5-2017

Joint Modelling of Longitudinal Measurements and Time-To-Event Data : Application to HIV Study

Mirna Walid Halawani
Montclair State University

Abstract

Longitudinal and survival data are frequently collected in biomedical studies. The research questions of interest in these studies often require separate analysis of the outcomes. But in many occasions interest also lies in studying their association structures, such as in biomarker research, where the clinical studies are designed to identify biomarkers with strong prognostic capabilities for event time outcomes. In the separate analyses, a linear mixed-effects model is used for modeling the longitudinal data to study the changing trend of the response overtime when controlling some covariates and a survival model is used to model the time-to-event data. A common issue in longitudinal studies is that informative dropout in the data can cause bias in the analysis. Associations between longitudinal and survival data can occur in the explanatory variables or through stochastic dependence between the subject-specific random effect component of the longitudinal model and the survival model. Ignoring the association between the longitudinal and survival data can result in biased inference. The joint model can account for these issues and simultaneously analyze the longitudinal and time-to-event data. This approach enables researchers to obtain more accurate inference regarding the survival probability to certain event when the longitudinal responses associated with the survival response or outcome-dependent study dropout.

In an HIV/AIDS study, our primary interest is to compare the survival for the patients with two antiretroviral drugs, Didanosine (ddI) and Zalcitabine (ddC) with some other risk factors. We also want to determine how the biomarker-CD4 lymphocyte cell counts changed over the period of the study. We use separate analysis and the joint model to analyze the survival and longitudinal outcome and then compare the two analysis results. In the longitudinal analysis, we used a linear mixed-effects model to fit the CD4 cell counts using a random intercept and slope for the observation time. In the survival analysis, we compared the survival between the two treatment groups by using a cox-proportional hazard model. Then a joint model was fitted by using the fitted longitudinal and survival objects. To compare the separate analysis and the joint analysis, we use the Akaike's Information Criteria (AIC). The joint model was shown to be better than the separate analyses of the longitudinal models and survival models with a smaller AIC value. Using the joint model for inference on the HIV study, Zalcitabine (ddC) was significantly effective in reducing a person's risk of death. The risk of death was 1.44 times as likely for patients assigned to ddI as compared to the patients assigned to ddC. The previous diagnosis result and observation time were significant predictors of the change in CD4 cell count at a 0.05 significance level. A patient having a previous diagnosis of AIDS at the study entry led to a decrease in CD4 cell counts thus, a patient was more likely to die or the disease progressed. The joint model showed a significant association between the CD4 count and survival: with higher CD4 count, the survival probability is also significantly higher (or the hazard of death is lower). The joint model approach provided more accurate inference than the separate approaches for the HIV study.

Keywords: longitudinal data, time-to-event data, linear mixed-effects model, survival model, cox-proportional hazard model, joint modelling, HIV/AIDS study

MONTCLAIR STATE UNIVERSITY

Joint Modelling of Longitudinal Measurements and Time-to-Event Data: Application to HIV Study

by

Mirna Walid Halawani

A Master's Thesis Submitted to the Faculty of

Montclair State University

In Partial Fulfillment of the Requirements

For the Degree of Master of Statistics May 2017

College/School The College of Science and Mathematics

Department Mathematical Sciences

Thesis Committee:

[redacted]

Dr. Haiyan Su
Thesis Sponsor

[redacted]

Dr. Andrew McDougall
Committee Member

[redacted]

Dr. Diana Thomas
Committee Member

JOINT MODELLING OF LONGITUDINAL MEASUREMENTS AND TIME-TO-EVENT DATA: APPLICATION TO HIV STUDY

A THESIS

Submitted in partial fulfillment of the requirements

For the degree of Master of Statistics

by

Mirna Walid Halawani

Montclair State University

Montclair, New Jersey

May, 2017

Acknowledgments

I would like to give special thanks to my thesis advisor Dr. Su for her assistance, knowledge and extreme patience. I also want to thank my committee Drs. McDougall and Thomas for supporting me and helping steer this thesis along in a short amount of time. Finally, I thank my friends and family for supporting me and being so understanding and encouraged me to pursue a career in bio-statistical research.

Contents

1 Introduction

Efforts to control the HIV/AIDS disease have increased significantly in recent years, but the virus continues to spread at an alarming rate. Clinical trials have been conducted to alter and enhance the standard HIV/AIDS treatment regimen. Researchers who conduct clinical trials often collect longitudinal and survival data. The goal of researchers who conduct longitudinal clinical trials on HIV/AIDS is to understand how the disease progresses over time and to identify risk factors for the disease. For example, in a study involving patients with HIV symptoms, researchers performed repeated measurements of patients' CD4 lymphocyte cell counts to understand the progression of disease with time and how risk factors such as gender, previous opportunistic infection, and Zidovudine (AZT) impact a person's chances of disease progression [1]. In the same study, researchers used survival data, also known as time-to-event data, to study treatment effects of HIV/AIDS medications over a span of time until an event of interest occurred.

Longitudinal and time-to-event data are frequently used in biomedical studies. Researchers may also use longitudinal and time to event data when studying the effect of endogenous time-dependent covariates measured repeatedly over time and when attempting to correct for nonrandom dropout. To analyze longitudinal and time-to-event data in HIV/AIDS studies, researchers use joint and separate models. A survival model combined with a model that enables researchers to study the effects of endogenous time-dependent covariates measured repeatedly and over time is an example of a joint model [7]. Alternatively, researchers conduct separate analyses in studies that include longitudinal and time-to-event data. A popular separate analysis is to analyzing longitudinal using a linear mixed-effects model and analyze time-to-event data using a survival model. A linear mixed-effects model is used to describe

the process of the repeated measurements over time and study for the treatment effect and the survival model is to analyze the treatment effect on survival up until an event of interest occurs [7]. The separate model approach does not enable researchers to establish if there is an association between the components of the two models.

The joint model approach is complex but enables researchers to establish if an association exists. Another advantage of using the joint model approach is that it enables researchers to conduct survival and longitudinal analyses simultaneously. Due to the complexity of the joint modeling approach, it is under-used in clinical research. The objective of this thesis is to provide insight and understanding into the joint modeling process. We will first discuss longitudinal studies and introduce linear mixed-effects models to analyze the longitudinal data. Next, we will introduce the survival model to analyze time-to-event data with nonparametric methods and the cox proportional hazard model. Furthermore, we will construct a joint model for longitudinal with time-to-event data. We will focus on a randomized clinical trial involving patients with HIV/AIDS to provide detailed analytic methods for joint modeling using the statistical software R [17].

In our study of interest, researchers collected longitudinal and survival data and compared the efficacy and safety of two antiretroviral drugs. The two antiretroviral drugs were used to treat patients who were intolerant to zidovudine (AZT) therapy [1]. A total of 467 patients with HIV or low CD4 counts were enrolled into the study. The 467 patients were each randomly assigned to one of two groups. One group of patients received an antiretroviral drug called Didanosine (ddI). The other group of patients received an antiretroviral drug called Zalcitabine (ddC). Researchers compared the effects of the two treatments and studied how the patients' CD4 cell counts changed over the course of the study. Our goal is to fit a joint model and conduct diagnostic analysis to ensure that the researchers of this study made valid inferences regarding

the effects of ddI and ddC on the survival of HIV/AIDS patients.

1.1 Clinical Trial

We will first provide background into biomedical research and the statistical methods used to analyze data. A clinical trial is conducted by researchers who want to study subjects impacted by or exposed to a disease or risk factor. Researchers use clinical trials to compare new intervention methods to existing treatment methods. Clinical trials are used to enhance medical intervention options for a disease or outcome, thus finding better ways to prevent, screen, diagnose, or treat a disease. To carry out a clinical trial, subjects are randomly assigned to different intervention groups or stratified according to different prognostic factors such as age or gender. Many clinical trials involve following up with patients for a long period of time. The follow-up time for the study may range from a few weeks to many years.

Different statistical procedures are used clinical trial studies. These procedures are useful in clinical research and provide vital information about intervention methods. For example, in the case of an oncology clinical trial, an intervention method is investigated to test the response on tumor shrinkage. In HIV/AIDS studies, different treatments are used to examine the change of CD4 cells on the survival time of patients with HIV. Typically, longitudinal data in HIV/AIDS studies consists of recording measurements of CD4 cell counts taken at various points in time throughout the study period. CD4 cell counts are considered important biomarkers of HIV disease progression because CD4 cell counts are an important part of the immune system, which begin to deplete as the virus infects the body.

1.2 Longitudinal Data Analysis

A longitudinal study is a type of randomized controlled experiment in which data is collected on repeated measurements for the same subject at a series of time points. In practice, the measurements are observed at discrete time points, usually including baseline measurements. In medical research, the focus is often on interrelationships between the variables of repeated measurements on a continuous response. A familiar example is that of HIV clinical trials, where covariates, including treatment assignment, demographic information, and measurements on immunologic and virologic status such as CD4 cell counts are recorded at baseline and then taken at subsequent clinic visits. Although it is common to assume independence between subjects, such repeated measurements are correlated within-subjects and therefore require special statistical methods for validity and inference. The researcher can establish sequences of events because longitudinal studies extend beyond a single moment in time.

1.2.1 Linear Mixed-Effects Models

We will use linear mixed-effects models to analyze the longitudinal data. The linear mixed-effects (LME) model is a type of model that uses fixed effects and random effects in the same analyses. A fixed effect contains covariates that relate to the i^{th} patient at time of their j^{th} measurement, where $i = 1, ...m$ and $j = 1, ..., n_i$. The primary interest in the model are the fixed effects, which include levels that could be used multiple times for repeated measurements. A random effect is a patient-specific coefficient that represents between-patients heterogeneity in an outcome variable that cannot be explained by measured covariates [4]. The LME model is widely used in statistical analyses of longitudinal data as it considers both the within-subjects and between-subjects variation. The LME model allows for a wide variety of correlation patterns. The LME model is also effective in modeling data that is missing at random.

