Dealing with Missing Data under Joint Modelling: Application to HIV Data

Tessie Atieno Akoko

138748

A Thesis Submitted in Total Fulfillment of the Requirements for the Degree of Masters of Science in Statistical Science of Strathmore University

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

June 2025

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Name: **Tessie Atieno Akoko**

Signature: 

Date: January 16, 2025

Approval

The thesis of Tessie Atieno Akoko was reviewed and approved by the following:

Dr. Evans Omondi

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

Dr. Collins Odhiambo

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

Dr. Godfrey Madigu

Dean, Institute of Mathematical Sciences, Strathmore University.

Prof. Bernard Shibwabo

Director of Graduate Studies, Strathmore University.

Abstract

Background: Addressing missing data poses a significant challenge in clinical research, particularly in studies like those focused on HIV. Traditional methods employed by researchers to tackle this issue often yield biased estimates and unreliable recommendations. In this study, we assess the effectiveness of an innovative strategy that integrates both longitudinal and survival processes to handle missing data, applied specifically to HIV data from Kenya obtained from the Kenya Health Information System (KHIS).

Methods: We conducted simulation studies by generating five datasets with varying percentages of missingness: 0% (representing a complete simulated dataset), 10%, 20%, 30%, and 40%. Subsequently, multiple imputation was performed on the four datasets containing missing values. This was followed by fitting a joint model to the imputed datasets. After the simulation studies, we applied the analysis to the real HIV data and conducted some diagnostic tests to the fitted joint models.

Results: The results indicate that there is no discernible difference in the models post-imputation across different percentages of missingness. Additionally, the joint model exhibits a good fit for our data compared to individual sub-models for longitudinal and survival analysis.

Conclusion: Joint modeling integrating survival and longitudinal models, emerges as a powerful statistical approach in clinical research, particularly in HIV studies, to address complex data structures and missing data challenges. This study concludes with insights into its significant contributions and future directions in clinical research.

Keywords: Missingness; Joint model; Time-to-event; Imputation; Simulations.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AIDS	Acquired Immune deficiency Syndrome	ANOVA	Analysis of Variance
EMR	Electronic Medical Records	FB	Full Bayesian
GEE	Generalized Estimating Equation	GLM	Generalized Linear Models
LOCF	Last Observation Carried Forward	MANOVA	Multivariate Analysis of Variance
MAR	Missing at Random	MCAR	Missing Completely at Random
MER	Mixed Effect Regression	MRM	Mixed Regression Model
MNAR	Missing Not at Random	NRTI	Nucleotide Reverse Transcriptase Inhibitor
IDU	Injection Drug Users	WHO	World health Organization

Acknowledgement

I extend my sincerest gratitude to my Lord Jesus for His unwavering love, strength, and faithfulness throughout my academic journey, providing me with immense courage, particularly during the completion of this work.

I also wish to express my heartfelt appreciation to my esteemed supervisors at Strathmore Institute of Mathematical Sciences (SIMS), Dr. Collins Odhiambo and Dr. Evans Omondi, for their invaluable guidance, encouragement, unwavering support, and thoughtful consideration throughout the process of my dissertation. Their wealth of knowledge, prompt responses to my inquiries, and frequent discussions significantly influenced the course of this study. I am grateful to the Strathmore Institute of Mathematical Sciences and the faculty for offering support and a conducive environment throughout my academic program.

Lastly, but most importantly, I am deeply grateful to my amazing, wonderful parents, siblings, and especially to my loving and supportive husband, who stood by me every step of the way, offering affection, inspiration, and steadfast assistance. Their continuous motivation has been a wellspring of strength, and I remain profoundly appreciative of their consistent dedication, love and support.

Dedication

This thesis is dedicated to God Almighty for his strength and love.

Chapter 1

Introduction

1.1 Background to the Study

Missing data are common in all fields of research such as cancer research, epidemiology, clinical trials, educational research, and general medicine among others (Little and Rubin, 2019). Missing data occur when some observations or subjects do not have values for a variable of interest, which can be a result of loss of subject (patient), subject refusing to respond, data entry error (mechanical error), or physician not ordering some tests. Missing data pose major problems to statistical analysis and inference making. For example, missingness can reduce the representativeness of a sample, reduce statistical power, or bias the estimates (Kang, 2013) —the magnitude of the missingness problem depends on the mechanisms of the missingness and the strategy used to handle it. Though strategies such as dropping observations with missing data, the so called complete case analysis (case-wise deletion), pair-wise deletion, mean imputation, regression imputation, and stochastic regression imputation have been long used to handle missing data (Little and Rubin, 2019), such methods result in too large (too small) standard errors, resulting in biased estimates, inferences and conclusions. In this thesis, we test the efficacy of a simultaneous modeling approach of dealing with missing data in longitudinal and survival datasets using the Human Immunodeficiency Virus (HIV) and acquired immunodeficiency syndrome (AIDS) data in a developing country context (Kenya).

1.1.1 Nature of Missing Health Data

Sub-Saharan Africa (SSA) accounts for the highest proportion of the global HIV burden (Jahagirdar et al., 2021); (Kharsany and Karim, 2016). Improved access to testing and

treatment has resulted in a decline in infections in the region over the years (Rapaport et al., 2023). However, tracking of progress at the regional level is a challenge mostly because of under(mis)reporting (Rosenberg et al., 2023). Though the HIV care continuum, which shows how individuals move from one stage of the continuum to another over time has been suggested as a convenient framework for tracking viral suppression (Lesko et al., 2016), its use is largely limited by lack of longitudinal cohorts. These trends are the same in Kenya, the focus of this study.

The level of HIV AIDs prevalence in Kenya was estimated at 1.4 million people aged between 15 and 64 years in 2021 (Young et al., 2023). Underreporting of the disease burden presents a major modelling challenge (Rosenberg et al., 2023). The "gold standard" for measuring incidence and prevalence is tracking a cohort of subjects over a long period (HIV care continuum)—measuring new infections, access to testing, access to treatment such the antiretroviral treatments (ARTs), risk behavior, and linking these trends to treatment and prevention (Lesko et al., 2016). However, such data are rare in developing countries including Kenya (Kim et al., 2011).

The Kenya AIDS Strategic Framework requires that Kenya tracks data on HIV/AIDS to effectively enhance evidence-based prevention and control. As such the country conducts periodic surveillance including aggregation of monthly facility data, annual mathematical modeling, Kenya demographic surveys that include HIV, antenatal surveillance, and AIDs indicator surveys (National AIDS and STI Control Programme; NASCOP, 2022). However, the biggest challenge with the surveillance system in Kenya is obtaining complete longitudinal data on antiretroviral treatment (ART) entry, viral suppression, and other relevant outcomes (MoH, 2020). This is because the national population-based surveys in Kenya are usually cross-sectional and are independently conducted, which makes harmonization of the data challenging (Achia et al., 2022). For example, the testing algorithms used in the 2007, 2012, and 2018 HIV prevalence tracking surveys were different (Young et al., 2023). In addition, weak health information systems, inadequate training of healthcare personnel, and logistical challenges in data collection and reporting contribute to missing HIV data (Rodríguez et al.,

2021). Inconsistent reporting practices, data entry errors, and a lack of standardized protocols for data collection further exacerbate the problem.

Some other factors that also contribute to the HIV/AIDS data problem in Kenya include the social stigma associated with HIV/AIDS which can cause individuals to be reluctant to disclose their HIV status or seek testing and treatment (Levy et al., 2021). This can lead to under-reporting of HIV cases in official records, resulting in missing data and inaccuracies in estimates of HIV prevalence and incidence (Vourli et al., 2021). An associated problem is the access to testing facilities. Evidence has shown that access to healthcare and testing facilities results in better economic status (Jakubowski et al., 2022). However, access to healthcare and testing facilities is a challenge in Kenya and SSA in general. Though access to HIV testing and canceling (HTC) has increased remarkably in Kenya (Ng'ang'a et al., 2014), evidence shows major disparities across regions and sociodemographic segments particularly rural populations (Mwangi et al., 2022), thus limiting the representativeness of the data.

1.1.2 Joint Modeling of Data

Traditionally, two distinct models are employed: one focusing on survival analysis to study the timing of an event, and another, a mixed effects model, to depict pattern of repeated observations across time. However, this approach may lead to flawed conclusions as it fails to account for the interconnectedness between these two aspects of the data concurrently (Lim et al., 2013). For clearer evaluation of the interlinkages connecting the repeated measure biomarker and time to event to appreciate better the outcome of a treatment a joint model that simultaneously brings together time to event and longitudinal model into a single model is considered (Ibrahim et al., 2010).

The joint model comprises two primary components: the event timing and the longitudinal aspect. Usually, a classical mixed effects model is applied to analyze repeated measurements, while the traditional Cox proportional hazard model is suitable for event timing data, especially concerning external time-varying covariates (Ibrahim et al., 2010).

In the joint modeling approach, the process involves integrating a longitudinal sub-model with a survival sub-model to create a unified joint model, followed by estimating coefficients within this combined model (Wulfsohn and Tsiatis, 1997).

The structure of the joint model entails:

$$\lambda_i(s|r_i,k_i) = \lim_{ds\to} Pr(s \leq S_i < s+ds|S_i \geq s, r_i, k_{ij})|ds \tag{1.1}$$

$$= \lambda_0(s)exp(r_i\alpha + k_{ij}\gamma), \tag{1.2}$$

Implying that:

$$\lambda\,(s|r_i,k_{ij}) = \lambda_0(s)exp(r_i\alpha + k_{ij}\gamma), \tag{1.3}$$

where s represents time, r_1 represents individual-specific covariates, and k_i represents group-specific covariates. This function describes the instantaneous rate at which an event occurs at time s for a specific individual i, given their covariate values r_i and k_i. The conditional intensity function is modeled using a Cox proportional hazards framework, where $\lambda_0(s)$ is the baseline hazard function, representing the hazard rate at time s for an individual with all covariates set to zero. The term $exp(r_i\alpha + k_{ij}\gamma)$ represents the relative hazard ratio associated with the individual-specific covariates r_i and the group-specific covariates k_i. The special feature of the joint model is that the parameters are estimated simultaneously.

In several distinct fields and medical studies, the use of the joint modeling approach has emerged progressively as a potent tool (Ibrahim et al., 2010). In earlier studies, the use of joint models for event-to-time and repeated measures analysis was provoked by HIV/AIDS clinical trials. Specifically, joint modeling of repeated measurements and event-to-time of CD4 counts. People living with HIV are normally investigated up until the time to event of interest in consideration takes place (Trickey et al., 2017). In such studies, the association between longitudinal bio-markers and time up until an event of interest is the

focal point of interest (Tiruneh et al., 2021). It's essential to evaluate how treatment impacts the survival biomarker process. This includes examining the variations in biomarker patterns over time and exploring the connection between the primary endpoint and characteristics of the longitudinal profiles (Lim et al., 2013). Despite the focus on the subject, two common hurdles often faced are measurement inaccuracies and bias resulting from incomplete data.

1.2 Statement of the Problem

In most biomedical and other types of research principally longitudinal studies incomplete data is a common challenge that significantly impact study outcomes (Ibrahim and Molenberghs, 2009). Missing data may arise from various sources, including participant dropout, non-response, or errors during data collection. When researchers choose to focus solely on complete cases, they risk introducing bias into their estimates, leading to potentially unreliable conclusions.

In this study, we aim to address this issue by developing and assessing the effectiveness of a joint model specifically designed to handle missing data. Joint models allow for simultaneously modelling of both the outcome and the missing data mechanism, providing a more comprehensive framework for analysis. Our research will explore how joint models effectively handle missing data, offering insights into their performance compared to traditional methods.

1.3 Research Objectives

1.3.1 Main Objective

The main objective of this study is to assess the performance of joint models of longitudinal and survival analysis in dealing with missing values and their application to HIV data.

1.3.2 Specific Objectives

1. To develop joint models for the analysis of HIV data.

2. To evaluate the performance of joint models with the application to HIV data.

3. To investigate the performance of joint models in dealing with missing values

1.4 Justification of the Study

In HIV research, patients undergo measurements of biomarkers like viral load until they experience outcomes such as death or AIDS. The main objective is to understand how treatment influences survival rates. However, a common obstacle in such analyses is the presence of missing data. In HIV prevention efforts, missing values due to dropout or non-participation can undermine the statistical integrity of conclusions if not properly addressed. Accurate estimates of HIV prevalence are crucial for monitoring ongoing programs, guiding HIV prevention and treatment initiatives, and efficiently allocating resources within countries (Muturi, 2005) there is a need for accuracy and credibility of outcomes for precise and proper policies when it comes to preventive measures.

1.5 Significance of the Study

Clinical trials, particularly those focusing on HIV/AIDS, are inherently sensitive due to their involvement with human subjects and their health outcomes, including recovery and survival. Consequently, the stakes are high in such trials. Given that these trials entail longitudinal observation of patients over extended periods, there is a significant probability of dropout or non-response, leading to missing data. Mishandling this missing data can result in erroneous conclusions and subsequent policy interventions for several reasons.

Firstly, extensive missing data diminishes statistical power, particularly when employing panel data analysis approaches that typically exclude unbalanced panels, thus compromising

the validity of the findings. Secondly, failing to comprehend the nature of missingness can introduce bias into the conclusions; for instance, certain interventions might disproportionately affect specific patient demographics, leading to their dropout from the study. Ignoring these considerations in the analysis of HIV data can have substantial repercussions.

Our study advocates for the adoption of joint modeling as a solution to address these challenges. Implementing this approach will enable researchers in clinical studies, especially those focusing on HIV/AIDS, to enhance statistical power in datasets with missing values and improve the validity of conclusions drawn from such data. This, in turn, will facilitate the formulation of accurate inferences and policy recommendations based on clinical datasets.

Chapter 2

Literature Review

2.1 Introduction

This chapter provides an overview of longitudinal and survival data, an Analysis of each data type, benefits and challenges of the data. We also review the issue of missing data in each data type and the different ways that have been employed when it comes to handling this occurrence under joint models.

2.2 Longitudinal Data

Longitudinal data involves the repetitive measurement of one or more variables gradually for a particular group of subjects. In a typical longitudinal study, observations are gathered for each unit at predetermined intervals. When the study design ensures that each unit is measured consistently at each interval, resulting in an equal number of observations for each unit, the responses for the i^{th} unit can be represented by a $T \times 1$ vector, denoted as Y_i where $i = 1, ..., N$. Additionally, we use X_i to represent a $T \times p$ design matrix for the i^{th} individual, which encompasses both individual characteristics and the temporal design (Laird, 1988). As an example of longitudinal data, HIV patients might undergo regular monitoring with monthly assessments of the viral load and CD4 count taken for purposes of the number of HIV particles in the blood and immune system evaluation in terms of disease progression respectively.

A unique feature of longitudinal data is the presence of correlation in the repeated measurements of a variable within an individual. An assumption of independence is made across

the individuals. For instance, throughout a hypertension clinical trial, individuals' blood pressure may be measured often. Within an individual, the measurements taken of blood pressure have a high likelihood of correlation because the present blood pressure value may rely on the preceding value. In a random selection, the assumption of independence holds for measurements between two individuals.

2.2.1 Longitudinal Data Analysis

Approaches such as analysis of variance (ANOVA) are starter methods in longitudinal studies analysis where the attention is on the comparison of group means but no insight is given on the trends and patterns in subjects over a period of time. This limitation inhibits the ANOVA approaches in proper handling of missing data and irregularly timed data. Multivariate analysis of variance (MANOVA) and ANOVA require fully complete data and measurement of participants at identical intervals respectively (Garcia and Marder, 2017). Therefore, utilizing ANOVA techniques on data with absent observations leads to skewed parameter estimations (Shaw and Mitchell-Olds, 1993).

