failure types identified influential subjects and observations (see Figures 7.2 and 7.3). We further carried out a sensitivity analysis by excluding influential subjects or observations (see Tables 7.1 and 7.2).

We have developed empirical influence diagnostics for the parameters of the joint model as a composite Cook's statistics, $\hat{\boldsymbol{\theta}}$. The subsets of the parameter space in the shared parameter joint model includes fixed effects from both the longitudinal and survival sub-models, variance components and association parameter represented as $\hat{\boldsymbol{\theta}}$. The focus of this thesis was on the composite Cook's statistics. However, these statistics can be decomposed to components which detect influential observations or subjects for fixed effects, variance components and association parameter of the joint model, all of which are estimated by the conditional joint distribution of the two processes i.e. longitudinal and survival process conditional on random effects.

Estimation of the joint model is done using optimisation algorithms, such as, the EM algorithm. The iterative nature of estimating the joint model makes it difficult to formulate analytical expressions for a Cook's distance as is the case in linear regression, where, a residual, mean square error and other quantities are calculated. The available software that fit the joint model are able to calculate the Cook's statistics based on the covariance matrix.

Implementing case deletion diagnostics can be resource intensive in terms of computing time due to the estimation of the joint model, which employs numerical integration techniques,

since the integral is intractable given the random effects. In our case, we implemented these diagnostics using parallel computing in order to be efficient.

8.4 Thesis contributions

In most clinical trials, the objective is to assess treatment effects and this therefore means estimating the effect, which depends on the data collected (Akacha et al. 2017; Council et al. 2010; Mallinckrodt et al. 2013). This implies that the validity of any statistical methods of analysis depends on study design and data collected from which estimates are obtained. In this thesis, we have considered diagnostics for joint survival models, which can be applied to IMPI trial dataset and similar trials. We have developed case deletion diagnostics for joint models for the IMPI dataset after evaluating them in simulation studies. The work on these case deletion diagnostics have been published (Singini, Mwambi, and Gumedze 2018).

Secondly, we have implemented a combined variance outlier model in a two-stage joint model using IMPI dataset. In this work, we have considered identifying influential subjects and/or observations in the longitudinal outcome and down-weighting them in the second stage of the two-stage joint model. The down-weighting of influential subjects and/or observations has to be sanctioned by the domain researcher. This work has been submitted for possible publication in the journal of Communications in Statistics - Simulation and Computation.

Thirdly, we have implemented case deletion diagnostics in multivariate joint model and joint models with multiple failure types (competing risk). These two settings are important in medical research as they account for correlation in the measurement model as well as assess association between repeated measures of biomarkers with risk of events (Rizopoulos 2012). We

have illustrated these diagnostics using IMPI dataset.

Finally, we have proposed extensions and R code for these case deletion diagnostics at various levels, for instance, fixed effects, variance components and association parameter depending on the investigator's interest.

8.5 Conclusions

In longitudinal studies, especially, in medical research, the response of interest could be a time-to-event outcome in the presence of a time varying endogenous time-varying covariate. The appropriate analysis method in this case is the shared parameter joint model, which adjusts for measurement error in the time-varying covariate and quantifies the association between the measurement model and risk of an event (Crowther, Abrams, and Lambert 2012; Henderson, Diggle, and Dobson 2000; Rizopoulos 2012; Tsiatis, Degruttola, and Wulfsohn 1995; Wulfsohn and Tsiatis 1997). Parameter estimates obtained from these models are unbiased compared to the two-stage approach (Dafni and Tsiatis 1998; Sweeting and Thompson 2011; Tsiatis and Davidian 2001). However, there is inadequate literature on model assessment based on the full conditional joint distribution, just as there is lack of influence diagnostics on the full conditional distribution of the joint model. In clinical trials, influential subjects and/or observations can over or under estimate treatment effects (Pregibon 1981). In order to draw valid inferences in the presence of influential subjects and/or observations using the shared parameter joint modeling framework, we have developed and implemented case deletion diagnostics on the full conditional joint distribution. These case deletion diagnostics were illustrated on the IMPI dataset with square root CD4 cell count and ART uptake as time-

varying covariates while constriction and death as survival outcomes (Mayosi et al. 2014).

This is the first study to develop and implement case deletion diagnostics used in joint models for longitudinal and survival data based on a full conditional joint distribution. The estimation of joint models has been an active research area in the past two to three decades, however, models assessment and diagnostics have been lacking except for separate analysis of residuals for the sub-models (Rizopoulos, Verbeke, and Molenberghs 2010). Development of full case deletion diagnostics was even difficult in the past because of computing power and unavailability of software (Dobson and Henderson 2003). Computer processing power has significantly improved in recent years (Dufour et al. 2012) and software is readily and freely available to implement full case deletion diagnostics on the joint model, for instance the JM Package by Rizopoulos (2010) in R.

In this thesis, we have endeavoured to develop case deletion diagnostics for the shared parameter joint based on Cook's statistics (Cook and Weisberg 1982). We have justified the theory for these diagnostics on the basis of work done by Zeng, Cai, et al. (2005) and made extensions to multivariate and multiple failure types in the joint model framework. We have also implemented a VSOM in the two-stage joint model, which is deemed to be easy to implement (Self and Pawitan 1992; Tsiatis, Degruttola, and Wulfsohn 1995), especially, in multivariate setting (Mauff et al. 2018, 2020). We are wary of the fact that our case study could not be generic for all clinical trial designs including the software being used to implement our developed case deletion diagnostics because joint model assessment and diagnostics are evolving. The main objective of this thesis is to develop diagnostics for the full conditional joint distribution for the joint model; evaluate these diagnostics in simulation studies and illustrate them using IMPI trial dataset.

We encourage researchers to be mindful when selecting a primary model and then consider performing a sensitivity analysis as studied by Mallinckrodt et al. (2013). In the case of influence diagnostics for a joint model, run the joint model on the full dataset, then implement case deletion diagnostics to identify influential subjects and/or observations and later run the same model excluding influential subjects and/or observation and compare parameter estimates. In this way, the researcher will be able to identify and initiate investigations on these influential subjects and/or observation to ascertain whether they are due to data entry errors or they are just influential or outlying. We further recommend that researchers make an effort to understand software implementation for case deletion diagnostics with the corresponding out-put from the joint model in order for researchers to be able to extract requisite components to be used for calculating Cook's statistics. We further note that, handling influential data points depends on many considerations as described by Welsch (1980). In any case the identification of these data points is valuable in that it demonstrates the instability of statistical inferences regarding the effect of regression coefficients, hence risk factors upon prognosis.