Most models exclude data on a subject if one measurement is missing. In cases where data on a subject is missing at random, the LME model includes the available data on that subject instead of excluding data on the subject all together. The LME model is also preferred when there is uneven spacing of repeated measurements. For example, in our application study, measurements of CD4 cell counts for HIV/AIDS patients were recorded at 0, 2, 6, 12, and 18 months.

A common model-based approach for longitudinal measurements assumes independence between subjects i, where each measurement is a realization of a Gaussian random variable [21]. The LME model can be used as an extension of the general regression model in equation (1). In equation (1), Y is a $N \times 1$ response vector and we denote y_{ij} as the jth measurement on the ith subject. X is a $N \times (p+1)$ matrix, where p are the number of explanatory variables. β is a $(p+1) \times 1$ dimensional vector of fixed-effects regression coefficients, and ε is a $N \times 1$ vector of the measurement errors, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. To extend the general regression model (1) to the LME model, we need to consider a random effect. In the equation we use Z to denote an $N \times q$ design matrix for the q-dimension random effects γ; in which Z could be a submatrix of the X matrix. γ is a $q \times 1$ vector of the random effects. We assume $\gamma \sim N(0, D)$ and γ, ε are independent. We denote Σ_i as the variance-covariance matrix of the response y_i with $var(Y_i) = \Sigma_i = Z_i D Z_i^T + \sigma_\varepsilon^2 I_{n_i}$.

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{1}$$

By inclusion of the random effects component, equation (1) is extended to the linear mixed-effects model in equation (2):

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon} \tag{2}$$

Different models for longitudinal data differ on the correlation structure for error term [21]. We will use the LME form given by equation (3) where Y_{ij} is the response variable measured on subject $i = 1, ..., m$ at time point t_{ij}, with $j = 1, ...n_i$. m is the number of subjects and n_i is the number of measurements for subject i. $W_{1i}(t_{ij})$ is the unobserved random process; where we change notation from Z in equation (2) to $W_{1i}(t_{ij})$. ε_{ij} are independent realizations of zero-mean Gaussian random variables with variance σ_ε^2 representing pure measurement error. $\mu_i(t_{ij})$ is the mean response for subjects i at time point t_{ij}, which represents the fixed effects and can have linear form such as $x_{ij}^T\beta$. From this point on we will assume that ε_{ij} represents pure measurement error.

$$Y_{ij} = \mu_i(t_{ij}) + W_{1i}(t_{ij}) + \varepsilon_{ij} \tag{3}$$

In this paper we will distinguish between the models of longitudinal process and of the survival process to construct the joint modelling of the two processes. We want to distinguish between the random effects component $W_i(t_{ij})$ in each process. The random effects component is subscript with a '1', $W_{1i}(t_{ij})$, to denote that it belongs to the longitudinal process of the two-stage joint modelling process. In the survival process, the random effects component of the survival model will be subscripted with a '2'. We will provide more details on the relationship of these components later in the paper.

Furthermore, some authors choose to decompose the unobserved random process $W_i(t_{ij})$ into two components in an additive way. For example, model (4) uses Diggle's, and Laird and Ware's proposed linear mixed-effects model [4] [9]. In this model $\boldsymbol{U}_i$ are m independent realizations of a r-dimension multivariate Gaussian random variable and $\boldsymbol{d}_i$ are r-dimensional vectors of explanatory variables for the random process $\boldsymbol{U}_i$.

The $V_i(t_{ij})$ are m independent realizations of a stationary Gaussian process.

$$W_i(t_{ij}) = \boldsymbol{d}_i'(\boldsymbol{U}_i) + V_i(t_{ij}) \tag{4}$$

Guo and Carlin also use Diggle's and Laird Ware's subject-specific (LME) model in an application to an HIV study. For our study of interest, we use the form of longitudinal model proposed in their paper [7] and use error structure proposed by Laird and Diggle in their paper [9]. We start with notation to the longitudinal model in which every parameter will have a subscript '1' for the combined methods later. The sequence of measurements $y_{i1}, y_{i2}, ..., y_{in_i}$ for the ith subjects at times $t_{i1}, t_{i2}, ..., t_{in_i}$ is modeled in the LME model (3), where $\mu_i(t_{ij}) = \boldsymbol{x}_{1i}^T(t)\beta_1$ is the mean response, $W_{1i}(t) = d_{1i}^T(t)U_i$ incorporates subject-specific random effects and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ is a sequence of mutually independent pure measurement errors account for variability between subjects. In application to the HIV study, $W_{1i}(t)$ is the true individual level CD4 trajectories after adjusting for the overall mean trajectory and the fixed effects. The vectors $\boldsymbol{x}_{1i}(t)$ and β_1 represent time-varying explanatory variables and their coefficients for the longitudinal process, respectively. The U_i are vectors of random effects corresponding to the explanatory variables $\boldsymbol{d}_{1i}(t)$. We will discuss a linear mixed-effects model with random intercept only and then a model with the combination of random intercept and random slope in our application to the HIV study.

1.3 Time-to-Event Data Analysis

In survival analysis, subjects are followed over a specified length of time in order to study the relationship of survival up to a certain event point with some risk factors. Time-to-event data may be based on events other than death, such as recurrence

of a disease, disease progression, or discharge from the hospital. For example, in biomedical studies we study treatment effects of HIV/AIDS medications over a span of time until death or disease progression. Survival data includes a censoring variable which indicates if an event occurred or did not occur during the observation time. Observations are called censored when the information about their survival time is incomplete; the most commonly encountered form is right censoring. Right censoring is indicated when a patient does not experience the event of interest for the duration of the study. The survival time for this person is considered to be at least as long as the duration of the study. Right censoring could also occur when a person drops out of the study before the end of the study observation time and did not experience the event. This person's survival time is said to be censored, since we know that the event of interest did not happen while this person was under observation. Censoring is an important issue in survival analysis, representing a particular type of missing data. Survival analysis requires censoring be random and non-informative to avoid bias. This means that the time-to-event and censored are independent.

In survival analysis, researchers can use life tables, Kaplan-Meier curves to describe the survival times of patients of some intervention groups. We use log-rank tests to compare the survival curves of two or more groups. To describe the effect of some explanatory variables on survival time, we can use cox proportional hazards regression or parametric survival models. We will provide a brief description of the study methods using the data from the HIV study.

1.3.1 Nonparametric Methods

The survival and hazard functions are key concepts in survival analysis for describing the distribution of event times. The main characteristics in survival functions are the response variable y as the surviving time until the occurrence of a well-defined event,

which could be censored, in the sense that for some units the event of interest has not occurred at the time the data are collected. First, we let T be a non-negative random variable representing the waiting time until the occurrence of an event; usually the time of failure. We let $f(t)$ be the density function of T. The survival function $S(t)$ is the probability of time-to-event, denoted by the random variable T, beyond some time point t. For example, we examine the probability of death or disease progression up to 18 months for different treatment groups to test the efficacy of the intervention method on patients with HIV. The survival function is defined on the domain $t \in [0, \infty)$ and has a probability range from [0,1]. We assume at $t = 0$, the probability of survival will be one ($S(0) = 1$) unless there is an immediate death, and $S(t)$ will approach zero as age increases without bound, indicating that life eventually ends.

The hazard function $\lambda(t)$, is defined as the instantaneous rate of death occurring at time t, given that failure time did not occur up to that time or before that time. We denote the survival function and hazard function respectively:

$$S(t) = P(T > t) \tag{5}$$

$$\lambda(t) = \lim_{\Delta \to 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{6}$$

We also denote the hazard function as a relation between the density function and survival function or the first derivative of the survival function and the survival function:

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{S'(t)}{S(t)} \tag{7}$$

Generally in survival studies, we wish to describe the relationship of a factor of interest (e.g. treatment) to the response, in the presence of several covariates, such

as age, gender, race, etc. In HIV studies, we analyze how the treatment affects survival. Methods to analyze the relationship of a set of predictor variables with the survival time include parametric, nonparametric and semiparametric approaches. This paper will consider nonparametric methods such as the Kaplan-Meier method and the semiparametric cox-proportional hazard model because they does not assume a specific distribution.

The logrank test and Kaplan-Meier method are widely used to compare and estimate survival probabilities as a function of time, respectively. The Kaplan-Meier method can be used to obtain univariate descriptive statistics for survival data, including the median survival time, and compare the survival probability for two or more groups of subjects. Graphically, Kaplan-Meier curves are useful in obtaining the probability of survival at different time points where life tables show us the number of patients at risk and survival probability for each observed time point in a table.

A life table is useful in survival analysis because it summarizes survival data in terms of the number of events and the proportion surviving at each event time point. For example, researchers could construct life tables to compare the amount of patients alive at various time points in each treatment group. The life table will provide information on the intervention method in cases of comparing the efficacy of two treatments or cases where the treatment is not working and the study needs to stop. For notation in our application we will denote a few variables. Let $t_{(1)} < t_{(2)} < ... < t_{(k)}$ be the ordered subset of $k \leq n$ unique observed failure times from the observed survival times. Let d_i be the number of failures which occurs at t_i and n_i be the number of patients at risk before time t_i. We also denote the risk set $R(t_i)$ and the relation d_i/n_i as the probability of failure at time t_i [21]. Based on the survival function previously mentioned we estimate the product-limiting estimate of

the survival function and hazard function respectively:

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_{(1)} \\ \prod_{t_{(i)} \leq t}(1 - \frac{d_i}{n_i}) & \text{if } t \geq t_{(1)} \end{cases} \tag{8}$$

$$\hat{\lambda}(t) = \frac{d_i}{n_i} \tag{9}$$

The variance of the product-limit estimate of the survival function can be obtained by

$$V(\hat{S}(t) = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{n_i - d_i} \tag{10}$$

Following the Kaplan-Meier method, researchers would then like to analyze the effect of predictor variables on the survival time. Kaplan-Meier curves do not work easily with quantitative predictors such as gene expression, CD4 count, or age. Cox-proportional hazard regression is usually used when the data contains quantitative predictor variables. It allows testing for differences in survival times of two or more groups of interest, while allowing to adjust for covariates of interest. The Cox proportional hazard model extends the logrank test by allowing the inclusion of additional covariates. The cox regression model use the hazard format instead of the survival probability, which makes it easy to interpret information regarding the relationship of the hazard function to predictors.

