Stockholms universitet

Joint Longitudinal and Survival Models to Predict Survival Outcomes

Julia Eriksson

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2021:2
Matematisk statistik
Februari 2021

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematiska institutionen

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2021:2**
http://www.math.su.se

Joint Longitudinal and Survival Models to Predict Survival Outcomes

Julia Eriksson*

February 2021

Abstract

Survival analysis is the common name of statistical methods where the time until an event is analysed. These methods are used extensively in medical research to analyse, for example, time until death or the development of a disease over time. Longitudinal data consists of repeated measurements taken over a period of time, for example blood pressure. Combining the Cox regression model used to analyse survival data with the linear mixed effect model for longitudinal data results in the joint longitudinal and survival model. In this thesis, the joint model is applied to a subset from the AMORIS (Apolipoprotein related mortality risk) cohort. The AMORIS cohort contains observations from subjects, collected between 1985 and 1996 in Stockholm, who provided blood and urine samples which were analysed. The subset used in this thesis includes all men aged 40-50 at observation who provided measurements of the four longitudinal biomarkers: Apolipoprotein A, Apolipoprotein B, total cholesterol and triglycerides. This resulted in a dataset containing 33 930 observations from 23 768 subjects. The joint model was applied to these data for each longitudinal biomarker separately, where the time until event in the survival submodel was if the subjects had died at the end of study. The Cox model and linear mixed effect model were fitted separately and then applied to the joint model following the adaptive Gauss-Hermite quadrature rule. The application of the joint model on these data allowed to predict conditional survival probabilities for the subjects who were still alive at the end of the study, following a Monte Carlo simulation scheme. In addition to the survival probabilities, subject specific dynamic survival probabilities were predicted, that is, how the survival probability change over time as more longitudinal observations are obtained. Martingale residuals for the survival part and marginal and subject specific residuals for the longitudinal part in the joint model were also computed and illustrated. The result of the joint model fit to the data indicated that a one unit increase of each of the four longitudinal biomarkers increase the risk of death.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: juliaagnes.eriksson@gmail.com. Supervisor: Taras Bodnar.

Acknowledgements

I would like to thank my external supervisor Professor Matteo Bottai at Karolinska Institutet who introduced me to the theory of predictive joint longitudinal and survival models. Thank you for your time and guidance throughout the process of writing this thesis. A warm thanks to the Unit of Biostatistics at Karolinska Institutet for welcoming me to your institution.

Thank you to Niklas Hammar for kindly letting me use the AMORIS data cohort and for your time to acquaint me with the data. A warm thanks to Mats Talbäck for the help to access and work with the data.

I am very grateful to my supervisor at Stockholm University, Professor Taras Bodnar, who, at numerous times, have read and commented my thesis and provided valuable knowledge.

Finally, I would like to thank my parents and our cat for the support and for keeping me company throughout the process of writing my thesis during this autumn in the pandemic.

Contents

List of Tables

List of Figures

1 Introduction

The desire to measure the time until a certain event goes back to as early as the 17th century when in 1662, John Graunt made the first life table [5]. A life table contains probabilities of survival before the next birthday for people at each age [26]. The life table was later developed a century later by Daniel Bernoulli in 1760 as he wanted to calculate and predict the life expectancy if the disease smallpox became extinct [25]. Time-to-event analysis or survival analysis has since been developed and used extensively in medical research to calculate, for example, time until death or time until relapse of a disease.

Common in research is that repeated measurements are observed for the same individual at different time points. In this thesis, these individuals that are a part of the analyses are referred to as subjects. In medical research, repeated measurements could for example be blood pressure or cholesterol levels. These measurements are known as longitudinal biomarkers [54]. The need to perform an analysis of change is essential in many research fields and one of the first methods used to measure change was the analysis of variance known as ANOVA. The original method for analysing longitudinal data was a mixed effects ANOVA that included a single random subject effect which resulted in a positive correlation between the repeated longitudinal measurements for each subject [21]. The British astronomer George Biddel Airy was the first to define a linear mixed model in 1861 [3], when he developed a model for the errors from the astronomy observations. This model was later defined theoretically by R. A. Fisher in 1918 [19] and 1925 [20] within the ANOVA model.