We have demonstrated that case deletion diagnostics in joint models can reveal influential subjects and/or observations. These influential subjects and/or observations have different degrees to which the affect parameter estimates obtained from the joint model. In most cases, the parameter estimates that were sensitive to influential subjects were from the longitudinal component of the joint model. We observed subtle changes in parameter estimates from the survival component except in Chapter 5 Section 5.6, where there is about 10% difference in the parameter estimate for the treatment effect. This justifies the importance of performing case deletion diagnostics for joint models especially in clinical trials.

In as much as we have highlighted the importance of case deletion diagnostics for joint models

with extensions to different settings, this thesis, also stresses on the practicality and importance of implementing these diagnostics, for instance in multivariate setting where there are evident computational challenges due to the dimensions of random effects (see Chapter 6).

The trial dataset that we illustrated case deletion diagnostics on, was from a cardiology multi-centre clinical trial, which also collected HIV related data (CD4 cell count and ART uptake). The HIV data was not routinely collected as would have been if the focus was on HIV and hence has missing CD4 cell count and ART uptake. Our analysis of HIV data in the IMPI trial, despite it being a cardiology study, provided useful information concerning the effect of prednisolone on CD4 cell count and ART uptake trajectories overtime. The missingness mechanisms of both CD4 cell count and ART uptake, have underlying assumptions that can be investigated using sensitivity analysis (Iddrisu and Gumedze 2019; Mallinckrodt et al. 2013). The missingness mechanisms for both CD4 cell count and ART uptake are beyond the scope of this thesis.

Carefully conducting model assessment and implementing of case deletion in a joint model diagnostics provides useful tools to assess different aspects of the model on the full conditional joint distribution. In the case of influence diagnostics, an investigation can be instituted on influential subjects and/or observation to reveal underlying characteristics of these subjects and/or observations. Once the underlying characteristics are revealed, knowledge could be gained by domain experts on the possible causes and further, the effect this has on parameter estimates.

In the IMPI trial most subject identified as influential had low CD4 cell count, reported minimal or no exposure to ART and neither experienced the events of interest. It would, therefore, seem plausible to report these subjects to domain experts (cardiologists) for further clinical

investigations.

Analysts conducting case deletion diagnostics do not necessarily have to perform these diagnostics on all parameters individually, i.e. fixed effects, variance components or association parameter but on clinically/ study relevant parameter(s).

8.6 Further research

The shared parameter joint model considered in this study is mainly used to assess the strength of association between a longitudinal biomarker and risk of an event (Baghfalaki, Ganjali, and Verbeke 2017; Diggle 2002; Henderson, Diggle, and Dobson 2000; Rizopoulos 2012; Tsiatis, Degruttola, and Wulfsohn 1995). The diagnostics developed in this joint model may further be investigated in other joint model formulations e.g. latent class joint model (Proust-Lima et al. 2009, 2014).

In this thesis, we have used the current value parameterisation of the biomarker, it would be important to investigate other parameterisations to assess the performance of the developed diagnostics. These parameterisations are: interaction effects, lagged effects, time-dependent slope, cumulative effects and random effects (Rizopoulos 2012).

The joint model diagnostics based on Cook's statistics identifies influential subjects and/or observations. A sensitivity analysis helps to evaluate the level of effect these influential data points have on parameter estimates. An implementation of a VSOM with the guidance of the domain researcher could be worthwhile, to investigate its performance on the measurement sub-model with a possibility of introducing a frailty term on the survival sub-model.

Appendices

A.1 Joint model simulation data sets R code

```
library(MASS)
library(splines)

n <- 100     # number of subjects
K <- 5       # number of planned repeated measurements per subject,
 per outcome
t.max <- 10  # maximum follow-up time

################################################

# parameters for the linear mixed effects model
betas <- c("Group0" = 3.3493, "Group1" = 2.8444,
"Group0:Time1" = 1.5125,
"Group1:Time1" = 1.2533, "Group0:Time2" = 2.4739,
"Group1:Time2" = 2.2994,
"Group0:Time3" = 2.5379, "Group1:Time3" = 1.8489)
sigma.y <- 0.8034 # measurement error standard deviation

# parameters for the survival model
gammas <- c("(Intercept)" = -5.7296, "Group" = 0.4092)
# coefficients for baseline covariates
```

```
alpha <- 0.3917 # association parameter
phi <- 1.6458 # shape for the Weibull baseline hazard
mean.Cens <- 12 # mean of the exponential distribution
for the censoring mechanism

D <- diag(c(0.6968, 2.1259, 1.5260, 1.2329)^2)

################################################

Bkn <- c(0, 19.5)
kn <- c(2.1, 5.5)

# design matrices for the longitudinal measurement model
# but this can be easily generalized
times <- c(replicate(n, c(0, sort(runif(K-1, 0, t.max)))))
# at which time points longitudinal measurements are
supposed to be taken

group <- rep(0:1, each = n/2)
# group indicator, i.e., '0' placebo, '1' active treatment

DF <- data.frame(year = times, drug = factor(rep(group, each = K)))
```

```
X <- model.matrix(~ 0 + drug + drug:ns(year, knots = kn,
Boundary.knots = Bkn), data = DF)

Z <- model.matrix(~ ns(year, knots = kn,
Boundary.knots = Bkn), data = DF)

# design matrix for the survival model
W <- cbind("(Intercept)" = 1, "Group" = group)

################################################

#simulate random effects
b <- mvrnorm(n, rep(0, nrow(D)), D)

# simulate longitudinal responses
id <- rep(1:n, each = K)

# linear predictor
eta.y <- as.vector(X %*% betas + rowSums(Z * b[id, ]))
y1 <- rnorm((n-5) * K, eta.y, sigma.y)

# Inserting five outliers in the longitudinal dataset
```

```
y2 <- rnorm(5 * K, eta.y, (2*sigma.y))

y <- c(y1,y2)

# simulate event times
eta.t <- as.vector(W %*% gammas)
invS <- function (t, u, i) {
h <- function (s) {
group0 <- 1 - group[i]
group1 <- group[i]
NS <- ns(s, knots = kn, Boundary.knots = Bkn)
XX <- cbind(group0, group1, group0*NS[, 1],
group1*NS[, 1],

group0*NS[, 2], group1*NS[, 2], group0*NS[, 3],
group1*NS[, 3])
ZZ <- cbind(1, NS)
f1 <- as.vector(XX %*% betas + rowSums(ZZ * b[rep(i, nrow(ZZ)), ]))
exp(log(phi) + (phi - 1) * log(s) + eta.t[i] + f1 * alpha)
}
integrate(h, lower = 0, upper = t)$value + log(u)
}
u <- runif(n)
```

```
trueTimes <- numeric(n)
for (i in 1:n) {
Up <- 50
tries <- 5
Root <- try(uniroot(invS, interval = c(1e-05, Up),
u = u[i], i = i)$root, TRUE)

while(inherits(Root, "try-error") && tries > 0) {
tries <- tries - 1
Up <- Up + 200
Root <- try(uniroot(invS, interval = c(1e-05, Up),
u = u[i], i = i)$root, TRUE)
}
trueTimes[i] <- if (!inherits(Root, "try-error")) Root else NA
}
na.ind <- !is.na(trueTimes)
trueTimes <- trueTimes[na.ind]
W <- W[na.ind, , drop = FALSE]
long.na.ind <- rep(na.ind, each = K)
y <- y[long.na.ind]
X <- X[long.na.ind, , drop = FALSE]
Z <- Z[long.na.ind, , drop = FALSE]
DF <- DF[long.na.ind, ]
```

```
n <- length(trueTimes)

# simulate censoring times from an exponential distribution,
# and calculate the observed event times, i.e.,
min(true event times, censoring times)
Ctimes <- runif(n, 0, 2 * mean.Cens)
Time <- pmin(trueTimes, Ctimes)
event <- as.numeric(trueTimes <= Ctimes) # event indicator

################################################

# keep the nonmissing cases, i.e., drop the longitudinal measurements
# that were taken after the observed event time for each subject.
ind <- times[long.na.ind] <= rep(Time, each = K)
y <- y[ind]
X <- X[ind, , drop = FALSE]
Z <- Z[ind, , drop = FALSE]
id <- id[long.na.ind][ind]
id <- match(id, unique(id))

dat <- DF[ind, ]
dat$id <- id
dat$y <- y
```

```
dat$Time <- Time[id]
dat$event <- event[id]
dat.id <- data.frame(Time = Time, event = event, group = W[, 2])
names(dat) <- c("time", "group", "id", "y", "Time", "event")

#summary(tapply(id, id, length))
#table(event)
#n
#mean(event)



# delete all unused objects
rm(y, X, Z, id, n, na.ind, long.na.ind, ind, Ctimes, Time, event, W,
betas, sigma.y, gammas, alpha, eta.t, eta.y, phi, mean.Cens, t.max,
trueTimes, u, Root, invS, D, b, K,
times, group, i, tries, Up, Bkn, kn, DF)
```