The two most commonly preferred approaches for longitudinal data are the generalized estimating equations model and mixed effects regression. In the case of correlated data, GEE models offer an extension to generalized linear models (GLMs). Hence, these models are widely employed for analyzing count and categorical outcomes, though they can also be applied to continuous outcomes. When contrasting Generalized Estimating Equations (GEE) with Mixed Regression Models (MRMs), one distinction lies in GEE models being built on quasi-likelihood estimation, hence the complete likelihood of the data remains undefined. GEE models, referred to as marginal models, focus on the regression of y on x, treating the parameters' association independently (Gibbons et al., 2010).

The primary advantages of GEEs lie in their clarity of estimation and resilience to inaccuracies in specifying the correlation structure of repeated measures (Garcia and Marder, 2017). GEEs come with certain constraints: they primarily emphasize regression parameters rather than all model parameters, complicating the utilization of Bayesian/Akaike information criteria and

likelihood ratio tests for testing and comparing model fits. This is because GEEs don't allow hypothesis testing on correlation parameters. Another limitation is their assumption that missing data are missing completely at random, In contrast to models utilizing full-likelihood estimation, which presuppose that missing data occur randomly (Garcia and Marder, 2017).

The Mixed Regression Models (MRMs) represent the correlation of repeated measures through the use of random effects that offer a description of cluster-specific trends progressively. Mixed Effects Regression (MER) models typically provide details about both the correlation structure of repeated measures and the regression relationship between covariates and recurring feedback (Garcia and Marder, 2017). The application of generalized mixed-effects regression models can be implemented to both Gaussian distributed continuous results together with categorical outcomes and other not normally distributed outcomes (Gibbons et al., 2010).

Mixed Effects Regression (MERs) offers several advantages over GEEs. It facilitates hierarchical modeling across multiple levels of data, allowing predictions at each hierarchical level. Additionally, it enables hypothesis testing on correlation parameters since they are directly estimated. Traditional methods like likelihood ratio tests and Akaike/Bayesian Information Criteria can be used to assess and compare model fits because all model parameters, including regression and correlation parameters, are estimated. MER models also demonstrate greater resilience to missing data, assuming that missingness follows the Missing At Random (MAR) assumption, which is broader than the Missing Completely At Random (MCAR) assumption of GEEs. Nonetheless, MER models are computationally more intricate than quasi-likelihood methods, and the validity of their hypothesis testing inferences hinges on the accurate specification of the correlation structure and the mean of the repeated responses (Garcia and Marder, 2017).

2.2.2 Benefits of Longitudinal Studies

One significant advantage of longitudinal studies lies in their capacity to assess changes in outcomes and exposure on an individual basis. They offer the chance to observe unique

patterns of change over time. Specifically, a prospective longitudinal study tracks the onset of new diseases. This timing can be linked to recent alterations in patient exposure or to chronic exposure (Fitzmaurice, 2009).

In longitudinal studies, researchers have the ability to discern between age, period, and cohort effects. When examining changes over time, it's essential to consider various time scales. The cohort scale represents the year of birth, for instance, 1945 or 1963, while the period indicates the current time, like 2003. Age is derived from the period minus the cohort. In a longitudinal study with measurements taken at different times $t_1, t_2, \ldots, t_n$, researchers can simultaneously examine multiple time scales, including age and cohort effects, by integrating covariates derived from the calendar time of the visit and the participant's birth year (Lebowitz, 1996).

2.2.3 Complications in Longitudinal Data

In longitudinal studies, measurements are inherently correlated due to their design. This correlation stems from the repeated assessment of the same individual over time. Neglecting to account for the various sources of correlation in longitudinal studies can lead to significant ramifications, including increased rates of false positives and confidence intervals that are rendered invalid due to underestimated standard errors (Gibbons et al., 2010). Another concern about longitudinal studies is dropout, missing values, and attrition amongst subjects in the study. Missing data presents a prevalent issue in longitudinal studies, especially when the subject of study is an individual or when there are substantial intervals between observations (Laird, 1988). To be able to handle the issue of missing data correctly, consideration of the three missing data mechanisms has to be checked first. According to (Rubin, 1976), the three missing data mechanisms include; Missing Completely at Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). The dynamics of the three missing assumptions will be discussed further in this paper.

In longitudinal studies, there is also the challenge of measurement error in that errors might be found in the observed values making them not the true values. The two classifications of measurement errors include; systematic error and random error. In situations where

there is the presence of unsound measurement procedures such as imperfect calibration of a measurement apparatus or unsound observation, the error can be defined as systematic error. This type of error cannot be lessened by averaging measurements. Random errors typically arise spontaneously and can be mitigated by averaging multiple estimates or replicates (Yu, 2019).

2.3 Survival Data

Survival data, also referred to as time-to-event data, denotes the duration until a specific event occurs. Typically, survival data are analyzed in terms of two interrelated probabilities: survival and hazard. The survival probability, denoted by S(t) or survivor function, represents the likelihood that an individual survives from the starting point to a designated future time t. On the other hand, the hazard, often denoted by $h(t)$ or $\lambda(t)$, signifies the probability that an individual being observed at time t experiences an event at that precise moment. In simpler terms, it reflects the instantaneous event rate for an individual who has already survived up to time t (Efromovich, 2018).

The distribution of survival times of a survival data is normally skewed thus, the survival data does not follow a Gaussian distribution or any other systematic distribution. Another characteristic of survival data is the possibility that some individuals may not be observed for the full time to failure, a phenomenon known as censoring. This can also occur in longitudinal data where there is a limitation of appraising. It may occur due to the following reasons; A person withdraws from the study; a person cannot be found for follow-up during the study period and a person does not experience the event before the study ends. The reasons stated can be classified as right-censoring. By right censoring, survival time is only known to surpass a given estimate.

Under the censoring mechanism, there are two types namely; non-informative and informative censoring. In non-informative censoring, there is no basing of the likelihood of being censored on the expected failure time but can rely on the predictors not related to the event process

such as age whereas the informative censoring there is a relationship between the likelihood of being censored to the expected failure time (Yu, 2019).

2.3.1 Analysis of Time to Event Data

When analyzing survival data, three main methodological approaches are commonly recognized: Non-parametric methods: These techniques do not impose assumptions on the distribution of survival times or on the relationship between covariates and survival time. Examples include the Kaplan-Meier estimator and the log-rank test, Semi-parametric methods: These methods also do not assume a specific distribution of survival times but do assume a particular relationship between covariates and the hazard function, and consequently, the survival time. The Cox Proportional Hazards (PH) model, widely utilized in survival analysis, is a semi-parametric method and Parametric methods: These methods assume a distribution for survival times as well as a functional form for the covariates (Schober and Vetter, 2018).

With the Kaplan-Meier estimator, the survival probability is nonparametrically estimated based on observed survival times, whether they are censored or uncensored, employing the KM (or product limit) approach (Kaplan and Meier, 1958). Kaplan- Meier survival curves and Logrank test are methods used under univariate analysis. The two strategies are concerned about survival with consideration to the factor under evaluation assuming the impact of any other factors.

The Cox proportional hazards model, often referred to as the Cox PH model (Cox, 1972), is widely adopted as the primary multivariate method for examining survival time data in medical studies. It functions as a survival analysis regression model, delineating the connection between event occurrence, as depicted by the hazard function, and a collection of covariates (Bradburn et al., 2003). The Cox proportional hazards (PH) model assumes that covariates remain constant over time. In simpler terms, variables such as gender and age at diagnosis are expected to stay the same throughout the observation period. However, extensions of the Cox model exist to accommodate covariates that change over time, such as blood pressure measurements at different follow-up intervals (Clark et al., 2003).

2.3.2 Complications in Survival Data

When it comes to survival studies the primary focus on analysis of times is on the outcome. In the analysis of such studies missing values in the covariates are a challenge when it comes to handling time-dependent covariates and time-varying effects. Furthermore, when it comes to the proportional hazards assumption in the application of a Cox model questions on how best to approach the investigation of model assumptions are raised. The presence of missing data poses more hindrances beyond survival occurrences. These challenges include flexibility in the modeling of covariates and how to go about choosing covariates in a model (Carroll et al., 2020).

2.4 Missingness Patterns

There are two categories of missing data patterns: non-monotone and monotone. In a scenario where all other outcomes miss due to a missing outcome at some visit, such missingness is known as monotone missing data pattern. In this pattern the likelihood of a participant returning once they have dropped out is negligible. Practically, the chance of data having a monotone missing pattern is rare. A specific instance of a monotone missing data pattern is termed a uniform missing data pattern, where an individual's outcomes are either entirely missing or fully observed at all time-points. Non-monotone missing data involves random missing pattern. Unlike monotone missing pattern, in non-monotone pattern subjects can decide to return to the study on a later period.

2.5 Missing Data Mechanisms

According to (Rubin, 1976), the three main mechanism for missing values consists of MAR, MCAR and MNAR.

2.5.1 Missing Completely at Random

Under the assumption of missing completely at random (MCAR), an analysis can be performed while overlooking the missing data since the missing data are autonomous of unobserved data and observed data (Molenberghs et al., 2015). In cases where missingness follows the Missing Completely At Random (MCAR) condition, it means that the absence of data is unrelated to the response variables. For instance, in an HIV prevention study where HIV status is missing due to a random error in data entry, it can be categorized as MCAR, specifically due to non-response (Harel et al., 2012). Other common examples of situations where missing value is MCAR include; an arbitrary failure of an experimental equipment and migration of a study participant to a different region thus unable to finish the visits.

2.5.2 Missing not at Random

In the context of missing not at random (MNAR), missing data are influenced by both observed and unobserved values (Molenberghs et al., 2015). For example, if individuals with positive but unobserved HIV status are more inclined to have missing data on their HIV status, then the unobserved HIV status values are considered to be missing not at random (MNAR) (Harel et al., 2012). Another example of MNAR phenomena include: Upon achieving an outcome, a participant misses a visit.

2.5.3 Missing at Random

The Missing at Random (MAR) assumption is applied when the probability of a covariate being missing is solely reliant on observed values (Molenberghs et al., 2015). For instance, in the context of an HIV prevention trial, if the absence of HIV status data is related only to age and gender, having complete data for these variables signifies a MAR mechanism (Harel et al., 2012). Other instances of MAR include dropout due to documented side-effects and missing data stemming from aspects of the study design.

To sum up, any missing data mechanism can be characterized by defining the conditional distribution $P(S|Y)$ of the outcome indicator given the complete data. The specification of Missing Completely At Random (MCAR) is outlined as follows:

$$P(S|y) = P(S) \tag{2.1}$$

MAR is specified as;

$$P(S|y = P(S|y_{obs}) \tag{2.2}$$

In equation 2.1 S represents the missing data pattern and y represents the observed data. This equation indicates that the probability distribution of the missing data pattern S is unrelated to the observed data y. In equation 2.2, the outcome response S relies solely on the observed data y_{obs}. In equation 2.1, the response indicator S does not hinge on the complete data y. Any missing data where S is influenced by the values of y_{mis} is categorized as Missing Not At Random (MNAR).

2.6 Missing Data Handling in Longitudinal Studies

The various statistical procedures used in handling missing values include; Complete- Case analysis, imputation techniques, Maximum likelihood technique, generalized estimating equation, and bayesian method.

2.6.1 Case Deletion

Complete-case analysis is a straightforward and the most common approach where there is exclusion of all units that did not offer data at all the intended occurrences. The exclusion can be a result of missing values or dropouts of subjects. Case deletion offers the advantage of being applicable to any statistical analysis without the need for specialized computational methods. When data adhere to the Missing Completely At Random (MCAR) condition, this

method can produce unbiased results (Little and Rubin, 2014). However, the technique of complete case analysis is less feasible when the assumption of MCAR does not hold.

2.6.2 Imputation Methods

To address the difficulties associated with complete case analysis, missing values can be imputed before conducting the analysis. Imputation methods fall into two main categories: simple imputation and multiple imputation.

Simple Imputation Methods

Simple imputation involves substituting missing values with a single imputed value to generate a complete dataset, which is then analyzed using standard analytical techniques. Common methods of simple imputation include mean-value imputation, maximum-minimum imputation, best/worst value imputation, and last observation carried forward.

In the last observation carried forward (LOCF) method, the last recorded value of a subject before dropout is used to replace missing values for subsequent time points until the end of the follow-up period. While LOCF is applicable for chronic repeated measurements outcomes, it overlooks post-dropout subject variability and individual reasons for dropout. Bias can be introduced when using LOCF, particularly when the data is Missing Completely At Random (MCAR).

Mean-value imputation can be implemented in two ways: for missing values after dropout, the mean value of a subject's repeated measurement outcomes can be imputed, or for each time point, the mean value of the longitudinal outcome can be imputed. However, this method does not consider each patient's individual pattern but rather relies on outcome values before dropout. In analyses where the focus is on variation in outcome over time, this approach may not be suitable for a complete dataset. In cases where the data is discrete and mean imputed values are not integers, a decision needs to be made whether to round to the nearest whole number or to leave the value as is.

Multiple Imputation

Multiple imputation involves generating multiple values for each missing observation using an imputation model, fitting the model of interest to each imputed dataset, and then combining the parameter estimates to obtain a pooled estimate (Rubin, 1987). If the imputation model is accurately specified, multiple imputation (MI) yields consistent and asymptotically efficient parameter estimates under the Missing At Random (MAR) assumption (Carpenter and Kenward, 2013).

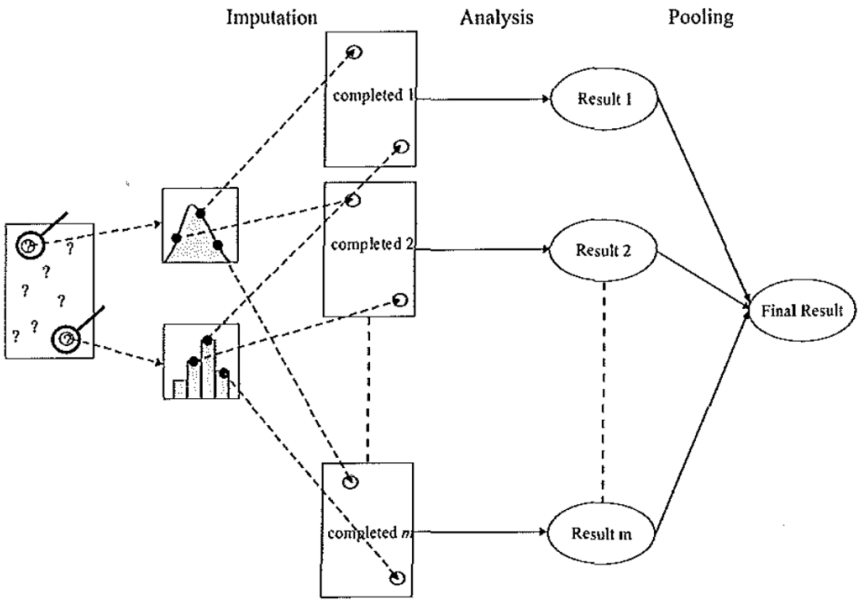

Figure 2.1: Concept of Multiple Imputation

From Figure 2.1 (Rubin, 1987), The original dataset contains missing values are represented by empty cells. These missing values can occur for various reasons, such as non-response, data entry errors, or study design. In the imputation step, multiple imputed datasets are created by replacing the missing values with plausible estimates. Each imputed dataset is generated using a statistical model that takes into account the observed data and the uncertainty associated with the missing values. Each imputed dataset is analyzed separately using the same statistical model. This step involves performing the desired analysis (e.g.,

regression analysis, hypothesis testing) on each complete dataset to estimate the parameters of interest.