1.3.2 Cox-Proportional Hazard Model

The Cox-proportional hazard model is used when the study interest is to get inference for the model parameters of a time-to-event process. In the Cox-proportional hazard model, the hazard of an individual with some covariates is proportional to a baseline function of time, where the baseline hazard function has no specified form. This

model allows for fixed covariates that do not change over time and parameters are estimated by maximizing the partial likelihood. The covariates can also include time-dependent variables. Their corresponding hazard function are given in equations (11) and (12) respectively, where $\lambda_0(t)$ is the unspecified baseline hazard function.

$$\lambda_i(t|X) = \lambda_0(t)exp\{X_i\beta\} \tag{11}$$

$$\lambda_i(t|X) = \lambda_0(t)exp\{X_i(t)\beta\} \tag{12}$$

We use the notations from Guo and Carlin's [7] in the cox proportional hazard model in equation (13), where the vectors $x_i^T(t)$ and β_2 are possibly time-dependent explanatory variables and their corresponding regression coefficients. $W_{2i}(t)$ includes subject-specific covariate effects and an intercept. Here, the random effects component $W_{2i}(t)$ in the survival model is subscripted with a '2' to denote that it belongs to the survival process. This is to distinguish between the random effects components in the construction of the two-stage joint modelling process in the next section.

$$\lambda_i(t) = \lambda_0(t)exp\{x_i^T(t)\beta_2 + W_{2i}(t)\} \tag{13}$$

2 Joint Modelling

2.1 Literature Review

It has become increasingly common in survival studies to record the values of key longitudinal covariates until the occurrence of survival time of a subject. This leads to informative dropout of the longitudinal data, which also complicates the survival analysis. Furthermore, in a survival analysis setting where the covariate of interest is time-dependent, either the entire history of the covariate measurement of every

subject, or, measurements of the covariate at each time of death occurrence or disease progression for all subjects in the risk set is needed. By modelling the covariates over time, we enhance the survival analysis since we can interpolate covariate values between the observed measurements to the specific times of death or disease progression, with use of the entire history of the subjects. Modelling the covariate also allows adjustment for covariate measurement error, which is known to result in biased estimates of relative risk parameters [6]. In addition, we can obtain improved covariate tracking estimates by adjusting for informative right censoring of the repeated measurements by the disease progression. Therefore to account for the association in the separate models researchers use joint modelling to handle irregularity and measure time-varying covariates correctly [2]. We defined a joint model in this paper as a survival model combined with a longitudinal model that enables researchers to study the effects of endogenous time dependent covariates measured repeatedly and over time.

Typically, a linear mixed-effects model is used first to describe the process of the repeated measurements over time and study for the treatment effect. A common problem in longitudinal studies is that informative dropout in the data could cause bias in the analysis. To account for informative dropout, a number of model-based approaches have been proposed to jointly model longitudinal outcome and the dropout mechanism (Wu and Carroll, DeGruttola and Tu, Little, Hogan and Laird [11]). We will use a linear mixed-effects model to analyze the repeated measurement of time-dependent variables over time due to its popularity and simplicity.

After the linear mixed-effects, researchers use survival models to analyze the treatment effect up until an event of interest occurs [7]. A widely used survival model is the proportional hazard model. Various approaches have been proposed under this framework including the regression method from Pawitan and Self, Tsiatis, DeGrut-

tola and Wulfsohn, the likelihood-based approaches from DeGruttola and Tu, Faucett and Thomas, Wulfsohn and Tsiatis [23], Henderson, Diggle and Dobson, and Song, Davidian and Tsiatis), corrected score (Wang) and conditional score (Tsiatis and Davidian, Song, Davidian and Tsiatis) approaches [20]. In HIV/AIDS studies it is known that the effect of antiretroviral treatments may decay after some time, therefore the traditional hazard assumption may be too restrictive in this case. An appealing alternative is the time-varying coefficient proportional hazards model proposed by Song and Wang [20], which allows the effect of the coefficients to vary over time.

Although there are many different approaches to construct models for the longitudinal and the time to event data, the separate model approach does not enable researchers to establish if there is an association between the components of the two models. Associations between longitudinal and survival data can occur in the explanatory variables or through stochastic dependence between the subject-specific random effects component of the longitudinal model and the survival model. We also assume associations between the drop-out process; when a missing longitudinal measurement terminates the sequence of longitudinal measurements, and the censoring process. When association between the two processes exists, we use a joint model to obtain less biased and more efficient inferences. The joint model approach is complex but enables researchers to establish if an association exists.

Model-based approaches for each type of analysis have been extensively described in the literature in HIV/AIDS studies. Clayton proposed a comprehensive model that combined the covariate tracking and disease risk models to estimate parameters of similar models. DeGruttola and Tu [3] and Tsiatis et al. [22] consider the progression of CD4 lymphocyte counts and survival time in patients of AIDS. DeGruttola and Tu assume that the joint distribution of log CD4 counts and some transformation of survival time are multivariate normally distributed. This formulation allows them to

use the modified EM algorithm from Laire and Ware [14] to fit the model. Using a Cox-proportional hazard model, Tsiatis et al. model the hazard of death as a function of the conditional expectation of the 'true' log CD4 counts given the history of the observed counts. They proposed a two-step procedure for fitting their model. First they assume a growth curve random components model with normal error for the true CD4 count and used a modified EM algorithm. Then they substituted these estimates into the proportional hazard model and used the cox regression to obtain estimates of the survival parameters.

Self and Pawitan proposed a similar two-step procedure for parameter estimation where they condition on the survival information when computing expected values of the covariates. They, like many others, used partial likelihood methods to obtain estimates of the disease risk parameters and maximum likelihood methods to model jointly immunologic markers, time to infection, and time to AIDS [12]. Wulfsohn and Tsiatis suggest that the joint maximum likelihood method is among the most satisfactory approaches to combine information. The approach described by Wulfsohn and Tsiatis is semiparametric in that no parametric assumptions are imposed on the baseline hazard function in the Cox model, while the random effects in the longitudinal component are assumed to be normally distributed. An attractive feature of this approach is its robustness against departure from the normal random effects assumption. It is said to be as efficient as a semiparametric random effects model proposed by Song, Davidian, and Tsiatis. Many also consider fully parametric Weibull regression models for the times to disease and infection. On the other hand, Faucett and Thomas used simulation studies to compare the analysis of the joint covariate tracking and disease risk model using Gibbs sampling to separate the analysis of each component.

To illustrate the association between the longitudinal component and survival

component, Henderson et al. proposed a LME model for the i^{th} subject via an unobserved or latent zero Gaussian process $W(t) = \{W_{1i}(t), W_{2i}(t)\}$, which is realized independently in different subjects. They assumed that the latent process forces a pair of linked longitudinal and survival sub-models. The longitudinal component has the following format:

$$Y_{ij} = \mu_i(t_{ij}) + W_{1i}(t_{ij}) + \varepsilon_{ij} \tag{14}$$

Where $\mu_i(t_{ij})$ is the mean response which could be described by a linear model $X_i(t)\beta_i$ $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ is a sequence of mutually independent measurement errors. They assumed $\mu_i(t)$ could be describe by a linear model where $u_i(t) = x_{1i}(t)^T\beta_i$. The survival component has the form as in equation (15), where $\lambda_i(t)$ is the hazard function and $\lambda_0(t)$ is the baseline hazard function. $x_{2i}(t)$ and β_2 represent time-varying explanatory variables and their coefficients for the survival process, respectively. $W_{2i}(t)$ is an unobserved random process for the survival process.

$$\lambda_i(t) = \lambda_0(t)\exp(x_{2i}(t)^T\beta_2 + W_{2i}(t)) \tag{15}$$

2.2 Formulation of the Joint Model

We will use the longitudinal model (14) and survival model (15) in the formulation of the joint model. The association between the longitudinal and survival components can arise in two ways. One is through common explanatory variables and the other is through stochastic dependence between the longitudinal and survival process. The association is thus between W_{1i} and W_{2i} in equation (14) and (15) as proposed by Henderson et al. They propose to jointly model the two processes via a latent zero-mean bivariate Gaussian process on $(W_{1i}, W_{2i})^T$, which is independent for different subjects. $W_{1i}(t)$ and $W_{2i}(t)$ link the longitudinal model in the LME from equation (14)

with the cox-proportional hazard model (15) together. The joint model is composed by the two linked sub-models, which they refer to as the longitudinal measurement model and the intensity survival model. When association between the two processes exists, the joint model provides less biased and more efficient inferences than the separate analysis.

3 Application to HIV Study

3.1 Clinical background

The availability of an increasing number of antiretroviral agents and the rapid evolution of new information has led to new treatment regimens for patients infected with HIV. A current treatment regimen used to treat patients with HIV/AIDS is Zidovudine (AZT) therapy. Anti-HIV drugs such as AZT slows down or prevents damage to the immune system. These drugs also reduces the risk of developing AIDS-related illnesses. It is known however, that patients can be intolerant to AZT or experience a 'failure'. In such cases, researchers provide alternative intervention methods or treatment such as the antiretroviral drugs: Didanosine (ddI) and Zalcitabine (ddC). Both drugs are commonly used to treat patients with HIV who cannot tolerate AZT or who had disease progression despite it [1]. It is also common in HIV studies to collect repeated measurements of important biomarkers of HIV progression such as CD4 lymphocyte cell count in a sample of blood and viral loads [15]. CD4 cell are an important part of the immune system, which begin to deplete as the virus infects the body. We note that a decrease in CD4 cell counts indicates the degree of immunosuppression. Our objective is to investigate the change of CD4 cell count over the period of the study to determine if the two drugs impact the survival of patients using the HIV dataset.