A.2 Summary statistics for variables used in the joint model

Table A.1: Summary statistics for candidate variables in the joint model for longitudinal and survival data.

Covariate	Category/Units	no. data (%) [b]	Prednisolone [a] (N=293)	no. data (%) [b]	Placebo [a] (N=294)
Age —mean (sd.)	years	293 (100%)	36 (9.4)	294 (100%)	34.6 (9.0)
Weight —mean (sd.)	kg	283 (96.6 %)	58.8 (1.3)	288 (98 %)	59.0 (1.2)
NYACLASS —no. (%)		293 (100 %)		294 (100%)	
	I-II		191 (65.2 %)		199 (67.7 %)
	III-IV		102 (34.8 %)		95 (32.3 %)
Creatinine —no. (%)		285 (97.3 %)		284 (96.6 %)	
	$\leq$105		250 (85.3 %)		252 (85.7 %)
	>105		35 (12.0 %)		32 (10.8 %)
Peripheral Oedema —no. (%)		293 (100 %)		294 (100%)	
	Yes		96 (32.8 %)		101 (34.4 %)
	No		197 (67.2 %)		193 (65.7 %)
Sex —no. (%)		293 (100 %)		294 (100 %)	
	Male		159 (54.3 %)		172 (58.5 %)
	Female		134 (45.7 %)		122 (41.5 %)
TB Status —no. (%)		293 (100 %)		294 (100 %)	
	Definite TB		64 (21.8 %)		65 (22.1 %)
	Probable TB		229 (78.2 %)		229 (77.9 %)

[a] Active treatment and Placebo arm
[b] Available data point for model fitting
[c] Subjects with at least two CD4 measurements were 587 and all were HIV positive

A.3 Influential subjects

The following tables present influential subjects identified by case deletion diagnostics based on Cook's distance.

Table A.2: Joint model vs Linear mixed model.

Linear mixed model	Joint model	
	Yes	No
Yes	174, 277, 511, 557	246
No	424	0

Table A.3: Joint model vs Two-stage joint model.

Two-stage joint model	Joint model	
	Yes	No
Yes	0	316, 320, 468, 506, 571
No	174, 246, 277, 511, 577	0

Table A.4: Joint model vs Standard Cox model.

Standard Cox model	Joint model	
	Yes	No
Yes	0	23, 205, 291, 316, 320, 468, 473, 474, 486, 578
No	174, 246, 277, 511, 577	0

Table A.5: Two-stage joint model vs Standard Cox model.

Standard Cox model	Two-stage joint model	
	Yes	No
Yes	316, 320, 468	23, 205, 291, 473, 474, 486, 578
No	506, 571	0

A.4 Cook's distance decomposition

In this Section we decompose the Cook's distance for the linear mixed model as a build up from the linear regression model. Using the following formulation for the linear regression model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon_{\mathbf{i}}. \tag{A.1}$$

The case-deletion diagnostics based on the Cook's distance for the simple linear regression model is presented as

$$\mathbf{C}^*_{\mathbf{k}} = \hat{\beta} - \hat{\beta}_{\mathbf{i}} = \frac{\mathbf{r_i}}{1-h_i}(X'X)^{-1}x_i', \tag{A.2}$$

where $\hat{\beta}$ is the ordinary least squares estimate for the full dataset, $\hat{\beta}_{\mathbf{i}}$ is the estimate after excluding the i^{th} case, $\mathbf{x_i}$ is the i^{th} row vector of $\mathbf{X}$, $\mathbf{r_i}$ is the residual and $\mathbf{h_i}$ is the leverage.

Leverage (h) is calculated as the partial derivative of the predicted value with respect to the corresponding dependent variable. This, therefore, means that the i^{th} leverage indicates how the predicted value of the i^{th} case is influenced by the i^{th} observation. For the simple linear regression model (A.1) leverage can be show to be

$$h_i = \frac{\partial \hat{y}_i}{\partial y_i} = \mathbf{X_i'}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{X_i}, \tag{A.3}$$

where $\hat{y} = \mathbf{X}\hat{\beta}$ and $\mathbf{X_i'}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{X_i}$ is commonly known as the hat matrix.