The joint longitudinal and survival model combines the survival analysis with the longitudinal observations allowing to make predictions of survival based on the longitudinal observations, when these are related to one another [54]. In a medical study with the aim to examine the treatment effect of a particular disease, the joint model will provide estimates of the treatment effects of the longitudinal markers, estimate the effects of the treatment on the time to event as well as reducing the bias of the estimates on how the treatment effects the survival and longitudinal markers [33]. Two early articles that are the basis of the development of joint longitudinal and survival models are DeGruttola and Tu in 1994 [11] and Tsiatis et al. in 1995 [72]. In these articles, the joint model was originally motivated by medical research on human immunodeficiency virus (HIV) where the aim was to examine the association between CD4 cell counts, which is an indicator of the wealth in the immune system that decreases for people diagnosed with HIV [32], and the time until the subjects were infected by Acquired immune deficiency syndrome (AIDS) [54].

1.1 Aim of thesis

The aim of this thesis is to learn about the statistical analysis method called joint longitudinal and survival models and to be able to implement this model on real data. The data are provided by Karolinska Institutet in Stockholm from the AMORIS (Apolipoprotein related mortality risk) cohort. The cohort contains observations from 812 073 subjects, 51 % women and 49 % men, collected between the years 1985 and 1996 [18]. The subjects provided blood and urine samples which were analysed in a lab resulting in 35 million values [16].

The intention is to apply the joint model on this data to predict the survival outcomes for a number of subjects based on the longitudinal biomarkers Apolipoprotein A, Apolipoprotein B, total cholesterol and triglycerides.

1.2 Examples of applications using joint longitudinal and survival models

Joint longitudinal and survival models are methods used for statistical analyses in many fields. In this section, some empirical examples are presented.

1.2.1 Ecology

A study performed by Lee et al. [40] in 2011 used joint longitudinal and time-to-event models to study tree growth and mortality where, for each tree, a latent feature of the tree has an influence on the growth and mortality. In this study, the trees are observed intermittently, meaning that the observation can cease for a time such that the measurements are not observed continuously [24]. This means that the exact time of events are not known and in this study the time between longitudinal observations can be more than a decade. The conclusion that was drawn from this study was that in an intermittent observation process, there are advantages to trace a longitudinal marker for imputation of lifetimes.

1.2.2 Insurance

In order to assess a reasonable insurance, the insurance companies collect information about their policyholders, in particular, information about their medical claims. A study by Piulachs et al. (2017) [49], used joint models as they explored the relationship between a policyholder's, aged 65 or older, demand of medical emergency claims per year, the longitudinal observation, and the time until death. The results obtained from this study were that a relatively high cumulative need for hospitalisation, the number of non-routine healthcare visits and the use of ambulance have a positive relationship with the health status for each subject as well as a higher mortality

risk. They also observed that a subject's most recent demand for critical medical care has the strongest influence on the survival.

1.2.3 Medicine

A web-based calculator was implemented by Taylor et al. (2013) [68] using joint longitudinal and survival models. The calculator was implemented to predict the probability of recurrence of prostate cancer for a new patient. Patients who have been treated for prostate cancer using radiation therapy attend regular check ups where they performed a laboratory test named Prostate Specific Antigen (PSA). An indication of a recurrence of prostate cancer is detected through these tests as the value of PSA increases. A group of these patients were used in this research to implement the calculator so that it will be able to accurately estimate the probability of a recurrence of prostate cancer.

1.2.4 Socio-economic status

A study about socio-economic inequalities was performed by Maharani (2019) [42] where the aim was to investigate how the association between socio-economic status and inflammation influence health in a population. Earlier research had indicated that inflammation and socio-economic status are associated with the health status in high-income countries. The longitudinal measurements in this study were the amount of C-reactive protein (CRP) and the education and wealth status for each individual in England and Indonesia. A C-reactive protein is produced in the liver and sent out to the body when there is an inflammation. A high level of CRP indicates that the body has an inflammation [78]. The results obtain was that in England, higher education and wealth were associated with a lover CRP value. In Indonesia, the socio-economic status had no significant relation with the CRP. Both countries indicated that physical activity is associated with lower CRP values.

1.3 Structure of the thesis

The structure of the thesis is as follows. In the first three sections, the theory of the joint longitudinal and survival model is described. In particular, the theory of survival analysis is introduced in Section 2 followed by the theory of longitudinal analysis in Section 3 and in Section 4 the joint model is defined. Section 5 presents the analysis of the data where the joint model is implemented on the AMORIS data with the results illustrated in Section 6. Final remarks and possibilities of future research are provided in Section 7.