3.2 Description of Dataset

The HIV study was a randomized clinical trial in which both longitudinal and survival data were collected to compare the efficacy and safety of two antiretroviral drugs in treating patients who had failed or were intolerant of Zidovudine (AZT) therapy [1]. The study enrolled 467 patients from December 1990 through September 1991. The patients were enrolled if they met the following criterion: they had an AIDS-defining condition or they had two CD4 counts of 300 cells or less per cubic millimeter, with either a positive serologic test for HIV or a clinician's working diagnosis of HIV infection; and they had undergone AZT therapy that led to intolerance of the drug or progression of disease during therapy[1]. The patients were then randomized to receive two antiretroviral drugs, either Didanosine (ddI) or Zalcitabine (ddC). However, a patient was allowed to switch drug treatment after 3 months with a washout period of at least 3 days. The patient data was then censored at the time of drug re-assignment.

The dataset consisted of 1408 observations on 9 variables. The data consisted of three continuous explanatory variables: the square root of the CD4 lymphocyte cell count, but for simplicity reasons we will refer to as CD4, Time (the time to death or censoring), and Obstime (the observed time points for CD4 measurements). CD4 cell counts were recorded at baseline (0 months) and at various time points (2, 6, 12 and 18 months) during the trial. The data also included six binary explanatory variables: Drug (ddC, ddI), Death (censoring=0, death=1), Gender (male, female), PrevOI (previous opportunistic infection of AIDS diagnosis or no previous AIDS diagnosis at study entry), and AZT (failure or intolerance). In total, 230 patients received ddI and 237 received ddC. There were 45 females and 422 males involved in the study. A total of 160 patients did not have a previous diagnosis of AIDS while 307 patients did have a diagnosis of AIDS at beginning of the study. There were 292 patients who were intolerant to AZT and 175 experienced a 'failure' to AZT, meaning the disease

progressed despite AZT.

Let y_{ij} denote the square root of the jth CD4 count measurement on the ith patient in the trial, $j = 1, \ldots, n_i, i = 1, \ldots, m$. Four explanatory variables as main effects were included in the analysis: Drug(ddI, ddC), Gender(male, female), PrevOI(AIDS,noAIDS), and AZT(intolerance, failure). The main goal of the study is to analyze the association of among CD4 count and survival, druggroup, gender, AIDS-diagnosis at baseline and AZT intolerance, accounting for all relevant correlations and subject-specific random effects.

3.2.1 Exploratory Data Analysis

To visualize the dataset, we used exploratory plots of the CD4 cell counts at each observed time point in Figure 1 and the CD4 counts at each observed time separated by treatment in Figure 2. The figures allow us to study how the CD4 cell count changes over time and determine the shape of the distributions. Figure 1 shows that the median of the CD4 counts was greater for the beginning months of the trial. We observed that the CD4 cell counts decreased at each observed month for all patients. The median CD4 count at each time point was between 5 and 10. The summary statistics value the median of the CD4 cell count at 6.083. However, we see outliers at months 2, 6, and 18 months. The boxplot for month 12 presents the most variability represented by the IQR. The shape of each distribution is right skewed and contains outliers. The dataset used a square root transformation of CD4 cell count however still see some skewness. The boxplots show high outliers after the baseline measurements, meaning there were patients with exceptionally high CD4 cell counts. Furthermore, we wanted to investigate how the CD4 cell count changes for each treatment group at each time point to detect a pattern. In Figure 2, the patients who received ddC had a lower median CD4 cell count than the patients who

received ddI during months 0-12. However, at month 18, the median CD4 cell count was higher in the ddC group compared to the ddI. It appears that at the end of the study patients assigned to ddC had a better chance of surviving or the disease has not progressed. However, there does not appear to be a big difference between the two treatment groups.

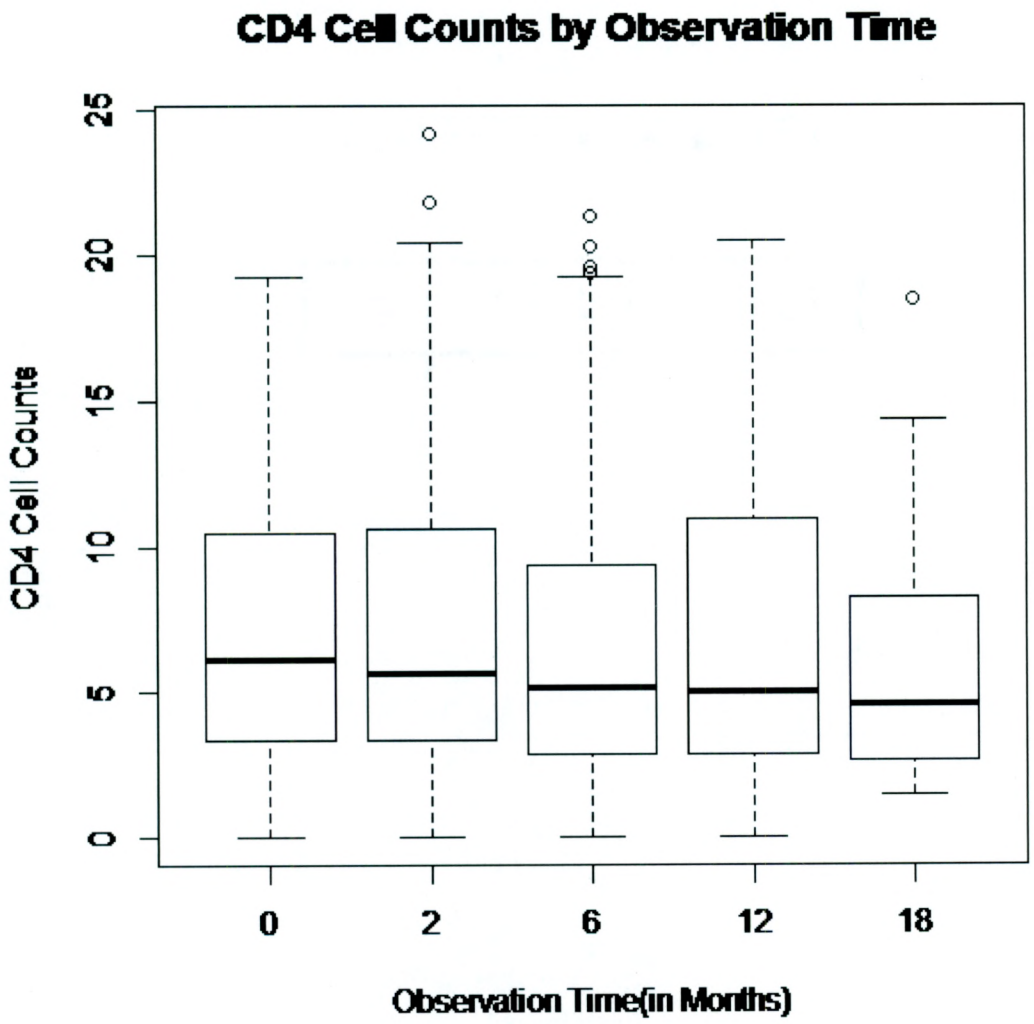

Figure 1: Boxplot illustrating the CD4 cell counts recorded at time points: 0, 2, 6, 12, and 18 months for all subjects

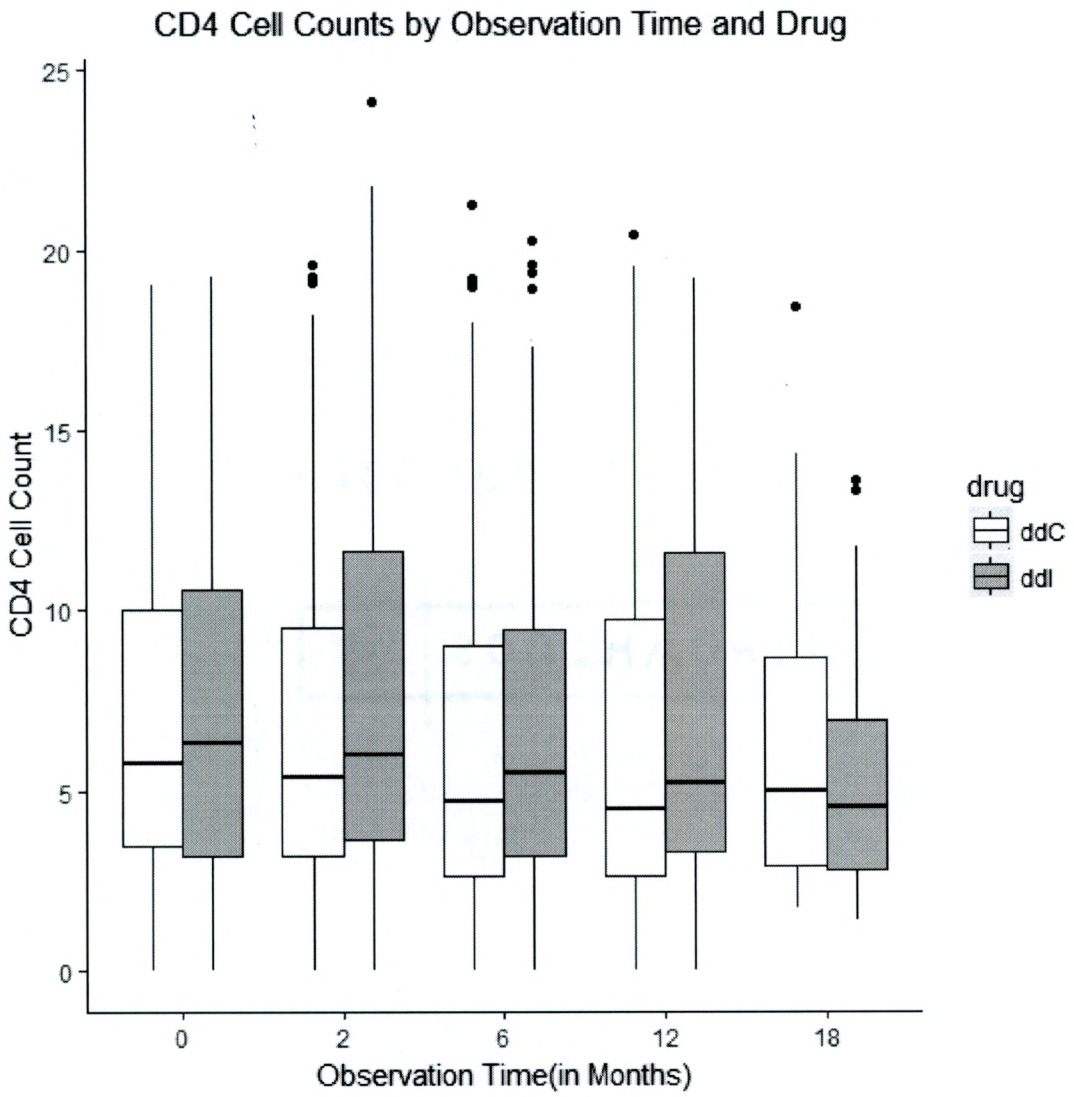

Figure 2: Side-by-side boxplots illustrating the CD4 cell counts recorded at time points: 0, 2, 6, 12, and 18 months for both treatment groups