The influence of a case i as proposed by Cook (1977) is measured by the squared distance between $\hat{\beta}$ and $\hat{\beta}_{\mathbf{i}}$ scaled by the variance presented as;

$$M_i^2 = \frac{1}{m\sigma^2}(\hat{\beta} - \hat{\beta}_{\mathbf{i}})'\mathbf{M}(\hat{\beta} - \hat{\beta}_{\mathbf{i}}) = \frac{\mathbf{1}}{\mathbf{m\sigma^2}}\frac{\mathbf{r_i^2}}{\mathbf{1-h_i^2}}\mathbf{X_i}\mathbf{M^{-1}}\mathbf{X_i'} \tag{A.4}$$

A decomposition of the Cook's distance is made to the following linear mixed model condi-

tioned on random effects;

$$\begin{cases} \mathbf{Y} = \mathbf{X}\beta + \mathbf{Zb} + \epsilon_i, \\ \mathbf{b} \sim \mathbf{N}(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{D}), \qquad \epsilon \sim \mathbf{N}(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{R}). \end{cases} \tag{A.5}$$

The assumptions about the random effects and error terms suggest that the variance covariance matrix of $\mathbf{Y}$ can be written as $\sigma^2\mathbf{V_i}$ where $\mathbf{V_i} = \mathbf{R} + \mathbf{Z_i}\mathbf{D}\mathbf{Z_i'}$. Under independence and normality assumptions of the random and error terms and if $\mathbf{D}$ is known the estimates of the fixed effects parameters can be obtained using generalized least squares as follows:

$$\hat{\beta} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{X_i'}\mathbf{V_i^{-1}}\mathbf{X_i} \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{X_i'}\mathbf{V_i^{-1}}\mathbf{y_i} \equiv \mathbf{M^{-1}}\mathbf{S}, \tag{A.6}$$

where $\mathbf{M} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{X_i'}\mathbf{V_i^{-1}}\mathbf{X_i}$ and $\mathbf{S} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{X_i'}\mathbf{V_i^{-1}}\mathbf{y_i}$ although estimation of β and $\mathbf{D}$ in linear mixed models is by maximum likelihood or restricted maximum likelihood.

Analogous to the linear regression (A.3) leverage matrix, Demidenko (2013) defined the $n_i \times n_i$ leverage matrix for the linear mixed model framework as follows:

$$\mathbf{H_i} = \frac{\partial \hat{\mathbf{y}}_\mathbf{i}}{\partial \mathbf{y_i}},$$

where $\hat{\mathbf{y}} = \mathbf{X_i}\hat{\beta} + \mathbf{Z_i}\hat{\mathbf{b}}_\mathbf{i}$ which is the predicted outcome conditional on the estimated random effects, $\hat{\mathbf{b}}_\mathbf{i}$. Assuming that $\mathbf{V_i}$ is fixed and after differentiating; the leverage matrix can be presented as the sum of three components as follows,

$$\begin{cases} \mathbf{H_i} = \mathbf{H_{i1}} + \mathbf{H_{i2}} + \mathbf{H_{i3}}, \\ \mathbf{H_{i1}} = \mathbf{X_i}\mathbf{M^{-1}}\mathbf{X_i'}\mathbf{V_i^{-1}}, \\ \mathbf{H_{i2}} = \mathbf{Z_i}\mathbf{D}\mathbf{Z_i'}\mathbf{V_i^{-1}}(\mathbf{I} - \mathbf{H_{i1}}) \qquad \text{and} \\ \mathbf{H_{i3}} = \mathbf{2}\mathbf{X_i}\mathbf{Z_i}\mathbf{D}\mathbf{Z_i'}\mathbf{V_i^{-1}}(\mathbf{I} - \mathbf{H_{i1}}) \end{cases} \tag{A.7}$$

In this framework, $\mathbf{H_{i1}}$ is the leverage for the fixed effects, $\mathbf{H_{i2}}$ is the leverage associated with the random effects and $\mathbf{H_{i3}}$ is the random effects variance, which is a measure of covariation between the a change in the average profile and a change in position of subject-specific profiles

relative to the average profiles.

The compact form of the above decomposition is presented as follows and available in most standard software.

$$CD_i(\theta) = (\hat{\theta}' - \hat{\theta}_{[-i]})\hat{\mathbf{M}}^{-1}(\hat{\theta} - \hat{\theta}_{[-i]}), \tag{A.8}$$

where θ is the parameter space and $\mathbf{M}$ is the variance covariance matrix with corresponding leverage.

A.5 Profiles of influential some selected subjects

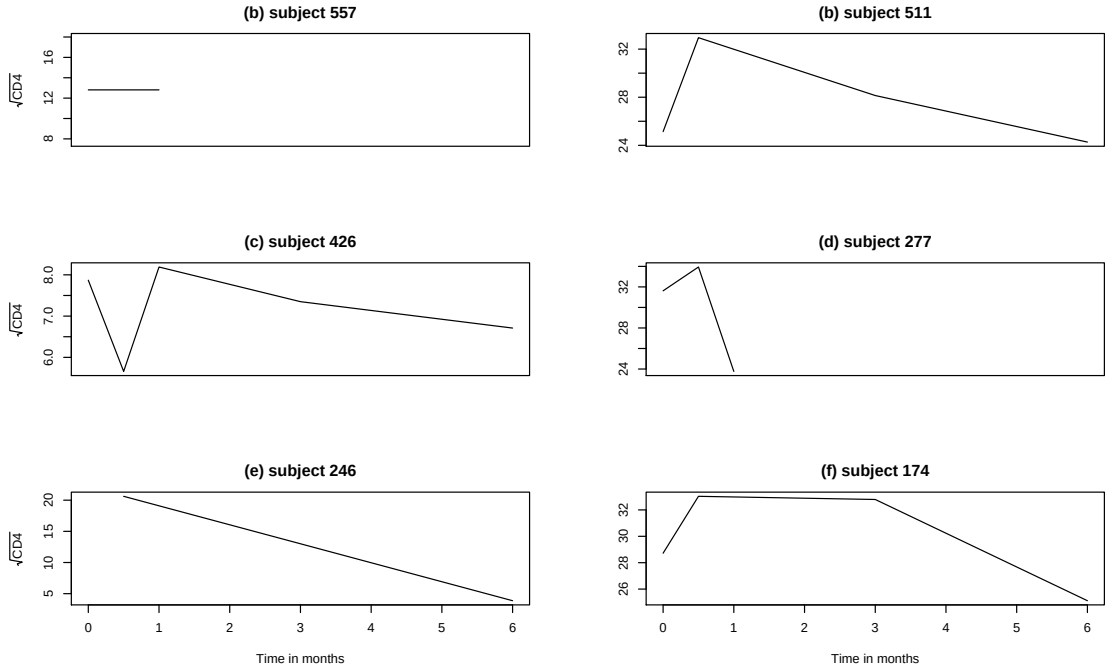


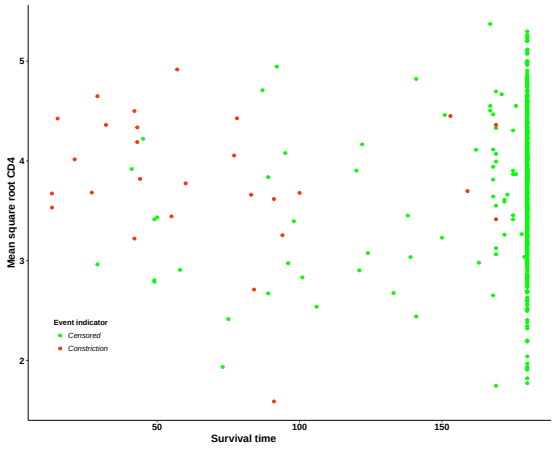
Figure A.5: CD4 cell count profiles for influential subjects