2 Survival Analysis

Survival data are a common measure when the aim is to study the time to an event. In medical research this event could be the time from birth until death, time until diagnosis of a disease to death or time from entry of a study until relapse. Some other examples are time from marriage to divorce, time from falling in love to the birth of first child, time from falling a sleep to waking up or the time for a light bulb to break [1]. Survival analysis, or time-to-event analysis measures the time until an event given a start time. These events are denoted as the time of failure [38].

In this section we delve into the fundamentals of survival analysis where we will get familiar with censoring and truncation, we define the four basic functions and then regard some methods and models to analyse survival data.

2.1 Censoring

This section follows Chapter 3 in Klein and Moeschberger's book *Survival Analysis: Techniques for Censored and Truncated Data* [38]. Analysing survival times are challenging and often not possible using linear regression or other statistical methods for data analysis. This is due to the fact that collecting data of survival times results in both complete and incomplete data. The incomplete data are denoted as censored survival times. There are three main types of censored times; right, left, and interval censoring. Censored data must be included in the analysis, otherwise the study is not complete which can lead to biased results.

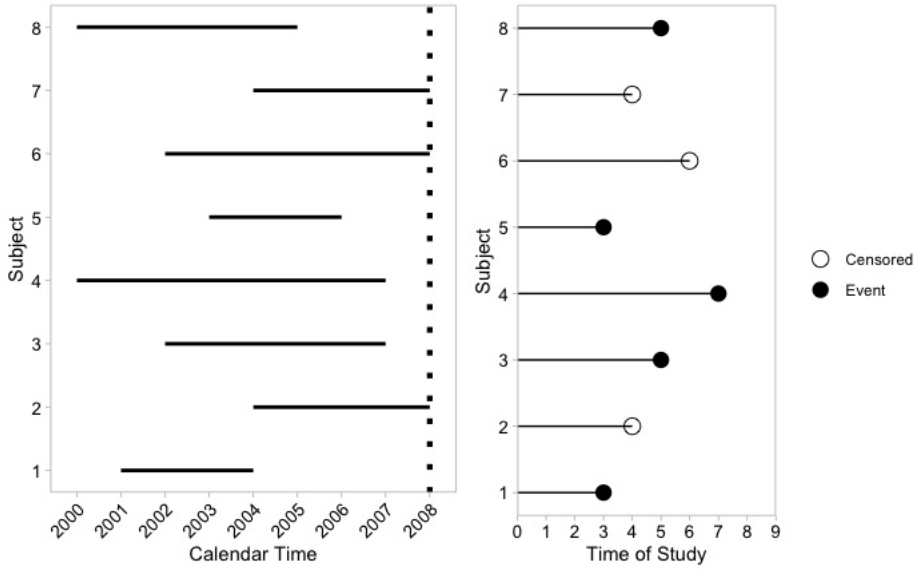
In a study where the time to an event is observed, not all subjects will have experienced the event at the end of the study. For example, assume a medical study where the time until infection is observed for 8 subjects. During this observation time, 5 subjects have been infected and hence experienced the event. The remaining 3 were not infected during this observation time. However, they might become infected later, after the end of the study, but this information will not be included in the study. These incomplete survival time observations are *right censored* C_r survival times. A visual example of right censoring is presented in Figure 1. In the left graph of the figure the entry time of the 8 subjects are displayed and the dotted vertical line indicates the end of the study. The right graph illustrates the individual time in the study for each subject where time 0 is the entry time. A filled circle is noted for 5 of the subject which indicates that they experienced the event before the end of the study. The end time of the remaining subjects are indicated with a white circle, which means that they did not experience the event during the time of study. These subjects are right-censored

Left censoring, C_l occurs if for example a subject is infected prior to the onset of the study and the time of infection is not noted.

When the study includes a follow-up after a certain time of an event, for example after an operation the follow-up is one year after, and a recurrence is detected a year after, the exact time point of the recurrence is not known. The only information known is that the recurrence happened in the interval of the left and right endpoint of the censored time $(L_i, R_i]$. This is known as *interval censoring*.

The censoring times described above are known as non-informative meaning that they are not related to the survival times of the individuals. *Informative censoring* is when certain individuals are removed from the study which could be due to a not wanted result of a treatment. Then the results must be analysed with caution.

Figure 1: The left graphs illustrates the entry time of a study of 8 subjects. The dotted vertical line indicates the end of the study. The right graph illustrates the time each subject was in the study and time 0 is the entry time. A filled circle indicates that the subject experienced the event before the end of the study. A white circle indicates that the subject did not experience the event during the time of study, these subjects are right-censored.