3.3 Longitudinal Data Analysis

3.3.1 Linear Mixed-Effects Model

We first examine the change of CD4 cell count throughout the time of the study and compare the two drugs when controlling other covariates. To analyze the repeated

CD4 cell measures we use linear mixed-effects models.

The linear mixed-effects models can be modeled as:

$$y_{ij} = \beta_{11} + \beta_{12}(t_{ij}) + W_{1i} + \varepsilon_{ij}, \tag{16}$$

where y_{ij} is the response of the j^{th} CD4 cell count measured on i^{th} patient in jth measurement, $j = 1, ..., n_i$ and $i = 1, ...n$ where $n = 467$ subjects followed over the period of the study [0-18 months]. β_{11} is the intercept and β_{12} is the observation time parameter and W_{1i} is the intercept random effect. The fitted model shows a significant negative effect of observation time on CD4 count. But this model only assumes subjects have different CD4 count at the baseline, their decreasing rate are all the same with time. We know that the AIC value for the model will be higher than the models we fit next. Therefore, we do not include the output table in the paper.

Next, in equation (17), we fit a linear mixed-effects model with random intercept and slope for Obstime. This model also assumes the change rate of CD4 with time are different with different subject besides assuming different baseline CD4 counts. From equation (16) we include the random effects for intercept and slope over time, $W_{1i}(t_{ij})$. We denote $W_{1i}(t_{ij}) = U_{1i} + U_{2i}(t_{ij})$. We include CD4 count and use the intercept and obstime as the random effects. The results are shown in Table 1; Obstime is a significant predictor of CD4 cell count ($t = -9.88$, p-value $= 0$). We denote p-value $= 0$ as $p < 0.0001$. The parameter for obstime is -0.1501, this describes the negative effect of obstime time on CD4 cell count. For every one month increase in observation time, the CD4 cell count decreases by 0.1501 cells per cubic millimeter of blood. A decrease of CD4 cell counts is harmful to a patient's autoimmune response in ability to fight diseases. This is evidence of death or disease progression occurring

throughout the study period.

$$y_{ij} = \beta_{11} + \beta_{12}(t_{ij}) + W_{1i}(t_{ij}) + \varepsilon_{ij} \tag{17}$$

Table 1: Linear Mixed-Effects Model with Random Intercept and Slope

LME Model	AIC	BIC	logLik
Random Int. and Slp.	7141.282	7172.76	-3564.641

Random Effect	StdDev	Corr
Intercept	4.5065	(Intr)
obstime	0.1729	-0.152
Residuals	1.7508	

	Estimate	Std.Error	DF	T-Value	P-value
Intercept	7.1890	0.2222	937	32.36	0
obstime	-0.1501	0.0152	937	-9.88	0

In addition, we examined the treatment effect on CD4 cell count by controlling other covariates in the model. This model (model 3) is a linear mixed-effects model with a random intercept and random slope for obstime and the inclusion of all the four covariates drug, gender, prevOI, and AZT. The output displayed in Table 3 shows that prevOI and obstme are significant predictors of the change in CD4 cell counts. In comparison to Models 1 and 2, the AIC value is smaller in the model with the added covariates ($AIC = 7020.004$) than the lme model with only a random intercept ($AIC = 7176.633$) and the lme model with a random intercept and slope ($AIC = 7141.282$). The estimate for prevOIAIDS is negative; this indicates that the CD4 cell count decreases more for the patients with a previous diagnosis of AIDS than the patients with no previous diagnosis of AIDS at the study entry. In agreement to our previous models, the variable obsime is significant ($p - value = 0$) and has a negative estimate. We conclude, as the observation time increases by month, the CD4 cell count decreases by 0.1524. The estimate for the obstime is similar to the

previous model outputs. In addition, the treatment term was found to be insignificant and reported an estimate of 0.4544. The CD4 cell count was higher in the ddI group than in the ddC group by 0.4544 cells. Males had lower CD4 cell counts than women by 0.3154. Patients who experience a disease progression despite AZT therapy had lower CD4 cell counts than patients who were intolerant to AZT by 0.2570.

Table 2: Linear Mixed-Effects Model with Random Intercept and Slope with Covariates

Full LME Model	AIC	BIC	logLik		
with covariates	7020.004	7072.439	-3500.002		
Random Effect	StdDev	Corr			
Intercept	4.0029	(Intr)			
obstime	0.1726	-0.18			
Residuals	1.7496				
	Estimate	Std.Error	DF	T-Value	P-value
Intercept	10.3869	0.6886	937	15.08	0 .0000
obstime	-0.1524	0.1514	937	-10.0686	0.0000
drugddI	0.4544	0.3803	462	1.19	0.2328
gendermale	-0.3154	0.6527	462	-0.48	0.6291
prevOIAIDS	-4.62561	0.4787	462	-9.66	0.0000
AZTfailure	-0.2570	0.4725	462	-0.54	0.5868

Next, we extend model 3 by including an interaction effect between drug and obstime. Model 4 is shown in equation (18). The output in Table 3 report the $AIC = 7026.648$, this value differs by 6.644. Although the interaction effect is not significant, we keep it in the model in case the investigator is interested in it. We then would consider Model 4 with in the interaction term in our construction of the joint model. The interaction between drug and obstime did not have a significant effect on the CD4 cell count. PrevOIAIDS and obstime were once again significant predictors on CD4 cell counts. Each estimate is similar to the output from model 3 and our conclusion remain the same. We conclude, as the observation time increases by month, the CD4 cell count decreases by 0.1628. The CD4 cell count was higher

previous model outputs. In addition, the treatment term was found to be insignificant and reported an estimate of 0.4544. The CD4 cell count was higher in the ddI group than in the ddC group by 0.4544 cells. Males had lower CD4 cell counts than women by 0.3154. Patients who experience a disease progression despite AZT therapy had lower CD4 cell counts than patients who were intolerant to AZT by 0.2570.

Table 2: Linear Mixed-Effects Model with Random Intercept and Slope with Covariates

Full LME Model	AIC	BIC	logLik		
with covariates	7020.004	7072.439	-3500.002		
Random Effect	StdDev	Corr			
Intercept	4.0029	(Intr)			
obstime	0.1726	-0.18			
Residuals	1.7496				
	Estimate	Std.Error	DF	T-Value	P-value
Intercept	10.3869	0.6886	937	15.08	0 .0000
obstime	-0.1524	0.1514	937	-10.0686	0.0000
drugddI	0.4544	0.3803	462	1.19	0.2328
gendermale	-0.3154	0.6527	462	-0.48	0.6291
prevOIAIDS	-4.62561	0.4787	462	-9.66	0.0000
AZTfailure	-0.2570	0.4725	462	-0.54	0.5868

Next, we extend model 3 by including an interaction effect between drug and obstime. Model 4 is shown in equation (18). The output in Table 3 report the $AIC = 7026.648$, this value differs by 6.644. Although the interaction effect is not significant, we keep it in the model in case the investigator is interested in it. We then would consider Model 4 with in the interaction term in our construction of the joint model. The interaction between drug and obstime did not have a significant effect on the CD4 cell count. PrevOIAIDS and obstime were once again significant predictors on CD4 cell counts. Each estimate is similar to the output from model 3 and our conclusion remain the same. We conclude, as the observation time increases by month, the CD4 cell count decreases by 0.1628. The CD4 cell count was higher

in the ddI group than in the ddC group by 0.3841 cells. Males had lower CD4 cell counts than women by 0.3180. The CD4 cell count decreases more for the patients with a previous diagnosis of AIDS than the patients with no previous diagnosis of AIDS at the study entry by 4.6281. Patients who experience a disease progression despite AZT therapy had lower CD4 cell counts than patients who were intolerant to AZT by 0.2538. As the observation time increases by month for the patients the CD4 cell counts are higher for the patients in the ddI group than the ddC group by 0.0217. Both outputs suggest that male patients in the ddI group, with a previous diagnosis of aids, and experienced a failure with AZT had worse survival outcomes.

$$y_{ij} = \beta_{11}+\beta_{12}(t_{ij})+\beta_{13}(t_{ij})\times Drug_i+\beta_{14}Gender_i+\beta_{15}PrevOI_i+\beta_{16}AZT_i+W_{1j}(t_{ij})+\varepsilon_{ij} \tag{18}$$

Table 3: Linear Mixed-Effects Model with Random Intercept and Slope with Interaction Effect

Full LME Model	AIC	BIC	logLik		
with Interaction	7026.648	7084.319	-3502.324		
Random Effect	StdDev	Corr			
Intercept	4.0036	(Intr)			
obstime	0.1734	-0.181			
Residuals	1.7488				
	Estimate	Std.Error	DF	T-Value	P-value
Intercept	10.4243	0.6906	936	15.09	0.0000
obstime	-0.1628	0.0210	936	-7.75	0.0000
drugddI	0.3841	0.3928	462	0.98	0.3287
drugddI:obstime	0.0217	0.0303	936	0.7146	0.4750
gendermale	-0.3180	0.6527	462	-0.49	0.6263
prevOIAIDS	-4.6281	0.4787	462	-9.67	0.0000
AZTfailure	-0.2538	0.4726	462	-0.54	0.5915

3.4 Survival Data Analysis

3.4.1 Summary Statistics

Next, we constructed a table of the number of patients at risk for each treatment at each observed time point (0, 2, 6, 12, 18 months) in Table 4. The number of patients at risk for the five time points are (230, 182, 153, 102, 20) for the ddI group and (237, 186, 157, 123, 14) for ddC group. The table demonstrates that the number of people at risk decreases at each time point for both treatments. The table also shows there is increasing missing rate due to death, dropouts, or missed visits.