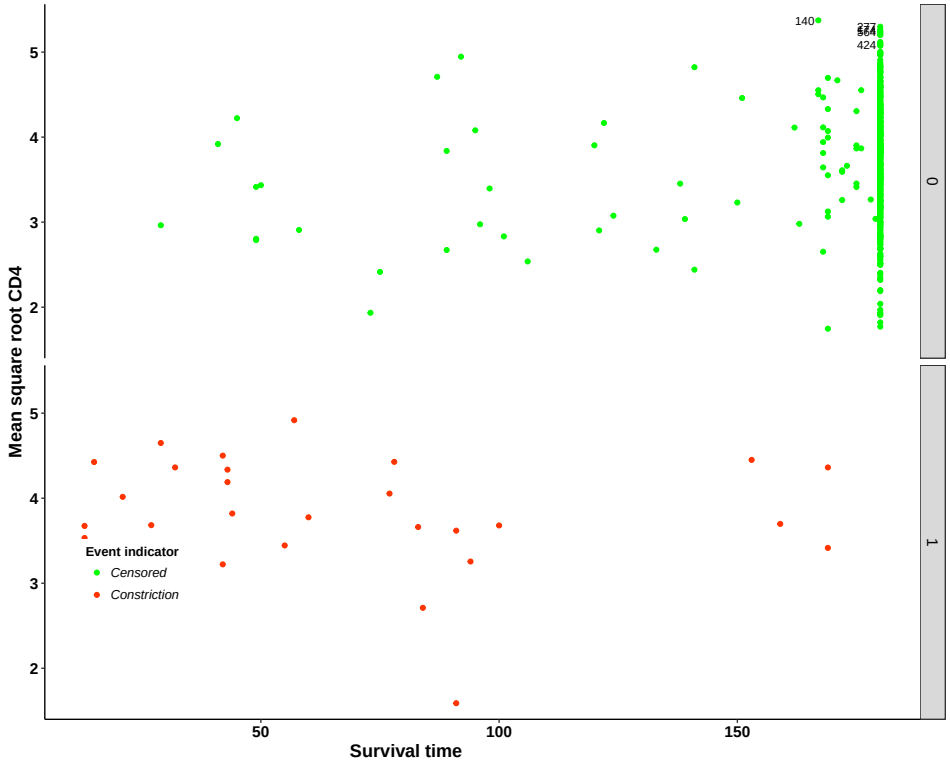
A.6 Distribution of mean square root CD4 over survival time

The survival times for subjects considered for death are displayed in the following figures. It is noted that some subjects with relatively high mean square root CD4 cell count had experienced an event earlier in the study i.e. less than one month. There are also subjects who had low mean square root CD4 cell count who had experienced an event late into the study and most of them were censored at the end of the study.

.6 Distribution of mean square root CD4 over survival time

The survival times for subjects considered for death are displayed in the following figures. It is noted that some subjects with relatively high mean square root CD4 cell count had experienced an event earlier in the study i.e. less than one month. There are also subjects who had low mean square root CD4 cell count who had experienced an event late into the study and most of them were censored at the end of the study.

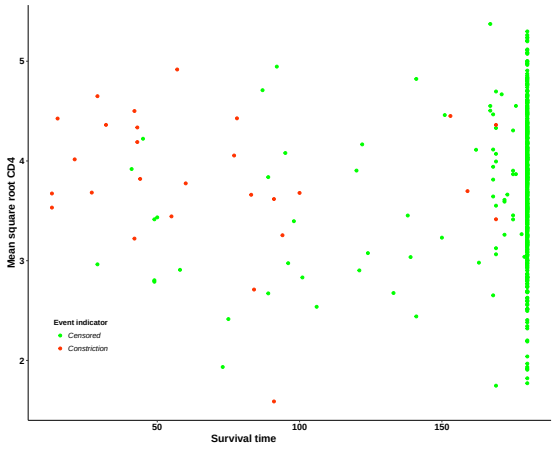


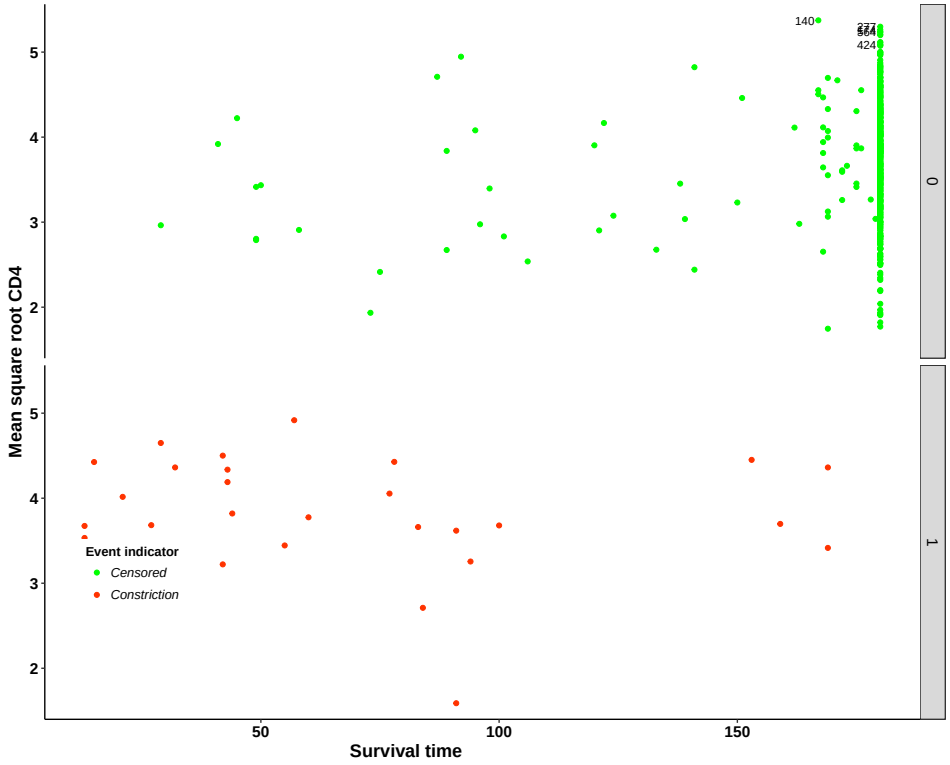


The survival times for subjects considered for constriction are displayed in the following figures. It is noted that some subjects with relatively high mean square root CD4 cell count had experienced an event earlier in the study i.e. less than one month. There are also subjects who had low mean square root CD4 cell count who had experienced an event late into the study and most of them were censored at the end of the study.