The results obtained from analyzing each imputed dataset are combined using specific rules, such as Rubin's rules. These rules take into account both within-imputation variability (variation between imputed datasets) and between-imputation variability (variation within imputed datasets), providing a single set of estimates and standard errors that incorporate the uncertainty due to missing data. The combined results are used for final inference, including hypothesis testing, model interpretation, and drawing conclusions. By incorporating information from multiple imputed datasets, the final estimates are more reliable and accurate compared to using a single imputed dataset or complete-case analysis.

The imputation model generates n (where $n \geq 2$) imputed values for each missing data point, distributed around its expected value. These n completed datasets are individually analyzed using the preferred statistical method for complete data. Subsequently, the n intermediate results are combined into a final result through explicit procedures. For pooling the completed data results, the analysis of complete data is represented by $(\hat{J}, U)$, where $\hat{J}$ represents a point estimate of the parameter of interest J, and U denotes the variance-covariance matrix of $\hat{J}$ from the n completed data results $(\hat{J}_1, U^*, ..., (\hat{J}_m, U_m^*))$. The unknown complete data results $\hat{J}$ and U are estimated by $(\bar{J}_m, \bar{U}_m)$, with $\bar{J}_m$ as given by:

$$\bar{J}_m = \frac{1}{m} \sum_{i=1}^{m} \hat{J}_i \tag{2.3}$$

and $\bar{U}_m$ given by

$$\bar{U}_m = \frac{1}{m} \sum_{i=1}^{m} U_i^* \tag{2.4}$$

The additional uncertainty in the inference regarding J arising from missing data is captured by the between-imputation variance-covariance matrix B_m, which is expressed as:

$$B_m = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{J}_i - \bar{J}_m)(\hat{J}_i - \bar{J}_m)^T \tag{2.5}$$

Using statistics such as $\bar{J}_m$, $\bar{U}_m$, and B_m, standard errors, confidence intervals, and p-values can be computed. The underlying assumption is that the sample size is large enough to approximate the distribution of $\bar{J}_m$ with a multivariate normal distribution (Rizopoulos, 2012).

2.6.3 Maximum Likelihood

Maximum likelihood is a statistical method suitable for large samples that aims to find parameter estimates with the highest probability of generating the observed data, according to a specified model. These estimated parameters are referred to as maximum likelihood estimates (Dempster et al., 1977; Little and Rubin, 1989, 2014). In maximum likelihood estimation, both fully observed and partially observed cases are utilized to estimate model parameters, with missing data values treated as random variables to be averaged over (Collins et al., 2001).

The principle of maximum likelihood starts from a sample $Y = (Y_1, ..., Y_n)$ of random variables that are determined in accordance with the family of probabilities γ_θ. Moreover, $f(y_\theta)$, where $y = (y_1, ..., y_n)$, denotes the density function for the data when θ represents the true state of nature.

The principle of maximum likelihood selects an estimator $\hat{\theta}$, which is the parameter value that maximizes the likelihood of the observed data. The likelihood function represents the density function considered as a function of θ.

$$L(\theta|y = f(y|\theta), \theta \in \Theta \tag{2.6}$$

The maximum likelihood estimator

$$\hat{\theta}(y) = \arg\max_{\theta} L(\theta|y) \tag{2.7}$$

2.6.4 Generalized Estimating Equation

For correlated data, generalized estimating equation (Liang and Zeger, 1986) became one of the most used procedures in practice. In this approach, the researcher defines a working correlation structure, but it doesn't need to precisely match the true correlation structure. This makes it a semi-parametric procedure, as the model for the data must be specified, but the correlations themselves are not the primary focus. Unfortunately, GEE for incomplete data is only unbiased under the Missing Completely At Random (MCAR) assumption. However, there is an infrequently employed extension known as weighted GEE (Robins et al., 1994), allows missing data under the MAR condition (Harel et al., 2012).

2.6.5 Bayesian Method

Complete Bayesian techniques are commonly executed through iterative algorithms like Markov Chain Monte Carlo (MCMC) methods (Brooks, 2011). When using weakly informative prior distributions, inferences derived from Full Bayesian (FB) methods rely solely on the observed data and remain valid under the Missing At Random (MAR) assumption (Heitjan, 2009).

In this study, we will use multiple imputation approach in handling missing values under the joint model.

2.7 Joint Model of Longitudinal and Survival Data

Joint modeling is an advanced and intricate method, yet it allows for the simultaneous modeling of both longitudinal repeated biomarker measurements and survival processes, considering their interdependencies (Lim et al., 2013). When estimating parameters, joint models of survival and longitudinal data can enhance efficiency and mitigate bias (Rizopoulos, 2012).

The desirability of joint model of survival and longitudinal data is frequent under the following occurrences: Firstly, a joint model can be used to analyze and understand the association between survival and longitudinal processes. In such a scenario, both the two processes are of chief attentiveness. The primary goal is to achieve an effective inference. For example, a latent process such as disease progression can be used to control both survival and longitudinal processes. Secondly, Longitudinal analysis with informative dropouts. In such a scenario, the primary aim is repeated measurements analysis, but there is a need to model the time to drop out so as to minimize possible biases related to informative dropouts (Yu, 2019).

Additionally, survival models incorporate time-dependent covariates, which could be subject to measurement errors or missing values. While the primary focus remains on survival analysis, longitudinal models are incorporated to account for these issues related to time-dependent covariates.

(Rajan and Leurgans, 2010) employed a multiple weighting approach under missing at random assumption when simultaneously modeling for non-participation and death. (Chen et al., 2014) used a joint model for the analysis of longitudinal and survival data accounting for both missingness and left censoring in the longitudinal covariates. Yuriko Takedo did an estimation by Bayesian approach and evaluated the performance of the method through Monte Carlo simulation. Using a logit model (Lu et al.,2017) did an analysis of a joint model strategy for skew longitudinal survival data with mismeasured covariates and missingness. (Parzen et al., 2006) used a pseudo-likelihood approach applied for repeated measurement binary data with non-ignorable missing responses and covariates. For the missingness in the predictor variables, a separate model has also been contemplated.

Despite the previous research work, there is still the issue of biased results when it comes to choosing the appropriate missingness assumption and also the need to extend various methods in handling missingness to different application situations. (Harel et al., 2012) suggested the need for researchers to attend more closely to missing values in HIV prevention trials.

2.8 Approaches Proposed to Address Missing Data in Joint Modeling Frameworks

Several approaches have been proposed to address missing data in joint modeling frameworks.

Complete Case Analysis (CCA): This approach involves analyzing only those individuals with complete data on all variables of interest. However, CCA may lead to biased estimates if missingness is related to the outcome or predictors.

Available Case Analysis (ACA): ACA involves including all available data in the analysis, which maximizes the sample size but may introduce bias if missingness is related to the outcome.

Multiple imputation (MI): Multiple imputation (MI) is a commonly employed technique that entails generating several plausible imputed datasets. In this method, missing values are substituted with estimated values derived from observed data and a predetermined imputation model. Analyses are then conducted on each imputed dataset, and results are combined to obtain overall estimates and standard errors. MI can provide unbiased estimates under the missing at random (MAR) assumption.

Selection Models: These models explicitly account for the selection mechanism that determines which subjects have complete versus incomplete data. They model the probability of missingness as a function of observed data and can provide valid estimates under certain assumptions.

Pattern Mixture models: These models classify individuals based on their patterns of missingness and allow for different parameter estimates for each missingness pattern. They provide flexibility in modeling different missing data mechanisms but may be challenging to implement and interpret.

Shared Parameter Models: These models incorporate latent variables or random effects to jointly model longitudinal and survival processes, allowing for flexible modeling of the relationship between them while accommodating missing data.

2.9 Gaps Identified

The literature under the joint models of longitudinal and survival data has highlighted previous work done. Need to say is that the idea that missing values are frequent in clinical trials is not a new thing under joint models. Some researchers would rather choose to ignore this phenomenon while others choose to employ various statistical methods in dealing with missing values. Still, there is no comprehensive remedy on the best approach to incorporate. Therefore, this study will offer an extension to the existing literature on the performance of joint models when dealing with missing values with application to HIV data.

Chapter 3

Methodology

3.1 Introduction

This chapter outlines information on the data source and study design, simulation studies, mathematical formulation of sub-models of longitudinal and time-to-event data. It details the development of the joint model which is applied to simulated data to assess the model performance and the original dataset to assess missingness as well as factors that influence the levels of viral load in HIV/AIDS study.

3.2 Study Design

This study was a retrospective, repeated measures case-based evaluation. It involved using a joint model to investigate the interlinkages of longitudinal biomarkers, in this case, viral load and the time to event- death. The missing values were imputed by multiple imputation using the mice function in R (version 4.4.0, R Core Team) and R Studio (version 2024.04.0+735, R Studio Team). The analysis involved Secondary data from Kenya Health Information Services (KHIS).

3.3 Data Source

In this analysis, we used data extracted from the Kenya Health Information System (KHIS). Retrospective data was extracted between Jan 2015 to July 2022, the two regions under investigation were Nairobi and Kiambu counties.

3.3.1 Covariates and Study Outcomes

In the longitudinal sub-model, the primary response variable was the viral load alongside other covariates whereas under the survival sub-model, the response variable of interest was event of time.

The other variables used in the joint model included;

- Patient's unique identification generated from the Kenya Electronic Medical Records (EMR) system.
- *Episode* represented the visit. HIV patients are expected to do routine visits (typically every 6 months, unless the patient is viremia (viral load increases extraordinarily)).
- *RegimenLine*: HIV ART Regimen line (Typically according to WHO there are three Regimen lines. The first line consists of; Nucleotide Reverse Tran-scriptase inhibitors is for patients who are doing well. Second-line antiretroviral therapy includes two NRTIs plus ritonavir-boosted protease inhibitors for adverse patients. 3rd line is for salvage-patients who are not doing well.
- *TB-HIV-Coinfection*: Patients with HIV-TB Coinfection (Both HIV and TB infection) (1=Yes, 0=No).
- *Region* where 1 represents Nairobi and 2-Kiambu.
- *CensoredVar*: Lost to Follow-up.
- *Time*: Duration

3.4 Statistical Analysis

A joint model comprises two sub-models: one for the event occurrence time and another for a longitudinal process. It is typically assumed that the longitudinal process model and the survival time hazard are interrelated, both influenced by common underlying random effects.

For example, in a weight loss program, the survival data will be the drop-off of individuals, and the longitudinal data will involve the weights of individuals being measured during the program. Another example will include the frequent measurement of the tumor sizes of the same cancer patients (this is longitudinal data) to the event of interest, the time to death. In both examples, there is a likelihood of a relationship between the longitudinal and survival data. Thus, the feasibility of the use of joint modeling in analysis.

In the initial modeling step, the progression of longitudinal biomarker measurements is individually estimated using a random effects model based on a linear or polynomial function of time. This method allows for the estimation of biomarker changes for each patient at the point of death or censoring. In the subsequent step, a joint modeling approach employing either the Cox or Weibull parametric hazards model is employed (Lim et al., 2013).

3.4.1 Longitudinal Sub-model

In our analysis we used the linear mixed-effect model for longitudinal data where random and fixed effects are employed.

In this model, each patient's profile overtime is expressed as below:-

$$Z_{ij} = \beta_{i0} + \beta_{i1} r_{ij} + \varepsilon_{ij} \tag{3.1}$$

Where

- Z_{ij} is the j^{th} response of the i^{th} patient
- β_{i0} is the intercept and β_{i1} is the slope for patient i

The linear effect model for the real HIV data is formulated as below:-

$$Z_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) r_{ij} + \varepsilon_{ij} \tag{3.2}$$

Where:

- Z_{ij} is the j^{th} viral load value for the i^{th} patient measured with error
- $r_i j$ is the j^{th} observational timepoint for the i^{th} patient
- $b_i's$ are the random effects
- $\beta's$ are known as the fixed effects

In a general form, linear mixed model is expressed as:-

$$\begin{cases} z_i = X_i\beta + R_i b_i + \varepsilon_i \\ b_i \sim N(0, D), \qquad \varepsilon_i \sim N(0, \sigma^2 \iota_{ni}) \end{cases} \tag{3.3}$$

Where:

- b_i and ε_i are independent
- R is a design matrix for random effects b_i
- X is a design matrix for fixed effects β

3.4.2 Survival Sub-model

Survival analysis involves statistical methods used to examine non-negative random variables related to the duration from a specific time point until the happening of a particular event. In biological studies, this event is often linked to death, and the random variable is referred to as failure time, survival time, or event time.

The submodel for time-to-event data with hazard $h_i(t|b_i)$ is expressed as:

$$h_i(t|b_i) = h_0(t) exp(\gamma^T \omega_i + \alpha_E^T b_i) \tag{3.4}$$

Where:

- $h_0(t)$ is a baseline hazard

- ω_i is a baseline covariate vector, associated with a parameter γ and α_E is an association parameter

3.4.3 Mathematical Formulation of Joint Model

As an important instrument, the establishment of joint models was for the basis of analysis of data in situations where event times are computed against a repeated outcome In a joint model, the linkage between repeated measures and time to-event model is via a collaborative parameter that influences dependence between them (Rizopoulos, 2012). The joint models works on the assumption that the hazard function of the survival sub model at specific time t is associated with the value of the longitudinal process simultaneously. Joint model accounts for both endogenous and exogenous time depending covariates.

The likelihood method for joint models involves maximizing the logarithm of the likelihood function for the combined distribution of the time-to-event data T_i, event indicators σ_i, and longitudinal data y_i. The vector of time-independent random effects b_i captures the relationship between the longitudinal and event processes, as well as the correlation among repeated measurements in the longitudinal outcome (Rizopoulos, 2012). Essentially, given b_i, the longitudinal process and the survival process are conditionally independent.

$$p(T_i, \sigma_i, y_i | b_i; \theta) = p(T_i, \sigma_i | b_i; \theta) p(y_i | b_i; \theta) \tag{3.5}$$

where $p(.)$ represents the corresponding probability density function, and

$$p(y_i | b_i; \theta) = \prod p(y_i(t_{ij}) | b_i; \theta), \tag{3.6}$$

In this context, $\theta = (\theta_t^T, \theta_y^T, \theta_b^T)^T$ denotes the parameter vector comprising the event time outcome, the longitudinal outcomes, and the covariance matrix of random effects, respectively. Given the modeling assumptions previously detailed and the conditional independence

assumptions outlined earlier, the joint log-likelihood contribution for the i^{th} subject can be expressed as follows:

$$\log p(T_i, \sigma_i, y_i; \theta) = \log \int p(T_i, \sigma_i, y_i, b_i; \theta) db_i \tag{3.7}$$

$$= \log \int p(T_i, \sigma_i | b_i; \theta_t, \beta) \left[\prod p(y_i(t_{ij}) | b_i; \theta_y) \right] p(b_i; \theta_b) db_i \tag{3.8}$$

In this context, the likelihood of the survival component is expressed as:

$$p(T_i, \sigma_i | b_i; \theta_t, \beta) = h_i(T_i | M_i(T_i); \theta)^{\sigma_i} S_i(T_i | M_i(T_i); \theta), \tag{3.9}$$

The combined distribution for longitudinal responses along with the random effects is determined by the following expression:

$$\prod p(y_i(t_{ij}) | b_i; \theta_y) p(b_i; \theta_b) \tag{3.10}$$

$$= (2\pi\sigma^2)^{-n_i/2} \exp\left\{ -||y_i X_i \beta - Z_i b_i||^2 / 2\sigma^2 \right\} \tag{3.11}$$

$$\times (2\pi)^{-q_b/2} \det(D)^{-1/2} \exp\left(-b_i^T D^{-1} b_i / 2 \right), \tag{3.12}$$

Here, q_b represents the dimensionality of the random effects vector, and $||.||$ denotes the Euclidean vector norm. Subsequently, maximizing the log-likelihood with respect to θ for all observed data is formulated as:

$$\iota(\theta) = \sum \log p(T_i, \sigma_i, y_i; \theta), \tag{3.13}$$

requires a combination of numerical integration and optimization algorithms. The likelihood function in joint modeling typically involves multiple parameters and integrals, making it analytically intractable to find closed-form solutions. Optimization algorithms provide numerical methods for finding the maximum likelihood estimates of the parameters. Many joint models involve nonlinear relationships between the parameters and the likelihood

function. Optimization algorithms such as gradient descent, Newton's method, or quasi-Newton methods are used to iteratively update the parameter estimates until convergence to the maximum likelihood estimates. Overall, optimization algorithms play a crucial role in fitting joint models to data by efficiently maximizing the likelihood function and estimating the parameters that best describe the relationship between longitudinal and survival outcomes.