In the beginning months (0-6), the number of patients at risk for ddI was less than the number of people at risk for ddC. We do note a strange occurrence at month 12 where the number of patients at risk for ddI decreases drastically and at 18 months the number of patients remaining in the ddI group is greater than the number of patients remaining in the ddC group. However, we need to consider these values proportional to the amount of patients at baseline in each group. Thus, converting the numbers of patients at risk, proportional to the amount of patients in each group we obtain (1, 0.79, 0.67, 0.44, 0.09) for ddI and (1, 0.78, 0.66, 0.52, 0.06) for ddC. Proportionally, for all time points except month 12, more people were at risk in the ddI group compared to the ddC group. At the end of the study, there were 20 patients remaining in the ddI group and 14 remaining in the ddC. This suggests that ddI may be better than ddC however, we note that there was not a big difference in the number of remaining patients between both groups.

Table 4: Number of patients at risk at months 0, 2,6,12, and 18 months

	Time(Months)				
No. At Risk	0	2	6	12	18
Didanosine(ddI)	230	182	153	103	20
Zalcitabine(ddC)	237	186	157	123	14

3.4.2 Kaplan-Meier Survival Curve

To further investigate the efficacy of the two treatments, we looked at the Kaplan-Meier estimates in Figure 3. The survival rates are plotted against the observation times and separated by treatment group. The plot displays the survival curve for ddC in blue and the survival rates for ddI in red. In accordance with the life table in Table 4, in the beginning months (0-6) we see similar survival curves for ddC and ddI. During months 6-18, the survival rates for patients in the ddC group is higher than the survival rates for the patients in ddI. Therefore, we suspect ddC was as effective as ddI in delaying disease progression and death. The patients assigned to ddC had a slightly better chance of surviving than patients assigned to ddI before 18 months. After 18 months we see the survival rate for patients assigned to ddC become worse than the survival rate for patients assigned to ddI. It is important to note that we have not found sufficient evidence to suggest that there is a difference in treatment effects. To investigate the treatments further, we constructed a Cox regression model to compare the ddI and ddC treatments while controlling other covariates in the model such as gender, prevOI, and AZT.

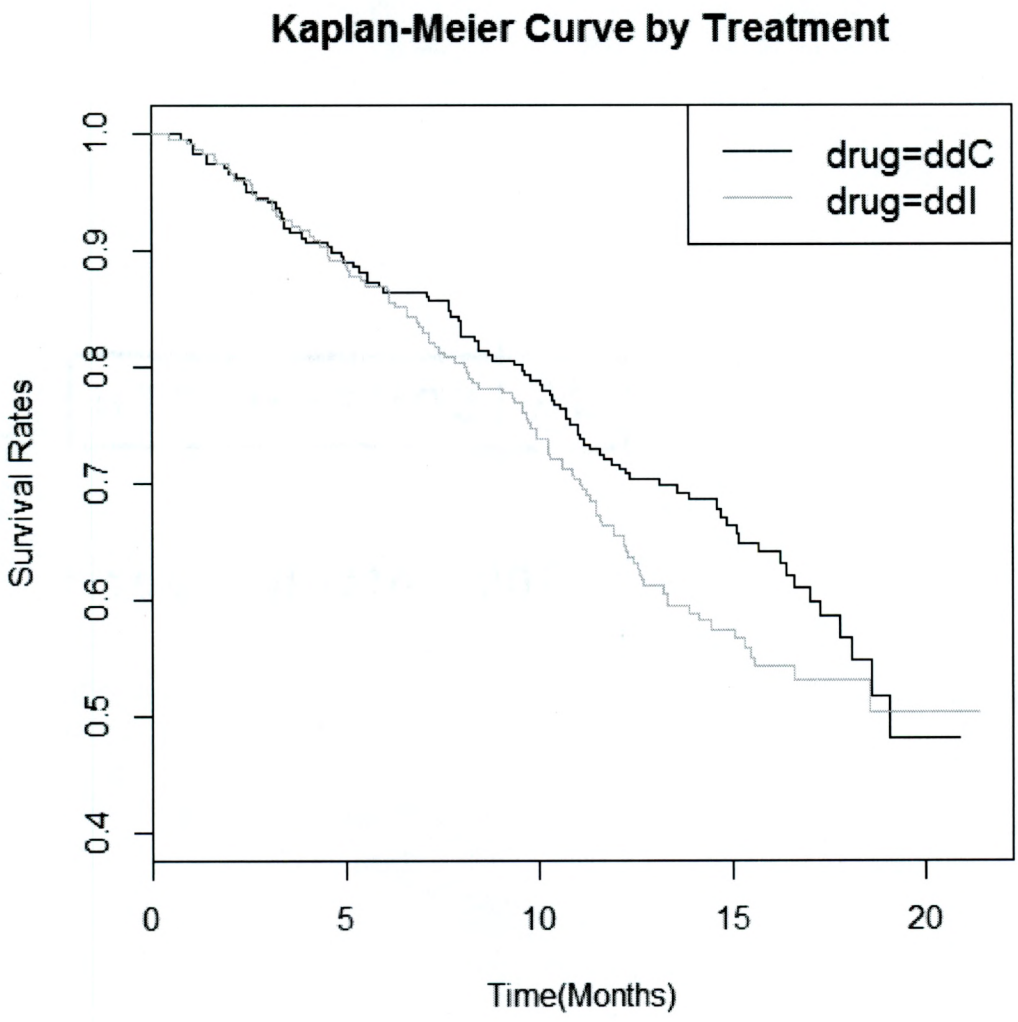

Figure 3: Kaplan-Meier survival curves for ddC (zalcitabine) and ddI (didanosine) from 0-24 months.

3.4.3 Cox-Proportional Hazard Model

In this section, we fit a cox-proportional hazard model to investigate the survival time on the drug term and other covariates gender, prevOI, and AZT. The event is defined as death. Using a significance level of 0.05, PrevOI is a significant predictor of hazard. We use the exponential of the estimates shown in column 2 in Table 5

to summarize our results. The expected hazard is 1.2423 times for the patients that were assigned ddI as compared to the patients assigned to ddC, while holding all other covariates constant. This indicates that the ddI group had worse survival than the ddC group, however this is not statistically significant. The male's expected hazard is 0.7104 times of females. The expected hazard is 3.6402 times for the patients with a previous opportunistic infection of AIDS diagnosis as compared to the patients with no previous diagnosis of AIDS. The expected hazard for patients with AZT failure is 1.1704 times of the patients that were intolerant to AZT. This suggests that the risk of death is greater in female patients who took drug ddI, had a previous diagnosis of AIDS at the beginning of the study and were failure to the drug AZT.

The covariates in the model are not time-varying therefore, the regression equation for the cox-proportional hazard model is given in equation (19).

$$\log(\lambda_i) = \beta_{21} + \beta_{22}Drug_i + \beta_{23}Gender_i + \beta_{24}PrevOI_i + \beta_{25}AZT_i \qquad (19)$$

Table 5: Expected Hazard on Parameter Effects

CoxPH Model	AIC 2113.514						
	Estimate	Exp(est)	SE(est)	Z	P-value	Lower L	Upper L
drugddI	0.2170	1.2423	0.1464	1.482	0.138	0.9324	1.655
gendermale	-0.3419	0.7104	0.2455	-1.393	0.164	0.4391	1.149
prevOI	1.2920	3.6402	0.2270	5.692	<0.0001	2.3330	5.680
AZTfailure	0.1575	1.1705	0.1634	0.964	0.335	0.8497	1.612

3.5 Joint Model

In the joint modeling, we combine the linear mixed-effects submodel (18) and cox-proportional hazard submodel (19)

$$\log(\lambda_i) = \beta_{21} + \beta_{22}Drug_i + \beta_{23}Gender_i + \beta_{24}PrevOI_i + \beta_{25}AZT_i + W_{2i}(t) \quad (20)$$

The association is between W_{1i} and W_{2i}, which linked the two processes together through the joint model.

In the longitudinal process, prevOI (p-value <0.0001, with 95% CI(-5.6315, -3.7498)) and observation time (p-value <0.0001, with 95% CI (-0.2239, -0.1398)) are still significant predictors of CD4 cell count at a 0.05 significance level. In accordance to the longitudinal results previously mentioned, for every one month increase in the observation time, the CD4 cell count decreases by 0.1819. The patients with a previous diagnosis of AIDS showed a decrease in CD4 cell count by 4.6906 compared to the patients without a previous diagnosis of AIDS. In the event process the treatment factor (p-value=0.0285, with 95% CI (1.0363,1.9044)), previous IO (p-value=0.0098 and 95% CI (1.1616,2.9791)), and the association component (p-value <0.0001, with 95% CI (0.7284,0.8407)) were significant predictors. The relative hazard is exp(0.3399)=1.40 for patients assigned to drug ddI as compared to 1.28 in the separate survival model. The relative hazard is exp(0.6207)=1.86 for the patients with a previous diagnosis of AIDS as compared to 2.19 in the separate survival model. The association term is the parameter that measures how strongly associated the CD4 cell count at any particular time point t is to the risk of death of disease progression. The association term (p-value <0.0001) is significant; the CD4 cell count was correlated to the risk of death or disease progression. The joint model parameter estimates are similar to the separate longitudinal and survival parameter estimates.