.6 Distribution of mean square root CD4 over survival time

The survival times for subjects considered for death are displayed in the following figures. It is noted that some subjects with relatively high mean square root CD4 cell count had experienced an event earlier in the study i.e. less than one month. There are also subjects who had low mean square root CD4 cell count who had experienced an event late into the study and most of them were censored at the end of the study.




Mean square root CD4
Survival time
Event indicator
Censored
Constriction
0
1
140
424

References

Akacha, Mouna et al. (2017). "Estimands and their role in clinical trials". In: *Statistics in Biopharmaceutical Research* 9.3, pp. 268–271.

Akaike, Hirotugu (1974). "A new look at the statistical model identification". In: *IEEE transactions on automatic control* 19.6, pp. 716–723.

Andersen Per Kragh, Gill and Richard D (1982). "Cox's regression model for counting processes: a large sample study". In: *The annals of statistics*, pp. 1100–1120.

Anderson, Richard Loree and Theodore Alfonso Bancroft (1952). *Statistical theory in research*. McGraw-Hill Book Company, Inc.; New York.

Atkinson, Anthony Curtis (1985). *Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis*. Clarendon Press Oxford.

Baghfalaki, Taban, Mojtaba Ganjali, and Geert Verbeke (2017). "A shared parameter model of longitudinal measurements and survival time with heterogeneous random-effects distribution". In: *Journal of Applied Statistics* 44.15, pp. 2813–2836.

Banerjee, Mousumi and Edward W Frees (1997). "Influence diagnostics for linear longitudinal models". In: *Journal of the American Statistical Association* 92.439, pp. 999–1005.

Barnett, V and T Lewis (1995). *Outliers in statistical data*.

Belsley, David A, Edwin Kuh, and Roy E Welsch (1980). "Wiley Series in Probability and Statistics". In: *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, pp. 293–300.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1, pp. 289–300.

Bernoulli, Daniel (1760). "Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir". In: *Histoire de l'Acad., Roy. Sci.(Paris) avec Mem*, pp. 1–45.

Bycott, Paul W and Jeremy MG Taylor (1998). "An evaluation of a measure of the proportion of the treatment effect explained by a surrogate marker". In: *Controlled clinical trials* 19.6, pp. 555–568.

Cain, Kevin C and Nicholas T Lange (1984). "Approximate case influence for the proportional hazards regression model with censored data". In: *Biometrics*, pp. 493–499.

Chatterjee, Samprit and Ali S Hadi (1988). "Impact of simultaneous omission of a variable and an observation on a linear regression equation". In: *Computational Statistics & Data Analysis* 6.2, pp. 129–144.

Chi, Yueh-Yun and Joseph G Ibrahim (2006). "Joint models for multivariate longitudinal and multivariate survival data". In: *Biometrics* 62.2, pp. 432–445.

Cho, Hyunsoon et al. (2009). "Bayesian case influence diagnostics for survival models". In: *Biometrics* 65.1, pp. 116–124.

Christensen, Ronald, Larry M Pearson, and Wesley Johnson (1992). "Case-deletion diagnostics for mixed models". In: *Technometrics* 34.1, pp. 38–45.

Collett, David (2015). *Modelling survival data in medical research*. CRC press.

Cook, R Dennis (1977). "Detection of influential observation in linear regression". In: *Technometrics* 19.1, pp. 15–18.

Cook, R Dennis (1987). “Influence assessment”. In: *Journal of Applied Statistics* 14.2, pp. 117–131.

Cook, R Dennis and Sanford Weisberg (1982). *Residuals and influence in regression*. New York: Chapman and Hall.

Council, National Research et al. (2010). *The prevention and treatment of missing data in clinical trials*. National Academies Press.

Cox, David R (1972a). “Regression models and life-tables”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2, pp. 187–202.

Cox, DR (1972b). “The statistical analysis of dependencies in point processes”. In: *Stochastic point processes*, pp. 55–66.

Cox, DR and DV Hinkley (1974). *D. V. HINKLEY, 1974 Theoretical Statistics*.

Crowther, Michael J, Keith R Abrams, and Paul C Lambert (2012). “Flexible parametric joint modelling of longitudinal and survival data”. In: *Statistics in Medicine* 31.30, pp. 4456–4471.

Dafni, Urania G and Anastasios A Tsiatis (1998). “Evaluating surrogate markers of clinical outcome when measured with error”. In: *Biometrics*, pp. 1445–1462.

Davison, AC, Gigli, and A (1989). “Deviance residuals and normal scores plots”. In: *Biometrika*, pp. 211–221.

Demidenko, Eugene (2013). *Mixed models: theory and applications with R*. John Wiley & Sons.

Demidenko, Eugene and Therese A Stukel (2005). “Influence analysis for linear mixed-effects models”. In: *Statistics in medicine* 24.6, pp. 893–909.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38.

Diggle, Peter (2002). *Analysis of longitudinal data*. Oxford University Press.

Diggle, Peter, Kung-Yee Liang, and Scott L Zeger (1994). “Longitudinal data analysis”. In: *New York: Oxford University Press* 5, p. 13.

Dobson, Angela and Robin Henderson (2003). “Diagnostics for joint longitudinal and dropout time modeling”. In: *Biometrics* 59.4, pp. 741–751.

Dufour, Christian et al. (2012). “Power system simulation algorithms for parallel computer architectures”. In: *2012 IEEE Power and Energy Society General Meeting*. IEEE, pp. 1–6.

Efron, Bradley (1981). “Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods”. In: *Biometrika* 68.3, pp. 589–599.

Escobar, Luis A and William Q Meeker Jr (1992). “Assessing influence in regression analysis with censored data”. In: *Biometrics*, pp. 507–528.

Faucett, Cheryl L and Duncan C Thomas (1996). “Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach”. In: *Statistics in medicine* 15.15, pp. 1663–1685.

Fitzmaurice, Garrett M, Nan M Laird, and James H Ware (2012). *Applied longitudinal analysis*. Vol. 998. John Wiley & Sons.

Fleming, T and D Harrington (1991). *Counting Processes and Survival Models*.

Gad, Ahmed M. and Niveen I. EL-Zayat (2018). “Fitting Multivariate Linear Mixed Model for Multiple Outcomes Longitudinal Data with Non-ignorable Dropout”. In: *International Journal of Probability and Statistics* 7.4, pp. 97–105.