3.5 Simulation Studies

To characterize the joint modelling we simulated a dataset that mimics the statistical distribution of the original HIV dataset. From the complete simulated dataset we fitted a joint model thereafter created random missingness of varying proportions of 10%, 20%, 30%, and 40% in the response variable in our case viral load for the longitudinal data. From literature, there is no predetermined cutoff of missingness thus the randomness in the choice of the random proportions (Dong and Peng, 2013) and the proportion of missingness has less impact on the results as compared to the pattern and mechanism of missingness (Tabachick et al.,2007). With the varying proportions we generated datasets with different portions of missing rates. We imputed values in each dataset with the varying proportions using the multivariate imputation by chained equations (MICE) package in r software. We then fitted joint models for the imputed datasets and compared the results. The last step taken was the application of analysis to the real HIV dataset.

3.6 Software Implementation

Statistical tests will be performed using R software version 4.3.3 (2024-02-29 ucrt).

Chapter 4

Results

4.1 Introduction

This chapter presents the findings of our analysis. Firstly, we discuss the results obtained from analyzing simulated data, followed by an examination of the results derived from analyzing real HIV data. For the simulated data, we generated a dataset comprising 1000 observations and introduced varying levels of missingness, specifically 10%, 20%, 30%, and 40%. The selection of these missingness levels was arbitrary. Subsequently, we employed the multivariate imputation by chained equations (MICE) package in R to impute values for the dependent variable and conducted analyses on the resulting five datasets. We then compared the regression models using imputed data with those using the complete dataset. Our assessment of model performance relied on the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

4.2 Simulation Studies

We used R software to simulate a data set that mimics the statistical distribution of the real HIV data set. The simulated dataset contained 1000 observations with 6 variables: "ID", "response", "time", "sex", "age" and "event". We fitted a joint model and the results are represented in Table 4.1.

Table 4.1 presents a joint model constructed using a complete simulated dataset. This model integrates both longitudinal and event processes and offers insights into the relationships between predictor variables and the outcomes of interest. The joint model has an AIC of

-261.518, a BIC of -223.835, and a LogLik of 140.759. Lower AIC and BIC values indicate better model fit, while a higher LogLik suggests the better likelihood of the model given the data. Under Longitudinal Process. "Intercept", "Observed time", and "Age" exhibit statistically significant positive effects on the longitudinal outcome, with p-values of 0.000, indicating strong associations. However, "Sex (Female)" does not significantly affect the longitudinal outcome, with a p-value of 0.427, suggesting that gender may not significantly influence the outcome in this context. Under the Event Process, the "Intercept" represents the baseline hazard, with a statistically significant negative effect on the event outcome (p-value = 0.000). "Sex (Female)" does not exhibit a significant effect on the event outcome, with a p-value of 0.441, indicating that gender may not significantly influence the event process. However, "Association" shows a statistically significant negative effect on the event outcome (p-value = 0.000), indicating a strong association between the longitudinal and event processes.

Table 4.1: Joint Model of Complete Simulated Dataset

Joint Model	**AIC**	**BIC**	**LogLik**	
	-261.518	-223.835	140.759	
Longitudinal Process				
	Estimate	**Std.Error**	**Z-value**	**P-value**
Intercept	11.897	0.047	255.724	0.000
Observed time	0.047	0.001	47.651	0.000
Age	0.237	0.001	183.095	0.000
Sex (Female)	0.014	0.017	0.795	0.427
Event Process				
	Estimate	**Std.Error**	**Z-value**	**P-value**
Intercept	-53.317	4.316	-12.352	0.000
Sex (Female)	-0.1599	0.2073	-0.771	0.441
Association	-2.930	0.191	-15.307	0.000

Table 4.2 presents a joint model constructed using simulated imputed data with 10% missingness. This model integrates both longitudinal and event processes and provides insights into the relationships between predictor variables and the outcomes of interest.The joint model has an AIC of -324.87, a BIC of -283.424, and a LogLik of 173.438. Under longitudinal Process "Intercept", "Observed time", and "Age" exhibit statistically significant positive

effects on the longitudinal outcome, with p-values of 0.000, indicating strong associations. However, "Sex (Male)" does not significantly affect the longitudinal outcome, with a p-value of 0.423, suggesting that gender may not significantly influence the outcome in this context. In the Event Process, "Intercept" represents the baseline hazard, with a statistically significant negative effect on the event outcome (p-value = 0.000). "Sex (Male)" exhibits a marginally non-significant positive effect on the event outcome (p-value = 0.101), suggesting a trend towards significance. "Age" demonstrates a statistically significant negative effect on the event outcome (p-value = 0.000), indicating its importance as a predictor."Association" shows a statistically significant negative effect on the event outcome (p-value = 0.000), indicating a strong association between the longitudinal and event processes.

Table 4.2: Joint Model of Simulated Imputed Data (10% Missingness)

Joint Model	**AIC**	**BIC**	**LogLik**	
	-324.87	-283.424	173.438	
Longitudinal Process	**Estimate**	**Std.Error**	**Z-value**	**P-value**
Intercept	11.902	0.049	245.421	0.000
Observed time	0.047	0.001	47.700	0.000
Age	0.238	0.001	181.646	0.000
Sex (Male)	-0.014	0.017	-0.802	0.423
Event Process	**Estimate**	**Std.Error**	**Z-value**	**P-value**
Intercept	-86.581	6.040	-14.335	0.000
Sex (Male)	0.351	0.214	1.639	0.101
Age	-0.421	0.051	-8.325	0.000
Association	-2.655	0.243	-10.906	0.000

The Table 4.3 offers a comprehensive overview of a joint model analysis applied to simulated imputed data, with a notable 20% missingness rate. we observe an AIC value of 267.98, a BIC value of 172.26 , and the LogLik value of -2.262 suggests that the model adequately describes the observed data.

Transitioning to the longitudinal process, the intercept, standing at 11.873, signifies the estimated response variable value when all predictors are zero. Notably, predictors such as observed time and age exhibit statistically significant effects on the response variable,

given their low p-values ($p < 0.05$). This implies that as observed time and age increase, the response variable tends to change accordingly. However, the predictor "Sex (Male)" appears to lack statistical significance, as evidenced by its relatively high p-value of 0.711, indicating that gender might not significantly influence the response variable in this context.

Turning to the event process, a similar pattern emerges. The intercept, with a value of -91.508, denotes the baseline level of the response variable when all predictors are zero. Notably, while "Sex (Male)" and "Age" exhibit potential influences on the response variable, as suggested by their respective estimates and p-values, the former fails to reach statistical significance ($p = 0.120$). Conversely, the predictor "Association" displays a significant impact on the response variable, given its low p-value ($p < 0.05$).

Table 4.3: Joint Model of Simulated Imputed Data (20% Missingness)

Joint Model	**AIC**	**BIC**	**LogLik**	
	267.98	272.26	-2.262	
Longitudinal Process	**Estimate**	**Std.Error**	**Z-value**	**P-value**
Intercept	11.873	0.050	235.730	0.000
Observed time	0.046	0.001	39.422	0.000
Age	0.238	0.001	175.842	0.000
Sex (Male)	-0.007	0.018	-0.370	0.711
Event Process	**Estimate**	**Std.Error**	**Z-value**	**P-value**
Intercept	-91.508	6.518	-14.039	0.000
Sex (Male)	0.330	0.212	1.556	0.120
Age	-0.4131	0.056	-7.436	0.000
Association	-3.066	0.266	-11.515	0.000

In table 4.4, the joint model has an AIC of -308.229, a BIC of -266.778, and a LogLik of 165.115. The longitudinal process includes predictors such as "Intercept", "Observed time", "Age", and "Sex (Male)". "Intercept", "Observed time", and "Age" exhibit statistically significant positive effects on the longitudinal outcome, with p-values of 0.000, indicating strong associations. However, "Sex (Male)" does not significantly affect the longitudinal outcome, with a p-value of 0.422, suggesting that gender may not significantly influence the outcome in this context. The event process includes predictors such as "Intercept",

"Sex (Male)", "Age", and "Association". "Intercept" represents the baseline hazard, with a statistically significant negative effect on the event outcome (p-value = 0.000). "Sex (Male)" exhibits a marginally non-significant positive effect on the event outcome (p-value = 0.105), suggesting a trend towards significance. "Age" demonstrates a statistically significant negative effect on the event outcome (p-value = 0.000), indicating its importance as a predictor. "Association" shows a significant negative effect on the event outcome (p-value = 0.000), indicating a strong association between the longitudinal and event processes.

Table 4.4: Joint Model of Simulated Imputed Data (30% Missingness)

Joint Model	**AIC**	**BIC**	**LogLik**	
	-308.229	-266.778	165.115	
Longitudinal Process				
	Estimate	**Std.Error**	**Z-value**	**P-value**
Intercept	11.879	0.049	244.874	0.000
Observed time	0.047	0.001	47.713	0.000
Age	0.238	0.001	182.104	0.000
Sex (Male)	-0.014	0.017	-0.803	0.422
Event Process				
	Estimate	**Std.Error**	**Z-value**	**P-value**
Intercept	-84.217	6.052	-13.916	0.000
Sex (Male)	0.345	0.213	1.620	0.105
Age	-0.406	0.052	-7.857	0.000
Association	-2.6067	0.246	-10.609	0.000

Table 4.5 presents a joint model constructed with simulated imputed data addressing 40% missingness. This model integrates both longitudinal and event processes and provides insights into the relationships between predictor variables and the outcomes of interest. The AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are measures used for model comparison, with lower values indicating better fit. The LogLik (log-likelihood) represents the maximized log-likelihood of the model, serving as a measure of model adequacy.The joint model has an AIC of -261.518, a BIC of -223.835, and a LogLik of 140.759. Lower AIC and BIC values suggest better model fit, while a higher LogLik indicates better likelihood of the model given the data.

The longitudinal process includes predictors such as "Intercept", "Observed time", "Age", and "Sex (Female)". "Intercept", "Observed time", and "Age" exhibit statistically significant

positive effects on the longitudinal outcome, with p-values of 0.000, indicating strong associations. However, "Sex (Female)" does not significantly affect the longitudinal outcome, with a p-value of 0.427, suggesting that gender may not significantly influence the outcome in this context. The event process includes predictors such as "Intercept", "Sex (Female)", "Age", and "Association". "Intercept" represents the baseline hazard, with a statistically significant negative effect on the event outcome (p-value = 0.000). "Sex (Female)" and "Age" do not exhibit significant effects on the event outcome, with p-values of 0.441 and 0.000, respectively. However, "Association" shows a statistically significant negative effect on the event outcome (p-value = 0.000), indicating a strong association between the longitudinal and event processes.

Table 4.5: Joint Model of Simulated Imputed Data (40% Missingness)

Joint Model	**AIC**	**BIC**	**LogLik**	
	-261.518	-223.835	140.759	
Longitudinal Process	**Estimate**	**Std.Error**	**Z-value**	**P-value**
Intercept	11.897	0.047	255.724	0.000
Observed time	0.047	0.001	47.651	0.000
Age	0.237	0.001	183.095	0.000
Sex (Female)	0.014	0.017	1.795	0.427
Event Process	**Estimate**	**Std.Error**	**Z-value**	**P-value**
Intercept	-53.317	4.3163	-12.352	0.000
Sex (Female)	-0.160	0.207	-0.771	0.441
Age	-0.426	0.167	-5.564	0.000
Association	-2.930	0.191	-15.184	0.000

Figure 4.1 represents the missing plot of all the combined datasets with varying percentages of missingness to visualize the patterns of missingness in the combined dataset.

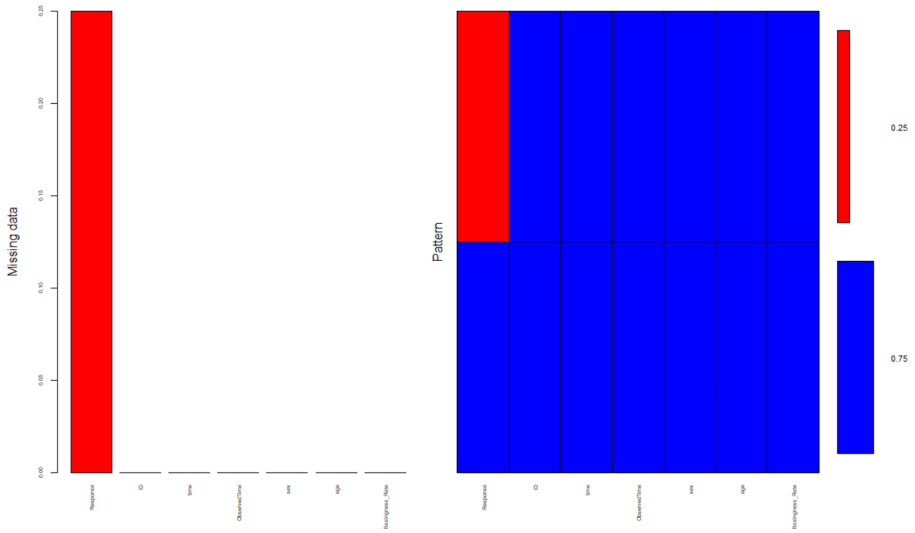

Figure 4.1: Missing Pattern Plot

4.3 Comparison of Joint Models

We evaluated the difference between the joint complete fitted model and the different varying proportions fitted joint models under two hypotheses; H_0: There is no model difference after imputation regardless of percentage missingness and H_1: There is a model difference after imputation regardless of percentage missingness.

From the results presented in Table 4.6; Table 4.7, Table 4.8, and Table 4.9 we can say that since the p-values are > 0.001 we fail to reject the null hypothesis thus concluding that there is no model difference after imputation using the different percentages at 10%, 20%, 30%, and 40%.

Table 4.6 provides a comparison between two joint models: one with complete data and another with a 10% missingness rate. Starting with the joint model with complete data, we observe an AIC value of -330.48 and a BIC value of -289.02, along with a LogLik value of 176.24. The joint model with a 10% missingness rate exhibits slightly higher AIC (-324.88) and BIC (-283.42) values, along with a marginally lower LogLik value of 173.44. The P-value associated with the comparison between the two models is 1.000, suggesting

that there is no statistically significant difference in model fit between the joint model with complete data and the one with a 10% missingness rate.