2.2 Truncation

For censored survival times, some information about the subject is known even if the data of this individual are incomplete, however these data are included in the observation. Truncated survival data are when only the observations where the event has occurred in a specified time period (Y_L, Y_R) are included. That is, all the subjects that did not experience the event in the given time interval are excluded [38].

If the time T of an event occurred after $Y_L, T > Y_L$ these individuals are only observed but not included in the survival data, this is known as *left truncation*. If time T of an event occurred before or at $Y_R, T \leq Y_R$, it is denoted *right truncation* [38].

2.3 Fundamental functions in survival analysis

Denote T as a nonnegative continuously distributed random variable that represents the time until an event, for example the time until onset of a disease, time until recovery or time until death. Then we can define four basic relationships that characterise the distribution of the random variable T. This section mainly follows Chapter 2 in Klein and Moeschberger (2003) [38], if nothing else is stated.

2.3.1 Cumulative distribution function

Consider a study that consists of a number of patients where the time until a possible event is observed. The start of the study is at time $t = 0$ and continues until time t. The cumulative distribution function of a random variable T is defined as the probability that the time of survival is less than time t, that is, the probability that the event has happened at time t [6]. It is given by

$$F(t) = P(T \leq t) = \int_0^t f(v)dv, \tag{2.1}$$

where $f(v)$ is the probability density function of T.

2.3.2 The survival function

The survival function,

$$S(t) = P(T > t) = 1 - F(t) \tag{2.2}$$

describes the proportion of the subjects that have not experienced the event at time t. These individuals have "survived" the event up until the time point t.

The survival function can also be written as the integral of the probability density function

$$S(t) = \int_t^\infty f(v)dv, \tag{2.3}$$

which can further be expressed as

$$f(t) = -\frac{dS(t)}{dt}. \tag{2.4}$$

This expression can be used to assess the probability that the event has happened in a small interval around time t [38].

Usually, as we increase t, the survival function will go to zero as most of the individuals will, after a certain time, experience the event. However, the survival function of events that, for example, are gender-related will tend to a positive value as we increase t as not every individual in the population will experience the event [1].

An important aspect that was addressed by Crowther (2014) [8] is the time of when a subject becomes at risk. The survival function assumes that the subject is at risk at time 0, but this is not always the case. Often in medical studies, a subject enters the study at time t_0, and then age is used as a timescale such that the age of when the subject is diagnosed correspond to the time when the subject is at risk. Then the survival function (2.2) must be conditioned on the survival of the event up to the entry time t_0. The probability of surviving up to time t, given survival up to time t_0 is

$$P(T > t|T > t_0) = \frac{S(t)}{S(t_0)}. \tag{2.5}$$

2.3.3 The hazard function

The hazard function $h(t)$ is defined as the probability of experiencing an event in the time interval $[t, t + \Delta t)$ conditioned on the unconditional probability that an event has not happened at time t. The function is expressed as [38]

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t|T \geq t)}{\Delta t}, \tag{2.6}$$

which can further be formulated as [38],

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \cap T \geq t)}{\Delta t P(T \geq t)} = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t P(T \geq t)} \\ &= \frac{f(t)}{S(t)} = -\frac{\frac{d}{dt}S(t)}{S(t)} = -\frac{d \log S(t)}{dt}. \end{aligned} \tag{2.7}$$

The last equality is interpreted as the failure rate at time t.

2.3.4 The cumulative hazard function

The cumulative hazard function is the integral of the hazard function and by (2.7) it can be written as [38]

$$H(t) = \int_0^t h(v)dv = -\log S(t). \tag{2.8}$$

This means that we can rewrite the survival function in terms of the hazard function as

$$S(t) = \exp\left\{-\int_0^t h(v)dv\right\}. \tag{2.9}$$

Then by Equation (2.7) we can present the probability density function as

$$f(t) = h(t)S(t). \tag{2.10}$$

2.4 Non-parametric models for survival analysis

There are two non-parametric estimators that are commonly used to fit survival data. From a sample of censored survival data, the Nelson-Aalen estimator estimates the cumulative hazard rate and the Kaplan-Meier estimator estimates the survival function. In this section these two estimators are described and illustrated, following Aalen et al. (2008) [1].