Table 6: Joint Model of Longitudinal and Time-to-Event

Joint Model	AIC	BIC	logLik			
	8546.979	8621.613	-4255.489			
Longitudinal Process						
	Estimate	Std.Error	Z-value	P-value	Lower L	Upper L
Intercept	10.4173	0.6912	15.0721	<0.0001	9.0626	11.7719
drugddI	0.3909	0.3911	0.9997	0.3175	-0.3755	1.1574
gendermale	-0.2515	0.6536	-0.3847	0.7004	-1.5326	1.0296
prevOI	-4.6906	0.4800	-9.7719	<0.0001	-5.6315	-3.7498
AZTfailure	-0.2779	0.4738	-0.5865	0.5575	-1.2065	0.6507
obstime	-0.1819	0.0214	-8.4790	<0.0001	-0.2239	-0.1398
drugddI:obstime	0.0103	0.0304	0.3386	0.7349	-0.0492	0.0698
Event Process						
	Estimate	Std.Error	Z-value	P-value	Lower L	Upper L
Intercept	-3.5044	0.4468	-7.8441	<0.0001	0.0125	0.0722
drugddI	0.3399	0.1552	2.1899	0.0285	1.0363	1.9044
gendermale	-0.3742	0.2573	-1.4542	0.1459	0.4153	1.1390
prevIOAIDS	0.6207	0.2403	2.5834	0.0098	1.1616	2.9791
AZTfailure	0.0904	0.1687	0.5360	0.5920	0.7865	1.5234
Assoct	-0.2452	0.0366	-6.7038	<0.0001	0.7284	0.8407

4 Model Selection and Accuracy

For model comparison we examine the AIC criterion of the joint model to see if it is smaller than the combination of the AIC criterion from the separate models. The joint model has an AIC of 8546.979 and the separate models together have an AIC= 9140.162 (7026.648 + 2113.514). The joint model does in fact do better than the separate approaches because the AIC criterion for the joint model is smaller than the AIC criterion for the separate models.

4.1 Diagnostics

We conducted model diagnostics to assess the validity of the joint model. We can check if a model works well for data in many different ways such as residual plots,

R^2 or AIC criterion that tell us how well a model fits the given data. We also assess the validity in exploration of the model's underlying statistical assumptions, an examination of the structure of the model by considering formulations that have fewer, more, or different explanatory variables, or looking for influential points that does not fit well represented by the model (outliers) or that have a relatively large effect on the model's predictions. Figure 4 showed a plot of the subject-specific residuals against the fitted values and a normal Q-Q plot of the standardize residuals and theoretical quantiles. We performed a Shapiro Wilk test of normality and plotted the marginal residuals against the fitted values to examine the validity of the Joint model. We also plotted the marginal survival curve and the marginal cumulative hazard rates against time.

The residuals vs fitted in top left corner in Figure 4 is a subject-specific residual plot. From this plot, we can see the residual vs fitted plot shows a constraint. In the left lower corner the residuals have a 45 degree downward slope, therefore the residuals are not random. The Normal Q-Q plot on the top right corner in Figure 4 is also a subject-level Q-Q plot. The Q-Q plot shows most of the residuals follow a straight line pattern but seems to have some extreme values deviated from the straight line pattern. With careful examination of the tail residuals shows strong evidence against Normality. These are not just extreme values as we may have thought. The pattern is consistent with a heavy-tailed distribution (such as a t-distribution).

The Shapiro-Wilk test of normality was used to determine if the residuals of the longitudinal process in the joint model meet the normality assumption. The test produced a test statistic W=0.9478 (p-value<0.0001), indicating the residuals are not from a normally distributed population. However, since the test is biased by sample size, the test may be statistically significant from a normal distribution in any large samples. Thus, we use the Q-Q plot in the top right corner for verification in

addition to the test.

From the marginal residual plot in Figure 5, we can see for small fitted values we have more positive than negative residuals. Small fitted values correspond to lower levels of square root CD4 cell count, which corresponds to a worsening of the patient's condition and subject to higher chance of dropout. Thus, the residuals corresponding to small fitted values are only based on patients with a 'good' health condition.

The problem in the residual plot and Q-Q plot is that the distribution of the residuals for the longitudinal process is affected by the dropout caused by the occurrence of events. When a patient experiences the event, it corresponds to a discontinuation of the collection of longitudinal information because either follow up measurements can no longer be collected or their distribution changes after the event occurred. The dropout mechanism implied by joint models is of a nonrandom nature, which implies the observed data, upon which the residuals are calculated, do not constitute a random sample of the target population [24]. This implies the residuals plots based on the observed data alone can be misleading because these residuals should not be expected to exhibit standard properties, such as zero mean and independence.

Even if we think there is a problem with the normality assumption, the inference is still valid because the joint model is a more robust procedure than the classical separate longitudinal and survival analysis and the sample size is large for this data. You can see that the model is robust because the Q-Q plot demonstrated heavy tailed distribution of residuals such as the t-distribution. A robust joint model still provides accurate inference on the study objectives despite having its assumptions altered or violated. A robust joint model works better than the separate models when a greater proportion of more extreme longitudinal outliers are present. Through t-distribution assumptions, longitudinal outliers are accommodated with their detrimental impact being down weighed and thus providing more efficient and reliable estimate.

Furthermore, the Marginal Survival curves shows a decreasing trend; as time increases the survival rate decreases. Lastly, examining the Marginal Cumulative Hazard curve, there is an increasing trend over time. As time increases the cumulative hazard rate increases exponentially. Both plots indicate that the disease progressed for the patients in the study.