Table 4.6: Comparison of Joint Models of Complete Vs Imputed (10% Missingness)

	AIC	**BIC**	**LogLik**	**P-value**
Joint Model Complete	-330.48	-289.02	176.24	
Joint Model (10%)	-324.88	-283.42	173.44	1.000

Table 4.7 provides a comparison between two joint models: one with complete data and another with a 20% missingness rate. Starting with the joint model with complete data, we observe an AIC value of -330.48 and a BIC value of -289.02, along with a LogLik value of 176.24. The joint model with a 20% missingness rate exhibits slightly higher AIC (-326.52) and BIC (-267.98) values, along with a marginally lower LogLik value of 172.26. The P-value associated with the comparison between the two models is 1.000, suggesting that there is no statistically significant difference in model fit between the joint model with complete data and the one with a 20% missingness rate.

Table 4.7: Comparison of Joint Models of Complete Vs Imputed (20% Missingness)

	AIC	**BIC**	**LogLik**	**P-value**
Joint Model Complete	-330.48	-289.02	176.24	
Joint Model (20%)	-326.52	-267.98	172.26	1.000

Table 4.8 illustrates the comparison between joint models constructed with complete data and those derived from imputed data with 30% missingness. For the joint model constructed with complete data: AIC: -330.48, BIC: -289.02 and LogLik: 176.24. For the joint model constructed with imputed data (30% missingness): AIC: -308.23, BIC: -266.78, LogLik: 165.11 and P-value for model comparison: 1.000. The associated p-value for model comparison is 1.000, suggesting no statistically significant difference between the complete and imputed models.

Table 4.8: Comparison of Joint Models of Complete Vs Imputed (30% Missingness)

	AIC	BIC	LogLik	P-value
Joint Model Complete	-330.48	-289.02	176.24	
Joint Model (30%)	-308.23	-266.78	165.11	1.000

Table 4.9 presents a comparison between joint models constructed with complete simulated data and those built after imputation with 40% missingness. Both the joint model with complete data and the one with imputed data exhibit identical values for AIC, BIC, and LogLik. The p-value associated with this comparison is 1.000, indicating no statistically significant difference between the two models. The identical values for AIC, BIC, and LogLik suggest that both the joint model with complete data and the imputed model perform similarly in terms of goodness of fit. The lack of statistical significance (p = 1.000) further confirms that there is no discernible difference between the two models based on these evaluation metrics.

Table 4.9: Comparison of Joint Models of Complete Vs Imputed (40% Missingness)

	AIC	BIC	LogLik	P-value
Joint Model Complete	-330.48	-289.02	176.24	
Joint Model (40%)	-330.48	-289.02	176.24	1.000

4.4 Application to Real HIV Dataset

4.4.1 Exploratory Analysis

Figure 4.2 gives a visual representation of the number of missing observations in the response variable -viral load. From the figure, the number of missing observations is 73% of the total observations and 27% observed entries.

Figure 4.2: Missingness Heatmap for Viral Load

In Figure 4.3, we employed the "md.pattern" function from the mice package to display the count of variables with specific patterns of missing values. Each row represents a particular missing data pattern, while the columns indicate the number of missing values in the respective variables. The last column tallies the total number of values missing for each pattern. Notably, there are no missing values for variables such as Patient ID, episode, time, regimen line, HIV-TB infection, and event. Missingness is solely observed in the viral load variable.

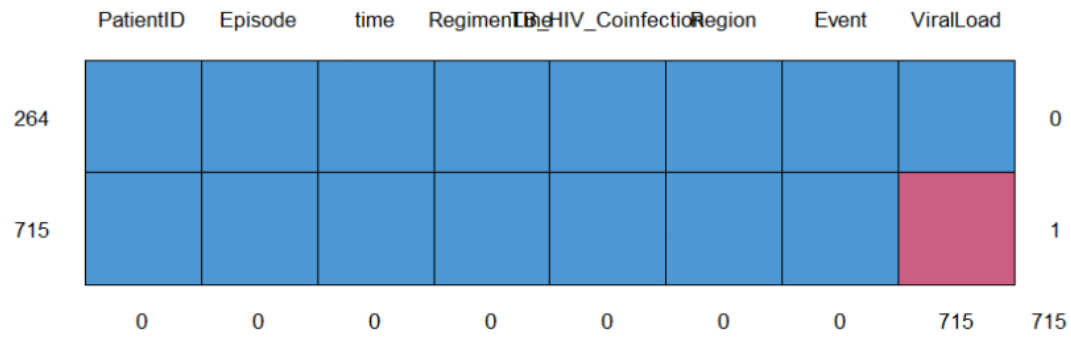

Figure 4.3: Missingness Data Pattern for all Variables

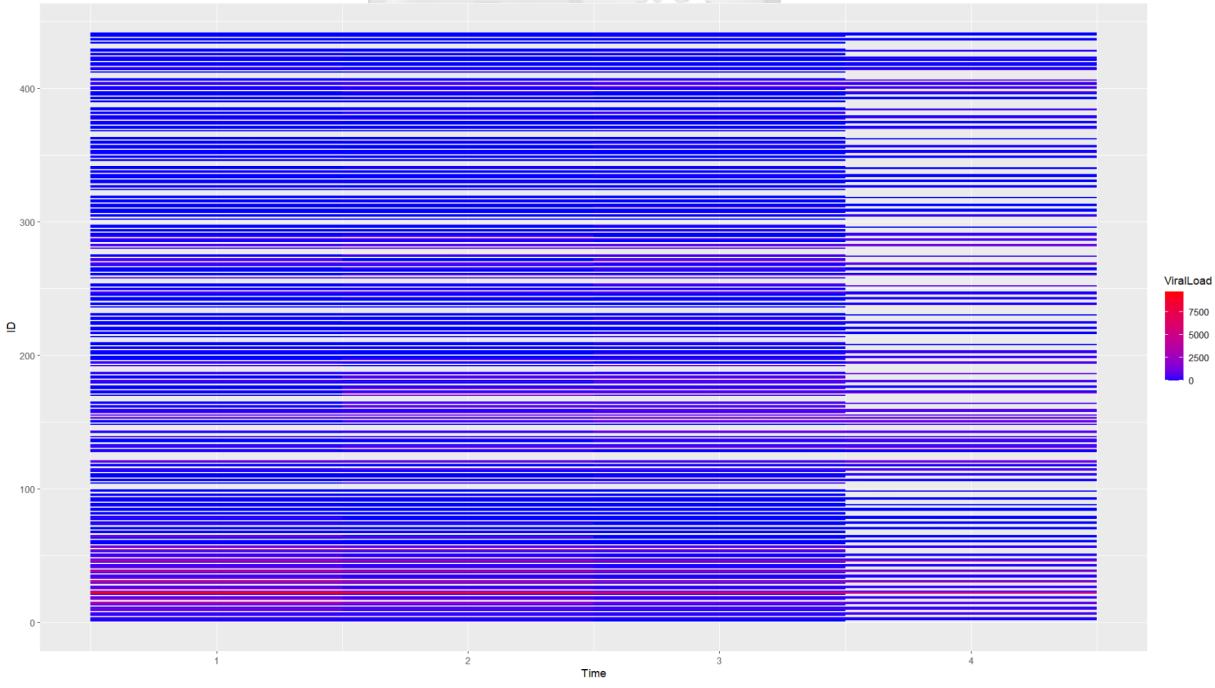

Figure 4.4: Missingness Data Pattern

Figure 4.4 illustrates the intensity of the patient's viral load against time. Blue color indicates low levels of Viral load whereas red indicates very high levels of viral load.

4.4.2 Linear Mixed Model

We start the analysis of the actual data using a linear mixed-effects model. As we discussed earlier, linear mixed models are applicable when the dependent variable is measured several times over a certain period for the same individual. Our sample consists of 271 patients and 979 observations. The dependent variable in this case is the viral load. Our interest is to assess the association between an individual's viral load and treatment (regimen line). We include other predictor variables namely episode (time when the viral load was measured), TB-HIV coinfection, and region (Nairobi or Kiambu) in the regression.

Our model assumes a random intercept. Table 4.10 shows the results from the linear mixed-effects model.

Table 4.10: Linear Mixed Model for Real HIV Data

Model Eva.Criteria	**AIC**	**BIC**	**LogLik**	
	14826.4	14865.44	-7405.2	
Longitudinal Process	**Estimate**	**Std.Error**	**t-value**	**P-value**
Intercept	91.323	60.523	1.509	0.132
Episode	-17.702	10.448	-1.694	0.091
Regimen line 2	-55.162	99.491	-0.554	0.580
Regimen line 3	593.436	112.550	5.272	0.000
TB-HIV Co-infection	160.676	63.989	2.511	0.123
Region (Nairobi)	171.149	66.757	2.564	0.011

The provided linear mixed model (LMM) results in Table 4.10 offer valuable insights into the longitudinal processes within real HIV data. The relatively low AIC and BIC values (14826.4 and 14865.44, respectively) suggest that the model adequately captures the underlying patterns in the data. The negative LogLik value (-7405.2) indicates the maximized log-likelihood of the model, indicating its appropriateness for the given data. The intercept represents the baseline value of the longitudinal outcome when all predictors are zero. The estimate for "Episode" suggests a potential decrease in the longitudinal outcome with each unit increase, although it does not reach statistical significance (p = 0.091). Similarly, the estimates for "Regimen line 2" and "episode" are negative, implying potential decreases in

the longitudinal viral load for patients on regimen line 2, though statistical significance is not achieved ($p > 0.05$). In contrast, the estimates for "Regimen line 3" and "Region (Nairobi)" are positive and statistically significant ($p < 0.05$), indicating higher viral load for patient's on regimen line 3 and those from Nairobi, respectively.

4.4.3 Survival Model

For the survival model, our interest is in determining the survival time or time to event. From our HIV data, the event of interest is death, under the three types of treatments (regimen line 1, regimen line 2, and regimen line 3). We use the proportional hazard model specified in the methodology section. The AIC value of 255.992 provides a measure of the model's goodness of fit. The estimates, along with their standard errors (Se), Z-values, and corresponding p-values, illuminate the influence of predictor variables on the hazard of the event of interest.

The negative estimate for "Regimen line 2" (-0.532) suggests a potential decrease in the hazard of the event for individuals on regimen line 2 compared to the reference group. However, this effect fails to reach statistical significance ($p = 0.294$). Conversely, the estimates for "Regimen line 3" and "TB-HIV Co-infection" exhibit statistically significant negative effects on the hazard of the event, with values of -2.377 and -1.510, respectively.

Table 4.11: Survival Model for the Real HIV Dataset

Cox-PH Model	**AIC**					
	255.992					
Event Process						
	Estimate	**Se (est)**	**Z-value**	**P-value**	**Lower L**	**Upper L**
Regimen line 2	-0.532	0.507	-1.049	0.294	0.2173	1.588
Regimen line 3	-2.377	0.586	-4.059	0.000	0.029	0.292
TB-HIV Co-infection	-1.510	0.383	-3.947	0.000	0.468	0.468

4.4.4 Joint Model for Real HIV Data

The analysis of joint model results in Table 4.12 offers a nuanced understanding of the dynamics underlying HIV-related outcomes, shedding light on the interplay between longitudinal processes and event occurrences. Moving to the longitudinal process estimates, the intercept term represents the baseline value of the longitudinal outcome, providing a reference point for interpretation. While the negative estimate for the "Episode" variable suggests a potential decrease in viral load with each unit increase. Similarly, the negative estimates for "Regimen line 2" imply potential decreases in viral load for individuals on regimen line 2 though statistical significance is not reached. Conversely, the positive and statistically significant estimates for "Regimen line 3" and "Region (Nairobi)" indicate higher viral load for individuals on regimen line 3 and those from Nairobi, respectively.

In the event process, the intercept term represents the baseline hazard, providing insights into the probability of event occurrence in the absence of other predictors. The negative and statistically significant estimates for "Regimen line 3" and "Association" suggest a decreased hazard rate for individuals on regimen line 3 and a negative association between the longitudinal and event processes, respectively.

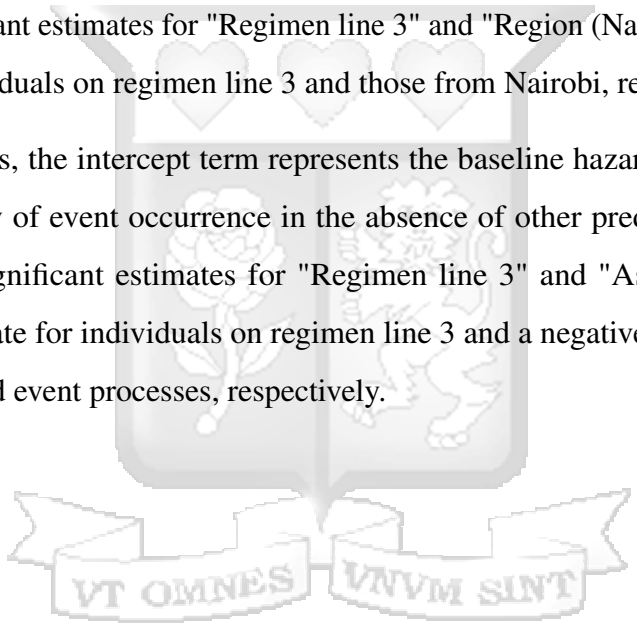

Table 4.12: Joint Model of Real HIV Data

Joint Model	**AIC**	**BIC**	**LogLik**	
	14991.62	15042.05	-7481.81	
Longitudinal Process				
	Estimate	**Std.Error**	**Z-value**	**P-value**
Intercept	91.748	60.324	1.521	0.128
Episode	-19.068	10.420	-1.830	0.067
Regimen line 2	-54.997	98.940	-0.556	0.578
Regimen line 3	593.672	112.243	5.289	0.000
TB-HIV Co-infection	172.206	63.635	2.706	0.007
Region (Nairobi)	164.300	65.987	2.490	0.013
Event Process				
	Estimate	**Std.Error**	**Z-value**	**P-value**
Intercept	-27.439	3.706	-7.405	0.000
Regimen line 2	-0.180	0.586	-0.308	0.758
Regimen line 3	-1.903	0.670	-2.720	0.000
TB-HIV Co-infection	-0.734	0.499	-1.472	0.141
Association	-0.003	0.000	-5.320	0.000

4.4.5 Model Selection and Accuracy

We must ascertain the most suitable model for our analysis among the three options: longitudinal, survival, and joint. Once again, we rely on the AIC values from each model. Specifically, the AIC for the longitudinal model stands at 14865.40, while for the survival model, it is 255.992. Adding these two AIC values together yields a total of 15121.392. In contrast, the joint model exhibits an AIC of 14991.62. Based on these comparisons, we determine that the joint model offers a better fit for our data.