2.4.1 The Nelson-Aalen estimator

To define the Nelson-Aalen estimator we first assume a sample n of censored survival data with $Q(t)$ as the number of individuals at risk at time point t and the ordered event times are denoted by $T_1 < T_2 < \cdots < T_J$. The hazard function $h(t)$ can be cumbersome to estimate if no parametric assumptions are made, as it then can be any nonnegative function. However, in Equation (2.8), the cumulative hazard function was calculated as the integral of the hazard function, where no parametric assumptions about $h(t)$ were made. This is similar to estimate the cumulative distribution function rather than to estimate the density function, which can be far more troublesome. This leads us to the Nelson-Aalen estimator, which proposes to estimate the cumulative hazard rate in the following way,

$$\hat{H}_{NA}(t) = \sum_{T_j \leq t} \frac{1}{Q(T_j)}. \tag{2.11}$$

The estimated variance for the estimator is

$$\hat{\sigma}^2_{NA}(t) = \sum_{T_j \leq t} \frac{1}{Q(T_j)^2}, \tag{2.12}$$

with a $100(1-\alpha)\%$ confidence interval given by

$$\hat{H}_{NA}(t) \pm z_{1-\alpha/2}\hat{\sigma}_{NA}(t), \tag{2.13}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. The interval can also be written on the log scale to obtain a better approximation to the normal distribution,

$$\hat{H}_{NA}(t)\exp\{\pm z_{1-\alpha/2}\hat{\sigma}_{NA}(t)/\hat{H}_{NA}(t)\}, \tag{2.14}$$

which results in better approximations for small samples.

If two events occur at the same time point, these are called tied events. If there are two or more events at a given time point T_j, the estimator must be extended to include the total number of events o_j at time T_j, where we assume that the event times are discrete:

$$\hat{H}_{NA}(t) = \sum_{T_j \le t} \frac{o_j}{Q(T_j)}, \tag{2.15}$$

with the estimated variance

$$\hat{\sigma}^2_{NA}(t) = \sum_{T_j \le t} \frac{(Q(T_j)-o_j)o_j}{Q(T_j)^3}. \tag{2.16}$$

2.4.2 The Kaplan-Meier estimator

The Kaplan-Meier estimator is defined as

$$\hat{S}_{KM}(t) = \prod_{T_j \le t} \left\{1 - \frac{1}{Q(T_j)}\right\}. \tag{2.17}$$

The estimator follows from the survival function $S(t)$ by dividing the time interval $[0,t]$ into smaller parts $0 = t_0 < t_1 < \cdots < t_K = t$ such that the survival function can be written as

$$S(t) = \prod_{k=1}^{K} S(t_k|t_{k-1}), \tag{2.18}$$

where $S(t_k|t_{k-1})$ is the conditional probability of the event to occur at time t_k given that it has not occurred at time t_{k-1}. The small intervals $0 = t_0 < t_1 < \cdots < t_K = t$ contains no more than one event so the estimate of $S(t_k|t_{k-1})$ will be the following,

$$S(t_k|t_{k-1}) = \begin{cases} 1, & \text{if no event in } (t_{k-1}, t_k] \\ 1 - \frac{1}{Q(t_{k-1})} = 1 - \frac{1}{Q(T_j)}, & \text{if an event at time } T_j \in (t_{k-1}, t_k]. \end{cases}$$

The estimated variance of the Kaplan-Meier estimator is given by

$$\hat{\sigma}^2_{KM}(t) = \hat{S}_{KM}(t)^2 \sum_{T_j \le t} \frac{1}{Q(T_j)^2}. \tag{2.19}$$

A more common way to estimate the variance of the Kaplan-Meier estimator is by Greenwood's formula which is expressed as

$$\hat{\sigma}^2_{KM}(t) = \hat{S}_{KM}(t)^2 \sum_{T_j \leq t} \frac{1}{Q(T_j)\{Q(T_j) - 1\}}. \tag{2.20}$$

The $100(1-\alpha)\%$ confidence interval is given by

$$\hat{S}_{KM}(t) \pm z_{1-\alpha/2}\hat{\sigma}_{KM}(t), \tag{2.21}$$

The interval can also be written by using the log-minus-log transformation to obtain a better approximation to the normal distribution as

$$\hat{S}_{KM}(t)^{\exp\{\pm z_{1-\alpha/2}\hat{\sigma}(t)/(\hat{S}_{KM}(t)\log\hat{S}_{KM}(t))\}} \tag{2.22}$$