Gail, Mitchell (1975). “A review and critique of some models used in competing risk analysis”. In: *Biometrics*, pp. 209–222.

Gill, Richard D, Niels Keiding, and PK Andersen (1993). *Statistical models based on counting processes*.

Gilmour, Arthur R, Robin Thompson, and Brian R Cullis (1995). “Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models”. In: *Biometrics*, pp. 1440–1450.

Gumedze, Freedom N and Tinashe D Chatora (2014). “Detection of outliers in longitudinal count data via overdispersion”. In: *Computational Statistics & Data Analysis* 79, pp. 192–202.

Gumedze, Freedom N et al. (2010). “A variance shift model for detection of outliers in the linear mixed model”. In: *Computational Statistics & Data Analysis* 54.9, pp. 2128–2144.

Hall, Gaineford J, William H Rogers, and Daryl Pregibon (1982). “Outliers matter in survival analysis”. In:

Hartley, Herman O and Jon NK Rao (1967). “Maximum-likelihood estimation for the mixed analysis of variance model”. In: *Biometrika* 54.1-2, pp. 93–108.

Harville, David A (1974). “Maximum likelihood approaches to variance component estimation and to related problems”. In: *Journal of the American Statistical Association* 72.358, pp. 320–338.

Henderson, Charles R (1950). “Estimation of genetic parameters”. In: *Biometrics*. Vol. 6. 2. INTERNATIONAL BIOMETRIC SOC 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210, pp. 186–187.

Henderson, Charles R (1953). “Estimation of variance and covariance components”. In: *Biometrics* 9.2, pp. 226–252.

Henderson, Charles R et al. (1959). “The estimation of environmental and genetic trends from records subject to culling”. In: *Biometrics* 15.2, pp. 192–218.

Henderson, Robin, Peter Diggle, and Angela Dobson (2000). “Joint modelling of longitudinal measurements and event time data”. In: *Biostatistics* 1.4, pp. 465–480.

Hippisley-Cox, Julia et al. (2007). “Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study”. In: *Bmj* 335.7611, p. 136.

Hosmer, David W, Borko Jovanovic, and Stanley Lemeshow (1989). “Best subsets logistic regression”. In: *Biometrics*, pp. 1265–1270.

Hsieh, Fushing, Yi-Kuan Tseng, and Jane-Ling Wang (2006). “Joint modeling of survival and longitudinal data: likelihood approach revisited”. In: *Biometrics* 62.4, pp. 1037–1043.

Iddrisu, Abdul-Karim and Freedom Gumedze (2019). “Application of sensitivity analysis to incomplete longitudinal CD4 count data”. In: *Journal of Applied Statistics* 46.4, pp. 754–769.

Jansen, Ivy et al. (2006). “The nature of sensitivity in monotone missing not at random models”. In: *Computational statistics & data analysis* 50.3, pp. 830–858.

Jennrich, Robert I and PF Sampson (1976). “Newton-Raphson and related algorithms for maximum likelihood variance component estimation”. In: *Technometrics* 18.1, pp. 11–17.

Khatri, Chinubhai G (1966). “A note on a MANOVA model applied to problems in growth curve”. In: *Annals of the Institute of Statistical Mathematics* 18.1, p. 75.

Laird, Nan M and James H Ware (1982). “Random-effects models for longitudinal data”. In: *Biometrics*, pp. 963–974.

Lange, Kenneth (2004). “Computational statistics and optimization theory at UCLA”. In: *The American Statistician* 58.1, pp. 9–11.

Lange, Kenneth and David R Hunter (2004). “A tutorial on MM algorithms”. In: *The American Statistician* 58.1, pp. 30–37.

Lee, Sang Hong and Julius HJ van Der Werf (2006). “An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree”. In: *Genetics Selection Evolution* 38.1, p. 25.

Lee, Youngjo, John A Nelder, and Yudi Pawitan (2006). *Generalized linear models with random effects: unified analysis via H-likelihood*. CRC Press.

Lesaffre, Emmanuel and Geert Verbeke (1998). “Local influence in linear mixed models”. In: *Biometrics*, pp. 570–582.

Lin, CY and AJ McAllister (1984). “Monte Carlo Comparison of Four Methods for Estimation of Genetic Parameters in the Univariate Case1”. In: *Journal of Dairy Science* 67.10, pp. 2389–2398.

Lin, Min and Rongling Wu (2006). “A joint model for nonparametric functional mapping of longitudinal trajectory and time-to-event”. In: *BMC bioinformatics* 7.1, p. 138.

Lindstrom, Mary J and Douglas M Bates (1988). “Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data”. In: *Journal of the American Statistical Association* 83.404, pp. 1014–1022.

Mallinckrodt, Craig et al. (2013). “Missing data: turning guidance into action”. In: *Statistics in Biopharmaceutical Research* 5.4, pp. 369–382.

Mauff, Katya et al. (2018). “Joint Models with Multiple Longitudinal Outcomes and a Time-to-Event Outcome”. In: *arXiv preprint arXiv:1808.07719*.

– (2020). “Joint models with multiple longitudinal outcomes and a time-to-event outcome: a corrected two-stage approach”. In: *Statistics and Computing*, pp. 1–16.

Mayosi, Bongani M et al. (2014). “Prednisolone and Mycobacterium indicus pranii in tuberculous pericarditis”. In: *New England Journal of Medicine* 371.12, pp. 1121–1130.

McLachlan, Goeffrey J (1999). “Mahalanobis distance”. In: *Resonance* 4.6, pp. 20–26.

Miller Jr, Rupert G (1974). “An unbalanced jackknife”. In: *The Annals of Statistics*, pp. 880–891.

Molenberghs, Geert and Geert Verbeke (2000). *Linear mixed models for longitudinal data*. Springer.

Pan, Jian-Xin and Kai-Tai Fang (2002). “Maximum likelihood estimation”. In: *Growth Curve Models and Statistical Diagnostics*. Springer, pp. 77–158.

Pan, Jianxin, Yu Fei, and Peter Foster (2014). “Case-deletion diagnostics for linear mixed models”. In: *Technometrics* 56.3, pp. 269–281.

Park, Ka Young and Peihua Qiu (2014). “Model selection and diagnostics for joint modeling of survival and longitudinal data with crossing hazard rate functions”. In: *Statistics in medicine* 33.26, pp. 4532–4546.

Patterson, H Desmond and Robin Thompson (1971). “Recovery of inter-block information when block sizes are unequal”. In: *Biometrika* 58.3, pp. 545–554.

Pettitt, AN and I Bin Daud (1989). “Case-weighted measures of influence for proportional hazards regression”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 38.1, pp. 51–67.