4.5 Diagnostics

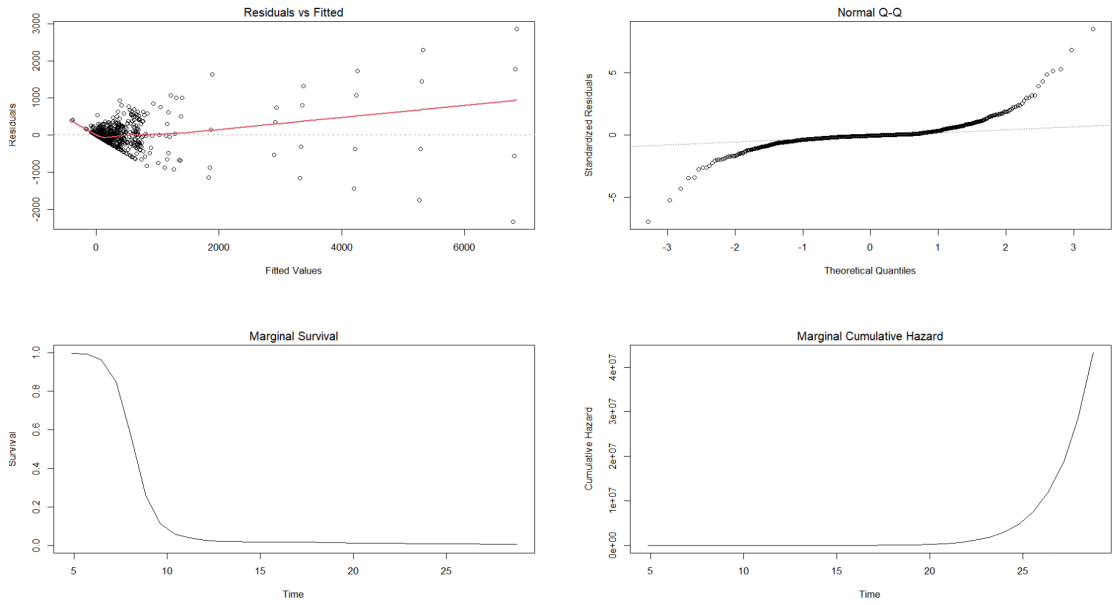

Figure 4.5: Residuals for the Joint Model Fit

We conducted an analysis to assess the effectiveness of the joint model, focusing on its diagnostic performance. Illustrated in Figure 4.5 are diagnostic visualizations representing the joint model's fit to an authentic HIV dataset. The top-right plot presents a normal Q-Q plot, showcasing standardized residuals specific to each subject's longitudinal process. On the top-left plot, subject-specific residuals for the longitudinal process are plotted against their predicted values. Furthermore, the bottom-right plot offers an estimation of the cumulative risk function for the event process, while the bottom-left plot illustrates an estimation of the survival function for the event process.

Analyzing the Marginal Cumulative Hazard curve reveals a consistent upward trend over time, indicating an exponential increase in the cumulative hazard rate as time progresses, suggesting disease progression among the patients in the study. Conversely, the Marginal Survival curves exhibit a declining trend, signifying a decrease in survival rate as time elapses.

While the Q-Q plot shows that most residuals conform to a linear pattern, there are some outliers deviating from this pattern. A notable challenge observed in the Q-Q plot and residual plot is that the distribution of residuals for the longitudinal process is influenced by dropout events. When patients experience events such as death, the collection of repeated measurements ceases, leading to a change in their statistical distribution Under the dropout mechanism in the joint models, there is the characteristic of a non-random nature (Verbeke et al., 2008). This suggests that residual plots derived solely from observed data may be deceptive since these residuals are not anticipated to demonstrate typical properties like zero mean and independence.

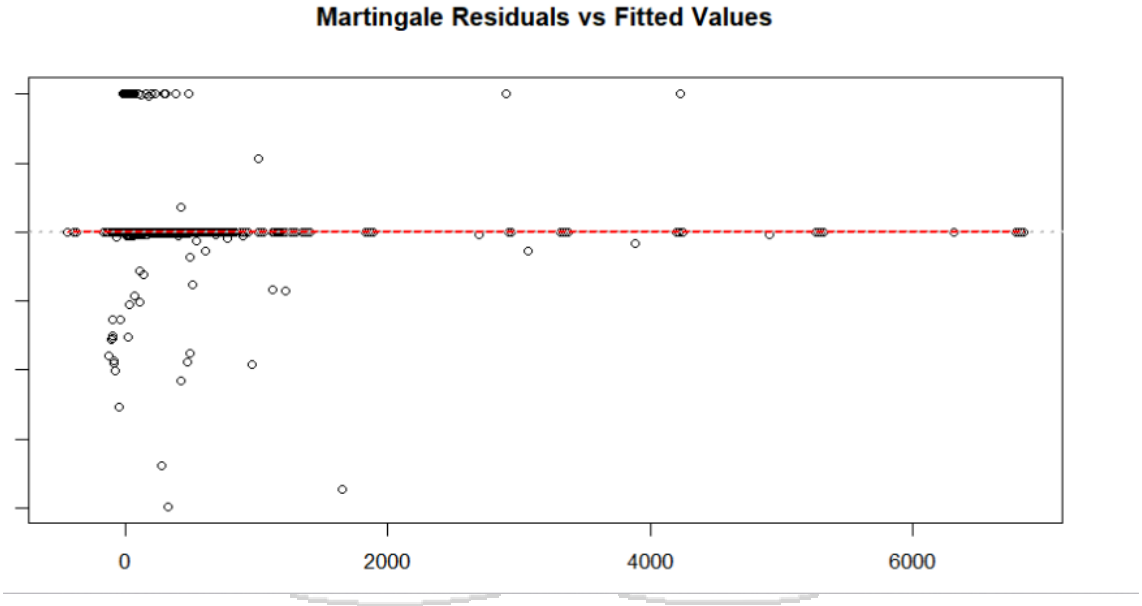

Figure 4.6: Martingale Residuals vs Fitted Values

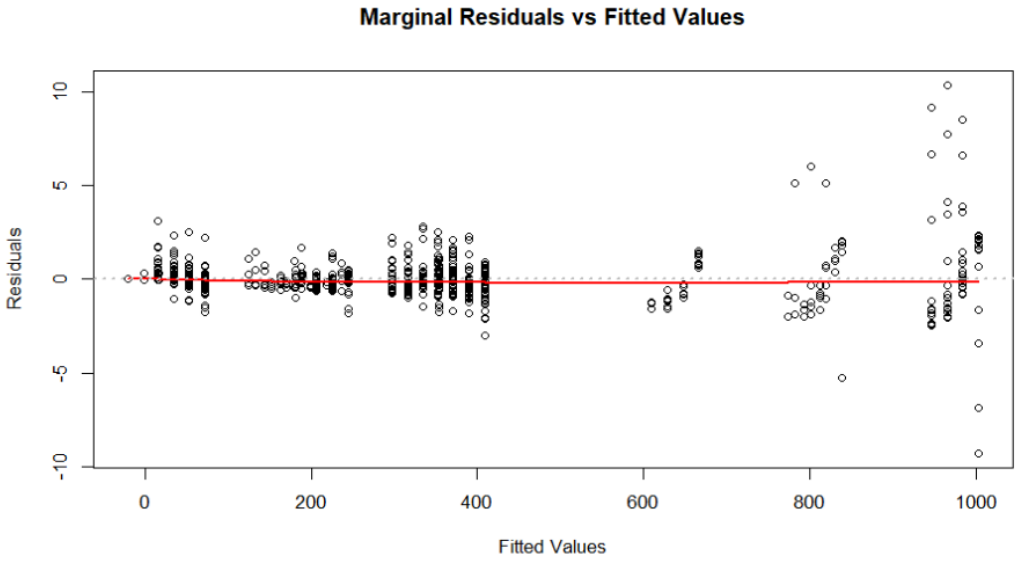

Figure 4.7: Marginal Residuals vs Fitted Values

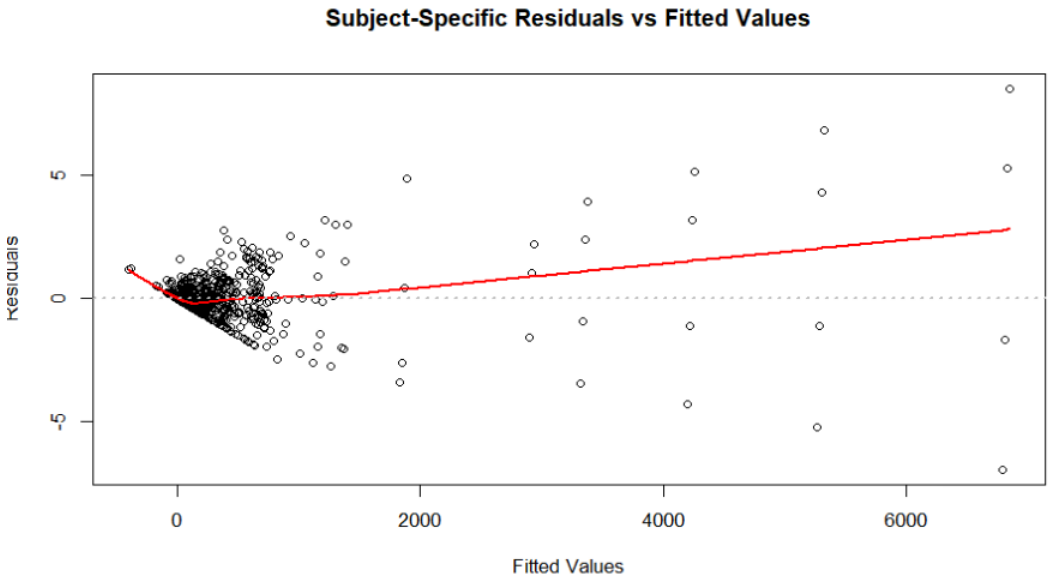

Figure 4.8: Subject-Specific Residuals vs Fitted Values

Figures 4.6; Figure 4.7, and Figure 4.8 depict additional residual plots. Initially, we computed the particular residuals of concern and subsequently plotted them against their corresponding fitted covariates. Figure 4.8 illustrates standardized subject-specific residuals,In Figure 4.7,

standardized marginal residuals are depicted against the fitted values for the longitudinal process, accompanied by martingale residuals. Remarkably, a distinct pattern emerges in Figure 4.7, revealing an abundance of positive residuals, especially prominent for lower fitted values.

Chapter 5

Discussion and Conclusion

5.1 Introduction

This chapter gives an illustration of the study findings with reference to what other scholars have done in similar area of study, strengths and limitations of the study, recommendations and conclusion.

5.2 Discussion

In prior research, scholars have typically employed distinct models for analyzing longitudinal data and time-to-event data. However, this separate analysis approach often faces significant hurdles. One common issue is the presence of correlations between the repeated measurements and the time-to-event data within the explanatory variables of both the survival model and the longitudinal model. These associations are not accounted for when using separate sub-models, potentially leading to biased results, particularly in longitudinal studies where dropouts may carry informative patterns. To address these challenges, we have introduced a joint model that concurrently addresses the relationship between survival and longitudinal data, as well as the issue of missing data. Joint models are increasingly recognized in clinical trials for their ability to provide more robust estimations of treatment effects on both time-to-event outcomes and longitudinal markers. Moreover, they help alleviate bias in estimating the overall treatment effect by considering both survival and longitudinal marker outcomes simultaneously.

In our study we started by simulating datasets with different proportions of missingness. The motivation of doing simulations is that data simulation facilitates the exploration of hypothetical scenarios and the assessment of statistical methods under various conditions. Researchers can systematically manipulate key parameters, such as sample size, effect size, and distributional assumptions, to investigate the robustness and performance of statistical techniques (Collins et al., 2001). By conducting simulation studies, researchers gain insights into the strengths, limitations, and assumptions underlying different analytical approaches, aiding in the selection of appropriate methods for real data analysis (Austin 2011). After the simulation process, we used multiple imputation then fitted joint models of the 4 datasets under the assumption of MAR. By generating multiple imputed datasets, uncertainty arising from the imputation process can be quantified, leading to more reliable estimates and standard errors. Rubin's rules can then be applied to appropriately combine the results from the analyses of imputed datasets, enhancing the validity of statistical inference.

For example, a study by (Buuren and Groothuis-Oudshoorn, 2011) evaluated the performance of multiple imputation techniques in handling missing data. They concluded that multiple imputation methods, particularly when implemented under MAR, yield unbiased estimates and valid statistical inferences when compared to complete case analysis or single imputation methods. Similarly, a review by (Horton and Kleinman, 2007) emphasized the importance of correctly specifying the imputation model and adherence to the MAR assumption. When these conditions are met, multiple imputation methods have been shown to produce accurate results and facilitate valid comparisons across imputed datasets.

Results of comparison of the fitted joint models using ANOVA highlighted that the p-value associated with the comparison was 1.000, indicating no statistically significant difference between the two models. These findings imply that the imputation process effectively captures data without significantly altering model outcomes, as evidenced by the similarity in model fit metrics. Researchers can have confidence in utilizing the imputed data for model building and analysis, as it produces comparable results to those obtained with complete data. The absence of a significant difference between the models suggests that the imputation

method employed is reliable and adequately addresses missingness in the dataset under joint models.

In the examination of the HIV data, we contrasted separate analyses of the longitudinal model and the survival model with the joint model. From the LME model, the findings highlight the influence of various factors, such as treatment regimen and geographic location, on longitudinal HIV-related outcomes. Understanding the factors driving longitudinal HIV outcomes is crucial for developing targeted interventions and personalized treatment approaches. The significant positive effect of Regimen line 3 on viral load suggests the importance of specific treatment regimens in managing HIV patients in Kenya. This finding underscores the need for effective antiretroviral therapy (ART) protocols, potentially indicating the superior efficacy or accessibility of certain treatment lines. The observed trend towards significance for TB-HIV co-infection highlights the complex interplay between tuberculosis (TB) and HIV in Kenya. Co-infection presents unique challenges in diagnosis, treatment, and disease management, necessitating integrated healthcare approaches (Ministry of Health, Kenya, 2019). Additionally, the regional disparities highlighted by the positive estimate for Nairobi underpins the importance of considering geographic factors in HIV care and resource allocation.

Under the survival model, the implications of the Cox proportional hazards (Cox-PH) model results, particularly concerning the event process predictors in HIV/AIDS management, are crucial for informing healthcare strategies and interventions aimed at reducing adverse outcomes among affected individuals. The significant negative association between Regimen line 3 and the hazard of the event suggests that individuals receiving this treatment regimen experience a lower risk of adverse outcomes. This finding underscores the importance of effective antiretroviral therapy (ART) protocols in improving clinical outcomes and prolonging survival among HIV-infected individuals (Ford, N., Darder, M., & Spelman, T., 2010). However, the lack of statistical significance for Regimen line 2 implies that certain treatment lines may not offer significant benefits in increasing a patient's survival time. This highlights the need for further research to evaluate the efficacy and tolerability of different

ART regimens, particularly in resource-limited settings where treatment options may be limited (Kibicho, W., Mburu, G., Nyanaro, G., Tiono, A. B., & Bekele, A., 2016).

The significant negative association between TB-HIV co-infection and the hazard of the event suggests that individuals with dual infection experience a higher risk of adverse outcomes compared to those without co-infection. This finding underscores the importance of integrated TB-HIV care and management strategies in improving clinical outcomes and reducing mortality among co-infected individuals (Suthar et al., 2012). Integrated care approaches, including collaborative TB-HIV activities, early detection, and treatment of both diseases, are essential for optimizing clinical outcomes and reducing the burden of TB-HIV co-infection in high-prevalence settings like Kenya .

The implications of the Joint Model results for real HIV data in Kenya are significant for informing healthcare policies, clinical practices, and public health interventions. The significant positive association between Regimen line 3 and both longitudinal outcomes and reduced hazard of adverse events suggests the efficacy of this treatment regimen in improving patient's outcomes. This finding emphasizes the importance of optimizing treatment protocols and ensuring access to effective antiretroviral therapy (ART) regimens in Kenya (World Health Organization, 2016).

The significant positive association between TB-HIV co-infection and longitudinal outcomes, coupled with a potentially protective effect on the hazard of adverse events, highlights the importance of integrated TB-HIV care and management strategies in Kenya. Coordinated efforts to address both diseases concurrently are crucial for improving patient outcomes and reducing morbidity and mortality rates (Ministry of Health, Kenya, 2018). The significant negative association between the association parameter and the hazard of adverse events suggests the importance of comprehensive care models that integrate longitudinal monitoring of viral load with event occurrence surveillance (time to event/survival time). Healthcare systems should adopt holistic approaches that consider both longitudinal health trajectories and acute event occurrences to provide patient-centered care and optimize health outcomes (Mwangi, L. W., Angwenyi, V., Ogendo, A., & Mbindyo, P., 2020).