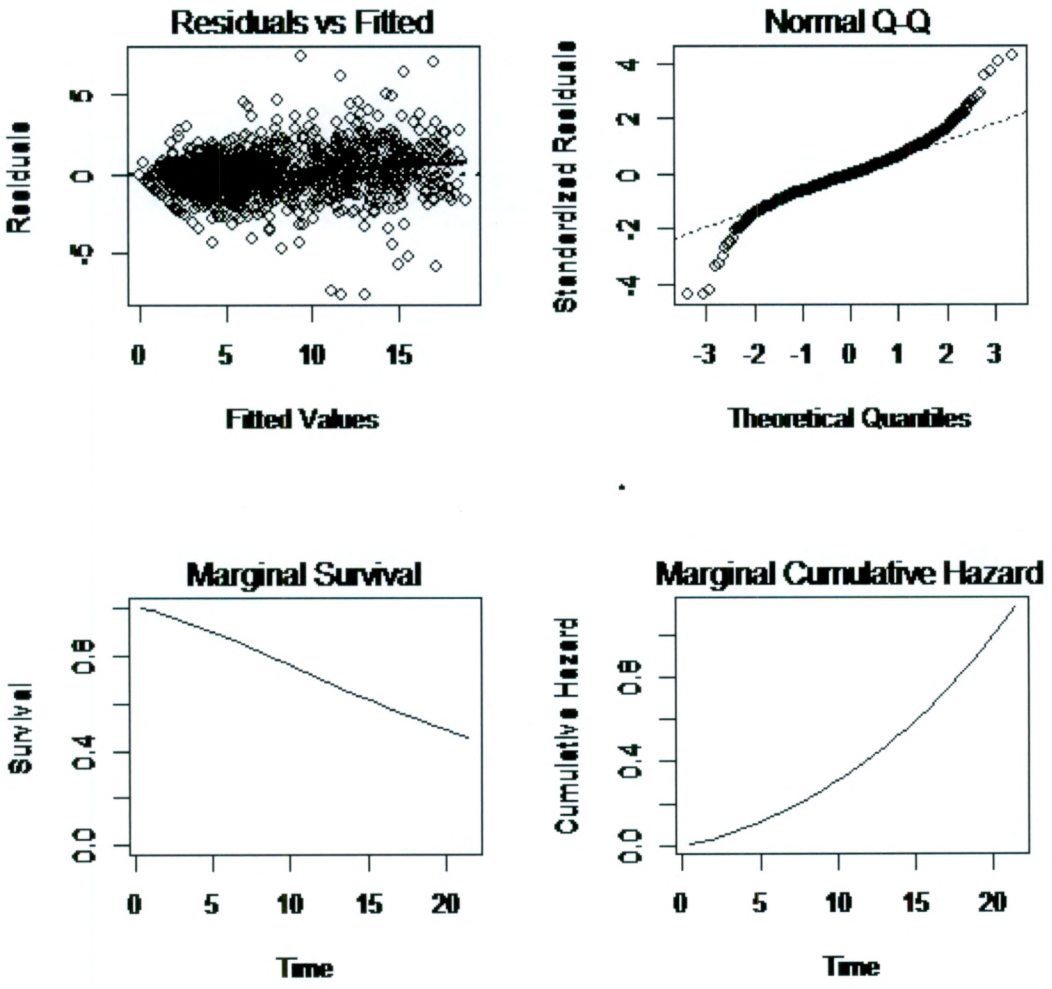

Figure 4: Residuals for the Joint Model Fit

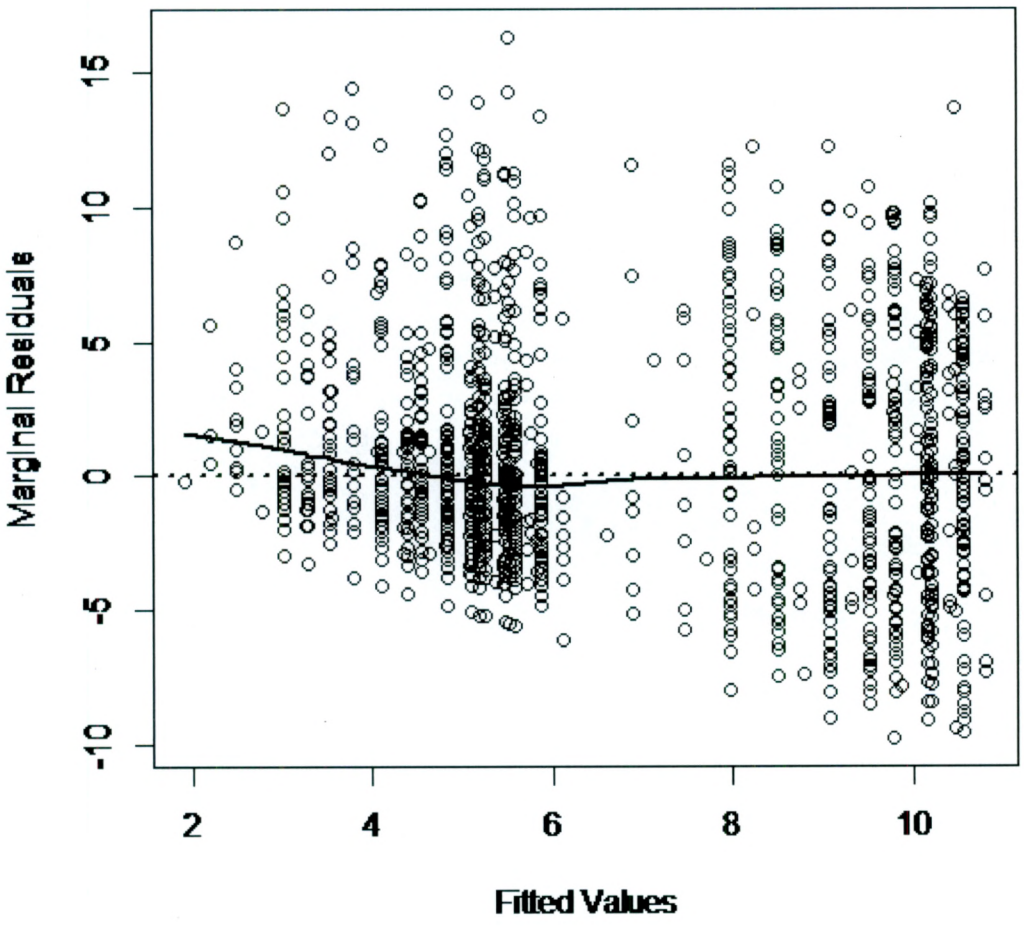

Figure 5: Marginal Residuals vs Fitted Values for the Joint Model Fit

5 Discussion and Conclusion

To summarize, previous authors have used separate analyses of longitudinal and time-to-event data. Common issues in longitudinal studies is that the data suffers from attrition, which can cause bias in the analysis if the dropout are informative. Also

associations between longitudinal and survival data can occur in the explanatory variables or through stochastic dependence between the subject-specific random effects component of the longitudinal model and of the survival model. We used a joint model to account for these issues and simultaneously analyze the longitudinal and time-to-event data. Our approach is important in many bio-statistical application areas, because we obtain accurate inference regarding longitudinal responses while adjusting for outcome-dependent study dropout. We can also apply these ideas to problems involving surrogate markers, where the focus is on using longitudinal measurements to improve prediction of survival prognosis.

We used the HIV study to investigate the efficacy and safety of the two drugs, Didanosine (ddI) and Zalcitabine (ddC), and how the CD4 cell counts changed over time. In the analysis of the HIV data, we compared separate analyses of the longitudinal model and survival model to the joint model. The separate analysis of the LME model with a random intercept and slope of obstime with all four covariates showed that observation time (p-value=0) had a significant negative effect on CD4 cell counts, indicating that for every one month increase for the observed time, the CD4 cell count decreased. PrevOIAIDS (p-value=0) had a significant negative effect on CD4 cell count, indicating that CD4 cell counts decrease for patients with a previous diagnosis of AIDS compared to patients with no previous diagnosis of AIDS at the study entry. The separate cox-proportional hazard model was used to investigate the effect the drug groups and additional covariates on the hazard. The cox-proportional hazard model showed prevOIAIDS was significant at the 0.05 significance level. We note that the drug term was not significant at 0.05 significance level but significant at 0.10 significance level. The expected hazard was 2.1886 times higher for patients who were diagnosed with AIDS compared to patients who didn't have a previous diagnosis of AIDS at baseline.

The joint model showed that prevOIAIDS and obstime were significant predictors of the change in CD4 cell count at a 0.05 significance level. Having a previous diagnosis of AIDS as time went on in the study led to a decrease in CD4 cell counts which means a patient was more likely to die or the disease progressed. The risk of death for patients assigned to ddI was 1.44 times as likely compared to the patients assigned to ddC with a significant p-value at 0.05 significance level. The association term between the CD4 count and the survival process (p-value <0.0001) is significant; the CD4 cell count was correlated to the risk of death. The joint model parameter estimates are similar to the separate longitudinal and survival parameter estimates. It also showed the two drug groups had significant hazard rate when controlling other covariates in the joint model. Model diagnostic showed the joint model fit the data well. With the model diagnostic results and a significant association between the longitudinal process and survival process, the joint model approach provides valid and accurate results for the HIV study. The joint model can be readily fit using the jointModel function under the JM package in R, thus avoiding the need for complex EM programming. This makes it more convenient to use in real data analysis.

6 References

References

[1] Abrams Donald I., Goldman Anne I, LaunerCynthia, Korvick Joyce A, Neaton James D, Crane Lawrence R., Grodesky Michael, Wakefield Steven, Muth Katherine, Kornegay Sandra, Cohn David L, Harris Allen, Luskin-Hawk Roberta, Markowitz Norman, Sampson James H, Thompson Melanie, and Deyton Lawrence. *A Comparative Trial of Didanosine or Zalcitabine after Treatment with Zidovudine in Patients with Human Immunodeficiency Virus Infection.* The New England Journal of Medicine 330.10, (1994): 657-62.

[2] Asar O, Ritchie J, Kalra P.A, and Diggle P.J. *Joint Modelling of Repeated Measurement and Time-to-event Data: An Introductory Tutorial* International Journal of Epidemiology 44.1 (2015): 334-44.

[3] DeGruttola, V. and Tu, X.M. *Modeling Progression of CD4 Lymphocyte Count and its Relationship to Survival Time* Biometrika 80 (1994):1003-1014.

[4] Diggle P.J., Heagerty P.J., Liang K-Y, and Zeger S.L *Analysis of Longitudinal Data.* Oxford University Press, (2002):2.

[5] Diggle P.J., Kenward M G. *Informative Dropout in Longitudinal Data Analysis.* Applied Statistics, 43, (1994):49-94.

[6] Faucett Cheryl L., and Thomas Duncan C. *Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates: A Gibbs Sampling Approach.* Statistics in Medicine, 15, (1996):1663-1685.

[7] Guo Xu, Carlin Bradley P. *Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages.* The American Statistician. 51.1, (2004): 1.

[8] Guo Wensheng, Radcliffe Sarah J., Ten Have Thomas T. *A Random-Pattern Mixed Model for Longitudinal Data with Dropouts.* Journal of the American Statistical Association, 99.468 (2004):929-937.

[9] Henderson Robin, Diggle Peter, and Dobson Angela. *Joint modelling of longitudinal measurements and event time data.* Biostatistics 1.4 (2000) 465-480.

[10] Hogan J and Laird N. *Mixture Models for the Joint Distribution of Repeated Measurements and Event Times.* Statistics in Medicine 16, (1997a):239-257.

[11] Hogan J and Laird N. *Model-Based Approaches to Analyzing Incomplete Longitudinal and Failure Time Data.* Statistics in Medicine 16, (1997b):259-272

[12] Hsieh Fushing, Tseng Yi-Kuan, and Wang Jane-Ling. *Joint Modeling of Survival and Longitudinal Data: Likelihood Approach Revisited.* Biometrics, 62,(2006):1037-1043.

[13] Lang Wu, Wei Liu, Grace Y. Yi, and Yangxin Huang. *Analysis of Longitudinal and Survival Data: Joint Modeling, Inference Methods, and Issues.* Journal of Probability and Statistics, (2012):17.

[14] Laird N, Ware J *Random-Effects Models for Longitudinal Data.* Biometrics,38 (1982):963-974.

[15] Lim Hyun J., Prosanta Mondal, and Stuart Skinner. *Joint Modeling of Longitudinal and Event Time Data: Application to HIV Study.* Journal of Medical Statistics and Informatics, 1.1 (2013): 1.

[16] Li Ning, Robert M. Elashoff, and Gang Li. *Robust Joint Modeling of Longitudinal Measurements and Competing Risks Failure Time Data* Biometrical journal. Biometrische Zeitschrift 51.1 (2009): 1930.

[17] R Core Team *@Manual, R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, (2017), https://www.R-project.org/

[18] Rizopoulos Dimitris *Package "JM"-Joint Modeling of Longitudinal and Survival Data.* 1, (2016):4-5

[19] Rizopoulos Dimitris *JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data.* Journal of Statistical Software, (2010):35(9), 1-33. URL http://www.jstatsoft.org/v35/i09/.

[20] Song Xiao, and Wang C.Y. *Semiparametric Approaches for Joint Modeling of Longitudinal and Survival Data with Time-Varying Coefficients.* Biometrics 64,(2008), 557-566.

[21] Sousa Ines. *A Review on Joint Modelling of Longitudinal Measurements and Time-to-Event.* REVSTAT- Statistical Journal, 9.01, (2011): 57-81.

[22] Tsiatis A.A, DeGruttola Victor, Wulfsohn M.S. *Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS.* Journal of the American Statistical Association, 90.429 (1995):27-37

[23] Wulfsohn M. S and Tsiatis, A. A. *A Joint Model for Survival and Longitudinal Data Measured with Error* Biometrics 53 (1997):330-339

[24] Verbeke Geert, Molenberghs Geert, and Beunckens Caroline *Formal and Informal Model Selection with Incomplete Data.* Statistical Science, 23.2, (2008):201-208