Pinheiro, José C and Douglas M Bates (2000). “Linear mixed-effects models: basic concepts and examples”. In: *Mixed-effects models in S and S-Plus*, pp. 3–56.

Post, FA, R Wood, and G Maartens (1996). “CD4 and total lymphocyte counts as predictors of HIV disease progression”. In: *QJM: An International Journal of Medicine* 89.7, pp. 505–508.

Pregibon, Daryl (1981). “Logistic regression diagnostics”. In: *The Annals of Statistics*, pp. 705–724.

Prentice, Ross L et al. (1978). “The analysis of failure times in the presence of competing risks”. In: *Biometrics*, pp. 541–554.

Proust-Lima, Cécile et al. (2009). “Joint modelling of multivariate longitudinal outcomes and a time-to-event: a nonlinear latent class approach”. In: *Computational statistics & data analysis* 53.4, pp. 1142–1154.

Proust-Lima, Cécile et al. (2014). “Joint latent class models for longitudinal and time-to-event data: A review”. In: *Statistical methods in medical research* 23.1, pp. 74–90.

Putter, Hein, Marta Fiocco, and Ronald B Geskus (2007). “Tutorial in biostatistics: competing risks and multi-state models”. In: *Statistics in medicine* 26.11, pp. 2389–2430.

R Brown, Elizabeth and Joseph G Ibrahim (2003). “A Bayesian semiparametric joint hierarchical model for longitudinal and survival data”. In: *Biometrics* 59.2, pp. 221–228.

Rabeneck, Linda et al. (1993). “A simple clinical staging system that predicts progression to AIDS using CD4 count, oral thrush, and night sweats”. In: *Journal of general internal medicine* 8.1, pp. 5–9.

Rao, C Radhakrishna (1971). “Estimation of variance and covariance components—MINQUE theory”. In: *Journal of multivariate analysis* 1.3, pp. 257–275.

Rizopoulos, Dimitris (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.

– (2017). *JM: Joint Modeling of Longitudinal and Time-to-Event Data in R*.

Rizopoulos, Dimitris, Geert Verbeke, and Geert Molenberghs (2010). "Multiple-Imputation-Based Residuals and Diagnostic Plots for Joint Models of Longitudinal and Survival Outcomes". In: *Biometrics* 66.1, pp. 20–29.

Rousseeuw, Peter J and Bert C Van Zomeren (1990). "Unmasking multivariate outliers and leverage points". In: *Journal of the American Statistical association* 85.411, pp. 633–639.

Schafer, Joseph L and Recai M Yucel (2002). "Computational strategies for multivariate linear mixed-effects models with missing values". In: *Journal of computational and Graphical Statistics* 11.2, pp. 437–457.

Schall, Robert and Timothy T Dunne (1988). "A unified approach to outliers in the general linear model". In: *Sankhyā: The Indian Journal of Statistics, Series B*, pp. 157–167.

Schwarz, Gideon (1978). "Estimating the dimension of a model". In: *The annals of statistics*, pp. 461–464.

Searle, Shayle R (1995). "An overview of variance component estimation". In: *Metrika* 42.1, pp. 215–230.

Searle, SR, G Casella, and Ch E McCulloch (1992). *Variance components: Wiley series in probability and mathematical statistics*.

Self, Steve and Yudi Pawitan (1992). "Modeling a marker of disease progression and onset of disease". In: *AIDS epidemiology*. Springer, pp. 231–255.

Self, Steven G and Kung-Yee Liang (1987). “Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions”. In: *Journal of the American Statistical Association* 82.398, pp. 605–610.

Singini, Isaac, Henry Mwambi, and Freedom Gumedze (2018). “Diagnostics for detecting influential observations in joint survival models”. In: *Annual Proceedings of the South African Statistical Association Conference*. Vol. 2018. Congress 1. South African Statistical Association (SASA), pp. 49–56.

Storer, Barry E and John Crowley (1985). “A diagnostic for Cox regression and general conditional likelihoods”. In: *Journal of the American Statistical Association* 80.389, pp. 139–147.

Swallow, William H and John F Monahan (1984). “Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components”. In: *Technometrics* 26.1, pp. 47–57.

Sweeting, Michael J and Simon G Thompson (2011). “Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture”. In: *Biometrical Journal* 53.5, pp. 750–763.

Tang, An-Min, Nian-Sheng Tang, and Hongtu Zhu (2017). “Influence analysis for skew-normal semiparametric joint models of multivariate longitudinal and multivariate survival data”. In: *Statistics in medicine* 36.9, pp. 1476–1490.

Therneau, Terry M (1997). “Extending the Cox model”. In: *Proceedings of the first Seattle symposium in biostatistics*. Springer, pp. 51–84.

Therneau, Terry M, Patricia M Grambsch, and Thomas R Fleming (1990). “Martingale-based residuals for survival models”. In: *Biometrika*, pp. 147–160.

Tsiatis, AA, Victor Degruttola, and MS Wulfsohn (1995). “Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS”. In: *Journal of the American Statistical Association* 90.429, pp. 27–37.

Tsiatis, Anastasios A and Marie Davidian (2001). “A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error”. In: *Biometrika* 88.2, pp. 447–458.

– (2004). “Joint modeling of longitudinal and time-to-event data: an overview”. In: *Statistica Sinica*, pp. 809–834.

Wang, Yan and Jeremy M G Taylor (2001). “Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome”. In: *Journal of the American Statistical Association* 96.455, pp. 895–905.

Weinberg, Jennifer L and Carrie L Kovarik (2010). “The WHO clinical staging system for HIV/AIDS”. In: *AMA Journal of Ethics* 12.3, pp. 202–206.

Welsch, Roy E (1980). “Regression sensitivity analysis and bounded-influence estimation”. In: *Evaluation of econometric models*. Elsevier, pp. 153–167.

Wulfsohn, Michael S and Anastasios A Tsiatis (1997). “A joint model for survival and longitudinal data measured with error”. In: *Biometrics*, pp. 330–339.

Xu, Jane and Scott L Zeger (2001). “Joint analysis of longitudinal data comprising repeated measures and times to events”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 50.3, pp. 375–387.

Zeng, Donglin, Jianwen Cai, et al. (2005). “Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time”. In: *The Annals of Statistics* 33.5, pp. 2132–2163.

Zewotir, Temesgen and Jacky S Galpin (2005). “Influence diagnostics for linear mixed models”. In: *Journal of Data Science* 3.2, pp. 153–177.

Zhang, Danjie et al. (2014). “Assessing model fit in joint models of longitudinal and survival data with applications to cancer clinical trials”. In: *Statistics in medicine* 33.27, pp. 4715–4733.

Zhu, Hongtu et al. (2012). “Bayesian influence measures for joint models for longitudinal and survival data”. In: *Biometrics* 68.3, pp. 954–964.