5.2.1 Strengths and Limitations

Joint modeling offers the ability to account for missingness mechanisms by providing a more realistic representation of the underlying data generating process. By incorporating information about why data are missing, joint modeling helps mitigate bias introduced by missing data and enhances the validity of study findings. This is particularly crucial in longitudinal studies where missing data are common due to participant attrition or intermittent data collection. Moreover, joint modeling increases statistical power by simultaneously analyzing longitudinal and survival outcomes within the same model. This enhances precision and reliability compared to separate analyses, especially in studies with limited sample sizes or high levels of missing data. Additionally, joint models are adept at handling correlated data structures inherent in longitudinal studies, allowing for the incorporation of within-subject correlation and leading to more efficient parameter estimation and improved model fit.

Despite its numerous strengths, joint modeling is limited in its complexity of implementation, which requires a thorough understanding of survival analysis, longitudinal data analysis techniques, and specialized software for fitting joint models. Interpretation of joint modeling results can also be challenging, especially when considering the interplay between longitudinal trajectories and survival outcomes. Effective communication of findings to diverse stakeholders, including clinicians and policymakers, may require careful consideration and clear explanation of model assumptions and results to ensure accurate interpretation.

5.3 Recommendations

5.3.1 Recommendation for Further Studies

Based on the discussion surrounding joint modeling in clinical research, particularly in the context of HIV studies more validation studies can be conducted to assess the performance and robustness of joint modeling approaches in diverse clinical settings and populations, Investigate additional longitudinal outcome measures beyond viral load, such as CD4 cell

count, clinical symptoms, or quality of life indicators, to comprehensively evaluate the impact of treatment regimens and interventions on HIV disease progression and patient well-being and investigate the impact of different missing data mechanisms (e.g., missing completely at random, missing at random, missing not at random) on joint modeling results and develop strategies to address missingness on both the dependent and predictor variables.

5.3.2 Recommendations for Policy

Based on the findings and implications of joint modeling in clinical research, particularly in HIV studies, Policy efforts should prioritize the integration of specific treatment regimens that have demonstrated significant positive associations with longitudinal outcomes and reduced hazards of adverse events. This entails ensuring access to effective antiretroviral therapy (ART) regimens, which has shown superior efficacy in managing HIV patients: continued support for research and innovation in statistical methodologies, such as joint modeling, is essential. Policy efforts should facilitate collaboration between researchers, clinicians, and policymakers to enhance the understanding and implementation of advanced analytical techniques that address the complex challenges of clinical data analysis; enhanced Surveillance and Monitoring: Policies should prioritize the implementation of comprehensive care models that integrate longitudinal monitoring of viral load with event occurrence surveillance. This entails adopting holistic approaches that consider both longitudinal health trajectories and acute event occurrences to provide patient-centered care and optimize health outcomes. By implementing these recommendations, policymakers can contribute to improving the quality of HIV care and treatment services, reducing disease burden, and ultimately advancing progress towards achieving HIV-related goals and targets..

5.4 Conclusion

Joint modeling offers several advantages, including accounting for missingness mechanisms, increasing statistical power, handling correlated data, and incorporating time-varying effects.

By simultaneously analyzing longitudinal and survival outcomes, joint models provide a holistic framework for understanding disease progression, treatment efficacy, and patient outcomes. Additionally, through the simulation of datasets and multiple imputation techniques, joint modeling facilitates robust estimations of treatment effects and enhances the validity of statistical inference. However, joint modeling also presents challenges, such as complexity of implementation, assumptions, computational intensity, and interpretation difficulties. Addressing these challenges requires expertise, careful model diagnostics, and adherence to assumptions, ensuring the reliability and validity of study findings.

The implications of joint modeling in clinical research are profound, informing healthcare policies, clinical practices, and public health interventions. From optimizing treatment strategies to promoting integrated care approaches and addressing regional disparities, joint modeling provides insights that drive evidence-based decision-making and improve patient outcomes. In conclusion, joint modeling represents a transformative approach in clinical research, offering a comprehensive framework for analyzing complex data structures and addressing missing data challenges. By leveraging its strengths, addressing its limitations, and pursuing further research, joint modeling has the potential to revolutionize clinical research practices, drive evidence-based healthcare decision-making, and improve patient outcomes on a global scale.

References

Achia, T., Cervantes, I. F., Stupp, P., Musingila, P., Muthusi, J., Waruru, A., Schmitz, M., Bronson, M., Chang, G., Bore, J., Kingwara, L., Mwalili, S., Muttunga, J., Gitonga, J., De Cock, K. M., and Young, P. (2022). Methods for conducting trends analysis: roadmap for comparing outcomes from three national HIV Population-based household surveys in Kenya (2007, 2012, and 2018). *BMC Public Health*, 22(1):1337.

Bradburn, M. J., Clark, T. G., Love, S. B., and Altman, D. G. (2003). Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. *British Journal of Cancer*, 89(3):431–436.

Brooks, S., editor (2011). *Handbook of Markov chain Monte Carlo*. CRC Press, Boca Raton. OCLC: 740896891.

Buuren, S. V. and Groothuis-Oudshoorn, K. (2011). **mice** : Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software*, 45(3).

Carpenter, J. R. and Kenward, M. G. (2013). *Multiple Imputation and its Application*. Wiley, 1 edition.

Carroll, O. U., Morris, T. P., and Keogh, R. H. (2020). How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review. *BMC Medical Research Methodology*, 20(1):134.

Chen, Q., May, R. C., Ibrahim, J. G., Chu, H., and Cole, S. R. (2014). Joint modeling of longitudinal and survival data with missing and left-censored time-varying covariates. *Statistics in Medicine*, 33(26):4560–4576.

Clark, T. G., Bradburn, M. J., Love, S. B., and Altman, D. G. (2003). Survival Analysis Part I: Basic concepts and first analyses. *British Journal of Cancer*, 89(2):232–238.

Collins, L. M., Schafer, J. L., and Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4):330–351.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the *EM* Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Dong, Y. and Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, 2(1):222.

Efromovich, S. (2018). *Missing and modified data in nonparametric estimation*. CRC Press, Boca Raton. OCLC: 1029246950.

Fitzmaurice, G. M., editor (2009). *Longitudinal data analysis*. Chapman & Hall/CRC handbooks of modern statistical methods. CRC Press, Boca Raton. OCLC: 227328498.

Ford, N., Darder, M., & Spelman, T. (2010). Systematic review of the efficacy and tolerability of antiretroviral therapy in HIV-1-infected patients in resource-limited settings. *Journal of Acquired Immune Deficiency Syndromes, 53(1), 13-19.*

Garcia, T. P. and Marder, K. (2017). Statistical Approaches to Longitudinal Data Analysis in Neurodegenerative Diseases: Huntington's Disease as a Model. *Current Neurology and Neuroscience Reports*, 17(2):14.

Gibbons, R. D., Hedeker, D., and DuToit, S. (2010). Advances in Analysis of Longitudinal Data. *Annual Review of Clinical Psychology*, 6(1):79–107.

Harel, O., Pellowski, J., and Kalichman, S. (2012). Are We Missing the Importance of Missing Values in HIV Prevention Randomized Clinical Trials? Review and Recommendations. *AIDS and Behavior*, 16(6):1382–1393.

Heitjan, D. F. (2009). Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis by DANIELS, M. J. and HOGAN, J. W. *Biometrics*, 65(1):328–328.

Horton, N. J. and Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1):79–90.

Ibrahim, J. G., Chu, H., and Chen, L. M. (2010). Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data. *Journal of Clinical Oncology*, 28(16):2796–2801.

Ibrahim, J. G. and Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *TEST*, 18(1):1–43.

Jahagirdar, D., Walters, M. K., Novotney, A., Brewer, E. D., Frank, T. D., and Carter (2021). Global, regional, and national sex-specific burden and control of the HIV epidemic, 1990–2019, for 204 countries and territories: the Global Burden of Diseases Study 2019. *The Lancet HIV*, 8(10):e633–e651.

Jakubowski, A., Kabami, J., Balzer, L. B., Ayieko, J., Charlebois, E. D., Owaraganise, A., Marquez, C., Clark, T. D., Black, D., Shade, S. B., Chamie, G., Cohen, C. R., Bukusi, E. A., Kamya, M. R., Petersen, M., Havlir, D. V., and Thirumurthy, H. (2022). Effect of universal HIV testing and treatment on socioeconomic wellbeing in rural Kenya and Uganda: a cluster-randomised controlled trial. *The Lancet Global Health*, 10(1):e96–e104.

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5):402.

Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kharsany, A. B. and Karim, Q. A. (2016). HIV Infection and AIDS in Sub-Saharan Africa: Current Status, Challenges and Opportunities. *The Open AIDS Journal*, 10(1):34–48.

Kibicho, W., Mburu, G., Nyanaro, G., Tiono, A. B., & Bekele, A. (2016). Evaluation of efficacy and tolerability of different antiretroviral therapy regimens in resource-limited settings: A systematic review. *International Journal of STD & AIDS, 27(6), 437-445.*

Kim, A. A., Hallett, T., Stover, J., Gouws, E., Musinguzi, J., Mureithi, P. K., Bunnell, R., Hargrove, J., Mermin, J., Kaiser, R. K., Barsigo, A., and Ghys, P. D. (2011). Estimating HIV Incidence among Adults in Kenya and Uganda: A Systematic Comparison of Multiple Methods. *PLoS ONE*, 6(3):e17535.

Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine*, 7(1-2):305–315.

Lebowitz, M. D. (1996). Age, Period, and Cohort Effects: Influences on Differences between Cross-sectional and Longitudinal Pulmonary Function Results. *American Journal of Respiratory and Critical Care Medicine*, 154(6_pt_2):S273–S277.

Lesko, C. R., Edwards, J. K., Moore, R. D., and Lau, B. (2016). A longitudinal, HIV care continuum: 10-year restricted mean time in each care continuum stage after enrollment in care, by history of IDU. *AIDS*, 30(14):2227–2234.

Levy, B., Correia, H. E., Chirove, F., Ronoh, M., Abebe, A., Kgosimore, M., Chimbola, O., Machingauta, M. H., Lenhart, S., and White, K. A. J. (2021). Modeling the Effect of HIV/AIDS Stigma on HIV Infection Dynamics in Kenya. *Bulletin of Mathematical Biology*, 83(5):55.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lim, H. J., Mondal, P., and Skinner, S. (2013). Joint modeling of longitudinal and event time data: application to HIV study. *Journal of Medical Statistics and Informatics*, 1(1):1.

Little, R. and Rubin, D. (2019). *Statistical Analysis with Missing Data, Third Edition*. Wiley Series in Probability and Statistics. Wiley, 1 edition.

Little, R. J. A. and Rubin, D. B. (1989). The Analysis of Social Science Data with Missing Values. *Sociological Methods & Research*, 18(2-3):292–326.

Little, R. J. A. and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons, Hoboken, second edition edition. OCLC: 934777983.

Ministry of Health, Kenya (2019). Kenya tuberculosis prevalence survey 2016. Technical report.

Molenberghs, G., Fitzmaurice, G. M., Kenward, M. G., Tsiatis, A. A., and Verbeke, G. (2015). *Handbook of missing data methodology*. CRC Press, Taylor & Francis Group, Boca Raton. OCLC: 1164748943.

Muturi, N. W. (2005). Communication for HIV/AIDS Prevention in Kenya: Social–Cultural Considerations. *Journal of Health Communication*, 10(1):77–98.

Mwangi, J., Miruka, F., Mugambi, M., Fidhow, A., Chepkwony, B., Kitheka, F., Ngugi, E., Aoko, A., Ngugi, C., and Waruru, A. (2022). Characteristics of users of HIV self-testing in Kenya, outcomes, and factors associated with use: results from a population-based HIV impact assessment, 2018. *BMC Public Health*, 22(1):643.

Mwangi, L. W., Angwenyi, V., Ogendo, A., & Mbindyo, P. (2020). Integrating TB-HIV care: Patient-centered approach. *Patient-centered approach. Journal of Public Health, 42(3), 515-522.*

Ng'ang'a, A., Waruiru, W., Ngare, C., Ssempijja, V., Gachuki, T., Njoroge, I., Oluoch, P., Kimanga, D. O., Maina, W. K., Mpazanje, R., and Kim, A. A. (2014). The Status of HIV Testing and Counseling in Kenya: Results From a Nationally Representative Population-Based Survey. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 66(Supplement 1):S27–S36.

Rajan, K. B. and Leurgans, S. E. (2010). Joint modeling of missing data due to non-participation and death in longitudinal aging studies. *Statistics in Medicine*, 29(21):2260–2268.

Rapaport, S. F., Peer, A. D., Viswasam, N., Hahn, E., Ryan, S., Turpin, G., Lyons, C. E., Baral, S., and Hansoti, B. (2023). Implementing HIV Prevention in Sub-Saharan Africa: A Systematic Review of Interventions Targeting Systems, Communities, and Individuals. *AIDS and Behavior*, 27(1):150–160.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: with applications in R*. Number 6 in Chapman & Hall/CRC biostatistics series. CRC Press, Boca Raton.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rodríguez, D. C., Mohan, D., Mackenzie, C., Wilhelm, J., Eze-Ajoku, E., Omondi, E., Qiu, M., and Bennett, S. (2021). Effects of transition on HIV and non-HIV services and health systems in Kenya: a mixed methods evaluation of donor transition. *BMC Health Services Research*, 21(1):457.

Rosenberg, N. E., Shook-Sa, B. E., Liu, M., Stranix-Chibanda, L., Yotebieng, M., Sam-Agudu, N. A., Hudgens, M. G., Phiri, S. J., Mutale, W., Bekker, L.-G., Moyo, S., Zuma, K., Charurat, M. E., Justman, J., and Chi, B. H. (2023). Adult HIV-1 incidence across 15 high-burden countries in sub-Saharan Africa from 2015 to 2019: a pooled analysis of nationally representative data. *The Lancet HIV*, 10(3):e175–e185.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley series in probability and mathematical statistics. Wiley, New York.

Schober, P. and Vetter, T. R. (2018). Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare. *Anesthesia & Analgesia*, 127(3):792–798.

Shaw, R. G. and Mitchell-Olds, T. (1993). Anova for Unbalanced Data: An Overview. *Ecology*, 74(6):1638–1645.

Suthar, A. B., Lawn, S. D., Del Amo, J., Getahun, H., Dye, C., Sculier, D., Sterling, T. R., Chaisson, R. E., Williams, B. G., Harries, A. D., and Granich, R. M. (2012). Antiretroviral Therapy for Prevention of Tuberculosis in Adults with HIV: A Systematic Review and Meta-Analysis. *PLoS Medicine*, 9(7):e1001270.

Tiruneh, F., Chewaka, L., and Abdissa, D. (2021). Statistical Joint Modeling for Predicting the Association of CD4 Measurement and Time to Death of People Living with HIV Who Enrolled in ART, Southwest Ethiopia. *HIV/AIDS - Research and Palliative Care*, Volume 13:73–79.

Trickey, A., May, M. T., Vehreschild, J.-J., Obel, N., Gill, M. J., Crane, H. M., Boesecke, C., Patterson, S., Grabar, S., Cazanave, C., Cavassini, M., Shepherd, L., Monforte, A. d., van Sighem, A., Saag, M., Lampe, F., Hernando, V., Montero, M., Zangerle, R., Justice, A. C., Sterling, T., Ingle, S. M., and Sterne, J. A. C. (2017). Survival of HIV-positive patients starting antiretroviral therapy between 1996 and 2013: a collaborative analysis of cohort studies. *The Lancet HIV*, 4(8):e349–e356.

Verbeke, G., Molenberghs, G., and Beunckens, C. (2008). Formal and Informal Model Selection with Incomplete Data. Publisher: [object Object] Version Number: 1.

Vourli, G., Katsarolis, I., Pantazis, N., and Touloumi, G. (2021). HIV continuum of care: expanding scope beyond a cross-sectional view to include time analysis: a systematic review. *BMC Public Health*, 21(1):1699.

World Health Organization (2016). *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach*. World Health Organization, Geneva, 2nd ed edition.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339.

Young, P. W., Musingila, P., Kingwara, L., Voetsch, A. C., Zielinski-Gutierrez, E., Bulterys, M., Kim, A. A., Bronson, M. A., Parekh, B. S., Dobbs, T., Patel, H., Reid, G., Achia, T., Keter, A., Mwalili, S., Ogollah, F. M., Ondondo, R., Longwe, H., Chege, D., Bowen, N., Umuro, M., Ngugi, C., Justman, J., Cherutich, P., and De Cock, K. M. (2023). HIV Incidence, Recent HIV Infection, and Associated Factors, Kenya, 2007–2018. *AIDS Research and Human Retroviruses*, 39(2):57–67.

Yu, T. (2019). Joint modelling of complex longitudinal and survival data, with applications to HIV studies. Publisher: University of British Columbia Version Number: 1.

Appendix A

Similarity Report

Tessie Atieno Akoko_Final.pdf

ORIGINALITY REPORT

10%	9%	10%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	eio.usc.es Internet Source	2%
2	digitalcommons.montclair.edu Internet Source	1%
3	repub.eur.nl Internet Source	1%
4	Ofer Harel, Jennifer Pellowski, Seth Kalichman. "Are We Missing the Importance of Missing Values in HIV Prevention Randomized Clinical Trials? Review and Recommendations", AIDS and Behavior, 2012 Publication	1%
5	livrepository.liverpool.ac.uk Internet Source	<1%
6	open.library.ubc.ca Internet Source	<1%
7	Yuriko Takeda, Toshihiro Misumi, Kouji Yamamoto. "Joint Models for Incomplete Longitudinal Data and Time-to-Event Data", Mathematics, 2022	<1%

	Publication	
8	m.moam.info Internet Source	<1 %
9	Tanya P. Garcia, Karen Marder. "Statistical Approaches to Longitudinal Data Analysis in Neurodegenerative Diseases: Huntington's Disease as a Model", Current Neurology and Neuroscience Reports, 2017 Publication	<1 %
10	faculty.washington.edu Internet Source	<1 %
11	hdl.handle.net Internet Source	<1 %
12	article.sapub.org Internet Source	<1 %
13	journals.lww.com Internet Source	<1 %
14	Submitted to University of Lancaster Student Paper	<1 %
15	Perlman, M.D.. "Lattice conditional independence models for contingency tables with non-monotone missing data patterns", Journal of Statistical Planning and Inference, 19990701 Publication	<1 %
	pergamos.lib.uoa.gr	

Appendix B

Ethical Clearance Confirmation

Strathmore UNIVERSITY

17th March 2023

Ms Akoko Tessie Atieno,
tessie.akoko@strathmore.edu

Dear Ms Akoko,

RE: Dealing with Missing Data under Joint Modelling: Application to HIV Data

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU- master's** research proposal. Your application reference number is **SU-ISERC1620/23.** The approval period is from **17th March 2023 to 16th March 2024.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 48 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 48 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-ISERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

for: **Dr Ben Ngoye,**
Secretary; SU-ISERC

Cc: Mr Ambrose Rachier,
Chairperson; SU-ISERC

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
17-Mar-2023
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

Appendix C

R Code

```
1
2
3  #Loading the required packages for data analysis.
4  library(JM)
5  library(survival)
6  library(nlme)
7  library(visdat)
8  library(mice)
9  library(ggplot2)
10 library(VIM)
11 library(survminer)
12
13 #Loading the datasets used for simulation analysis.
14 simlong <- read.csv("C:/Users/Tessie␣Akoko/OneDrive␣-␣CGIAR/Joint␣
       Modelling/SimLong.csv")
15 head(simlong)
16 sim_Surv <- read.csv("C:/Users/Tessie␣Akoko/OneDrive␣-␣CGIAR/Joint␣
       Modelling/Sim_Surv.csv")
17 head(sim_Surv)
18
19 #linear mixed model fit with treatment effect
20 Longitudinal_Fit= lme(Response ~ ObservedTime+age +sex, random = ~ 1|ID,
       data = simlong)
21 summary(Longitudinal_Fit)
22
```

```
# cox model fit with treatment effect
Survival_Fit= coxph(Surv(time, event) ~ sex+age, data = sim_Surv, x =
    TRUE)
summary(Survival_Fit)

# joint model fit with treatment effect
JM_Fit <- jointModel(Longitudinal_Fit, Survival_Fit, timeVar = "
    ObservedTime", control = list(optimizer = "BFGS"))
summary(JM_Fit)

# Plot the baseline survival function
ggsurvplot(survfit(Survival_Fit), data= sim_Surv$sex, palette = "#2E9FDF"
    ,
           ggtheme = theme_minimal())

##Creation of random missingness of varying proportions of 10%, 20%, 30%
    and 40% in the Response variable for longitudinal data
# Set seed for reproducibility
set.seed(123)

# Function to introduce random missingness in the response variable
generate_missingness <- function(data, missingness_rate) {
  # Generate random indices for missingness
  n <- nrow(data)
  missing_indices <- sample(1:n, size = floor(missingness_rate * n))

  # Introduce missingness in the response variable
  data_with_missingness <- data
  data_with_missingness$Response[missing_indices] <- NA

  return(data_with_missingness)
}
```

```
51
52 # Generate datasets with different missingness rates
53 missingness_rates <- c(0.1, 0.2, 0.3, 0.4)
54 datasets_with_missingness <- lapply(missingness_rates, function(rate) {
55   generate_missingness(simlong, rate)
56 })
57
58 # Access datasets with different missingness rates
59 dataset_10_percent_missing <- datasets_with_missingness[[1]]
60 dataset_20_percent_missing <- datasets_with_missingness[[2]]
61 dataset_30_percent_missing <- datasets_with_missingness[[3]]
62 dataset_40_percent_missing <- datasets_with_missingness[[4]]
63
64 # Combination of datasets with an indicator of missingness rates
65 dataset_10_percent_missing$Missingness_Rate <- "10%"
66 dataset_20_percent_missing$Missingness_Rate <- "20%"
67 dataset_30_percent_missing$Missingness_Rate <- "30%"
68 dataset_40_percent_missing$Missingness_Rate <- "40%"
69
70 # Combination of all datasets into one
71 all_datasets <- rbind(dataset_10_percent_missing,
72                       dataset_20_percent_missing,
73                       dataset_30_percent_missing,
74                       dataset_40_percent_missing)
75
76 # Visualization of missingness patterns
77 aggr_plot <- aggr(all_datasets, col = c("blue", "red"),
78                   numbers = TRUE, sortVars = TRUE,
79                   labels = names(all_datasets),
80                   cex.axis = 0.5, gap = 3,
81                   ylab = c("Missing␣data", "Pattern"))
82
```

```
# Print the missingness patterns plot
print(aggr_plot)

#Using Mice to fill in the missingness
# Impute missing values for each dataset
imputed_datasets <- lapply(datasets_with_missingness, function(dataset) {
  # Create mids object
  mids <- mice(dataset, method = "pmm", m = 5) # Use predictive mean
      matching

  # Generate completed dataset
  complete_data <- complete(mids)

  return(complete_data)
})

# Access imputed datasets
imputed_dataset_10_percent <- imputed_datasets[[1]]
imputed_dataset_20_percent <- imputed_datasets[[2]]
imputed_dataset_30_percent <- imputed_datasets[[3]]
imputed_dataset_40_percent <- imputed_datasets[[4]]

#####Fit Joint Models
# linear mixed model imputed_dataset_10_percent
Longitudinal_Fit10= lme(Response ~ ObservedTime+age +sex, random = ~ 1|ID
    , data = imputed_dataset_10_percent)
summary(Longitudinal_Fit10)

# cox model fit with treatment effect
Survival_Fit= coxph(Surv(time, event) ~ sex+age, data = sim_Surv, x =
    TRUE)
summary(Survival_Fit)
```

```
112
113 # joint model fit with treatment effect
114 JM_Fit10 <- jointModel(Longitudinal_Fit10, Survival_Fit, timeVar = "
    ObservedTime", control = list(optimizer = "optim", optCtrl = list(
    method = "BFGS")))
115 summary(JM_Fit10)
116
117 # linear mixed model imputed_dataset_20_percent
118 Longitudinal_Fit20= lme(Response ~ ObservedTime+age +sex, random = ~ 1|ID
    , data = imputed_dataset_20_percent)
119 summary(Longitudinal_Fit20)
120
121 # cox model fit with treatment effect
122 Survival_Fit= coxph(Surv(time, event) ~ sex+age, data = sim_Surv, x =
    TRUE)
123 summary(Survival_Fit)
124
125 # joint model fit with treatment effect
126 JM_Fit20 <- jointModel(Longitudinal_Fit20, Survival_Fit, timeVar = "
    ObservedTime", control = list(optimizer = "optim", optCtrl = list(
    method = "BFGS")))
127 summary(JM_Fit20)
128
129
130 #### linear mixed model imputed_dataset_30_percent
131 Longitudinal_Fit30= lme(Response ~ ObservedTime+age +sex, random = ~ 1|ID
    , data = imputed_dataset_30_percent)
132 summary(Longitudinal_Fit30)
133
134 # cox model fit with treatment effect
135 Survival_Fit= coxph(Surv(time, event) ~ sex+age, data = sim_Surv, x =
    TRUE)
```

```
summary(Survival_Fit)

# joint model fit with treatment effect
JM_Fit30 <- jointModel(Longitudinal_Fit30, Survival_Fit, timeVar = "
    ObservedTime", control = list(optimizer = "BFGS"))
summary(JM_Fit3)


#### linear mixed model imputed_dataset_40_percent
Longitudinal_Fit40= lme(Response ~ ObservedTime+age +sex, random = ~ 1|ID
    , data = imputed_dataset_40_percent)
summary(Longitudinal_Fit40)

# cox model fit with treatment effect
Survival_Fit= coxph(Surv(time, event) ~ sex+age, data = sim_Surv, x =
    TRUE)
summary(Survival_Fit)

# joint model fit with treatment effect
JM_Fit40 <- jointModel(Longitudinal_Fit40, Survival_Fit, timeVar = "
    ObservedTime", control = list(optimizer = "optim", optCtrl = list(
    method = "BFGS")))
summary(JM_Fit40)


#########Compare 5 JM using anova
# Compare models using ANOVA
anova(JM_Fit, JM_Fit10)
anova(JM_Fit, JM_Fit20)
anova(JM_Fit, JM_Fit30)
anova(JM_Fit, JM_Fit40)

```

```
163 ##Application to real HIV data
164
165 #Loading the datasets
166 RealData_Surv <- read.csv("C:/Users/Tessie␣Akoko/OneDrive␣-␣CGIAR/Joint␣
        Modelling/RealData_Surv.csv")
167 RealData_Long <- read.csv("C:/Users/Tessie␣Akoko/OneDrive␣-␣CGIAR/Joint␣
        Modelling/RealData_Long.csv")
168
169 # Plot missingness heatmap for one variable
170 # Create a dataframe with just the variable of interest
171 data <- data.frame(ViralLoad = RealData_Long$ViralLoad)
172
173 # Plot missingness heatmap for the variable
174 vis_miss(data)
175 # Create a matrix of missing data patterns
176 missing_patterns <- md.pattern(RealData_Long)
177
178 # Plot missingness heatmap for one variable
179 md.pattern(RealData_Long$ViralLoad)
180
181 # Perform multiple imputation
182 imputed_data <- mice(RealData_Long, method = "pmm", m = 5)
183
184 # Get the complete imputed datasets
185 completed_data <- complete(imputed_data)
186
187 # Access the imputed values for a specific variable (e.g., ViralLoad)
188 imputed_values_viral_load <- imputed_data$ViralLoad
189 View(completed_data)
190
191 #Visualize data
192 # Example using ggplot2
```

```
193 ggplot(completed_data, aes(x = Episode, y= PatientID, fill = ViralLoad))
        +
194   geom_tile() +
195   scale_fill_gradient(low = "blue", high = "red") +
196   labs(x = "Time", y = "ID")
197
198
199 # Use of 'completed_data' for analysis, which contains the imputed values
200 # linear mixed model fit with treatment effect
201 completed_data <- completed_data %>% mutate(RegimenLine_2 = as.factor(
        RegimenLine),
202                                             Region_2 = as.factor(ifelse(
                                                    Region=="1", "Nairobi", "
                                                    Kiambu")))
203
204 RealData.Long= lme(ViralLoad ~ Episode+RegimenLine_2+TB_HIV_Coinfection+
        Region_2,
205                    random = ~ 1 | PatientID, data = completed_data)
206 summary(RealData.Long)
207
208 # cox model fit with treatment effect
209 RealData_Surv <- RealData_Surv %>% mutate(RegimenLine_2 = as.factor(
        RegimenLine),
210                                           Region_2 = as.factor(ifelse(
                                                Region=="1", "Nairobi", "
                                                Kiambu")))
211
212 RealData.Surv= coxph(Surv(time, Event) ~ RegimenLine_2+TB_HIV_Coinfection
        , Region,
213                      data = RealData_Surv, x = TRUE)
214
215 summary(RealData.Surv)
```

```
216
217 # joint model fit with treatment effect
218 RealData_Joint <- jointModel(RealData.Long, RealData.Surv, timeVar = "
      Episode", cluster = "PatientID")
219 summary(RealData_Joint)
220
221 #Extraction of AIC
222 extractAIC(RealData.Surv)
223
224
225 ##Diiagnostics
226
227 ####
228 # the standard call to the plot() method for objects of class 'jointModel
      '
229 par(mfrow = c(2, 2))
230 plot(RealData_Joint)
231
232 # a useful function used in the residual plots below
233 plotResid <- function (x, y, ...) {
234   plot(x, y, ...)
235   lines(lowess(x, y), col = "red", lwd = 2)
236   abline(h = 0, lty = 3, col = "grey", lwd = 2)
237 }
238
239 par(mfrow = c(2, 2))
240
241 # Subject-Specific Residuals vs Fitted Values
242 resSubY <- residuals(RealData_Joint, process = "Longitudinal", type = "
      stand-Subject")
243 fitSubY <- fitted(RealData_Joint, process = "Longitudinal", type = "
      Subject")
```

```
244 plotResid(fitSubY, resSubY, xlab = "Fitted␣Values", ylab = "Residuals",
245          main = "Subject-Specific␣Residuals␣vs␣Fitted␣Values")
246
247
248 # Marginal Residuals vs Fitted Values
249 resMargY <- residuals(RealData_Joint, process = "Longitudinal", type = "
       stand-Marginal")
250 fitMargY <- fitted(RealData_Joint, process = "Longitudinal", type = "
       Marginal")
251 plotResid(fitMargY, resMargY, xlab = "Fitted␣Values", ylab = "Residuals",
252          main = "Marginal␣Residuals␣vs␣Fitted␣Values")
253
254
255 # Martingale Residuals vs Fitted Values
256 resMartT <- residuals(RealData_Joint, process = "Event", type = "
       Martingale")
257 fitSubY <- fitted(RealData_Joint, process = "Longitudinal", type = "
       EventTime")
258 plotResid(fitSubY, resMartT, xlab = "Fitted␣Values", ylab = "Residuals",
259          main = "Martingale␣Residuals␣vs␣Fitted␣Values")
260
261
262 #shapiro wilks test
263 shapiro.test(RealData_Long$ViralLoad)
```