For tied events, the Kaplan-Meier estimator is defined as

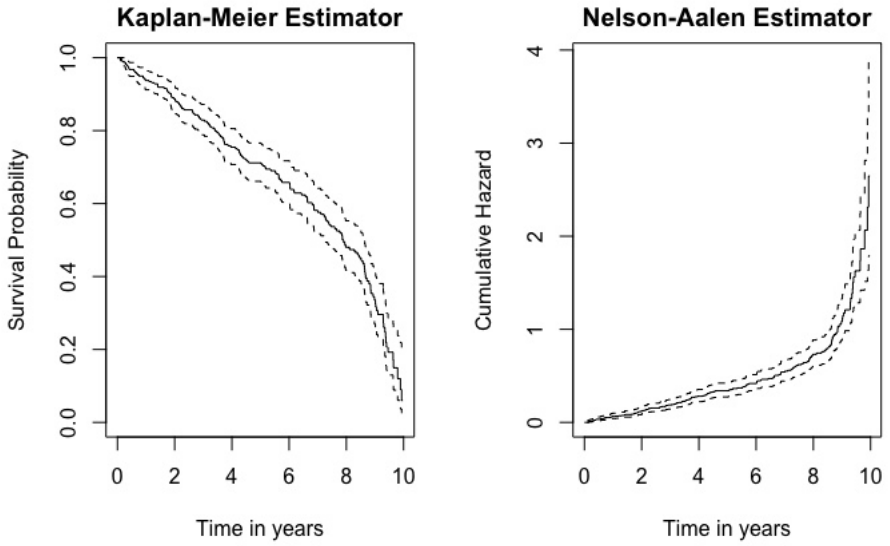
$$\hat{S}_{KM}(t) = \prod_{T_j \leq t} \left\{1 - \frac{o_j}{Q(T_j)}\right\} \tag{2.23}$$

with estimated variance, where we assume discrete event times,

$$\hat{\sigma}^2_{KM}(t) = \hat{S}_{KM}(t)^2 \sum_{T_j \leq t} \frac{o_j}{Q(T_j)\{Q(T_j) - 1\}}. \tag{2.24}$$

Figure 2 below illustrates the Kaplan-Meier and Nelson-Aalen estimators from a random generated dataset with event indicator and time of event. The estimations were calculated using the function **survfit()** from the package **survival** [69] in R. From the left graph illustrating the Kaplan-Mier estimator, we can observe that the y-axis is in the interval $[0, 1]$ as it represents the probability of survival. The probability is 1 at the beginning at time 0 and then goes quite steep to 0 as time increases. The right graph illustrates the Nelson-Aalen estimator for the cumulative hazard over time. The cumulative hazard is 0 at the initial time point and then increases with time which means that the risk of experiencing the event increases with time.

Figure 2: The left graphs illustrates the Kaplan-Meier estimator of the survival probability over time. The right graph illustrates the Nelson-Aalen estimator of the cumulative hazard over time.



2.5 Parametric models for survival analysis

The survival function can be modelled using different assumptions imposed on the hazard function. In this section we are going to define some of the models, namely, the exponential, Weibull and Gompertz distribution. The section follows Chapter 2.5 of Moeschberger & Klein [38] and the selection of distributions is based on Crowther (2014) [8].

2.5.1 Exponential distribution

The definitions of the survival and hazard functions of an exponential distribution are the following

$$S(t) = \exp(-\lambda t), \qquad h(t) = \lambda. \tag{2.25}$$

Then it follows from Eq. (2.10) that the density function is given by $f(t) = \lambda \exp(-\lambda t)$. From the definitions we can state that the exponential distribution assumes that the hazard rate is constant over time and also the mean residual life $E(T) = 1/\lambda$ is constant. Another property is that the exponential distribution is memoryless, that is

$$P(T \geq t + s | T \geq t) = P(T \geq s). \tag{2.26}$$

2.5.2 Weibull distribution

The survival and hazard functions of Weibull distribution are defined as,

$$S(t) = \exp(-\lambda t^{\gamma}), \qquad h(t) = \lambda\gamma t^{\gamma-1}, \tag{2.27}$$

and it follows that the density function is given by $f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma})$. The Weibull distribution is introduced as a more flexible distribution than the exponential distribution for survival data. In addition to the hazard rate being constant over time, it can be both monotone increasing or decreasing. Note that at a constant shape parameter, $\gamma = 1$, the Weibull distribution will equal the exponential distribution.

2.5.3 Gompertz distribution

The last distribution that we define is the Gompertz distribution, which is mainly used when describing mortality curves and when the hazard rate have an exponential change over time.

The survival and hazard function are the following

$$S(t) = \exp\left(\frac{\lambda}{\gamma}(1 - e^{\gamma t})\right), \qquad h(t) = \lambda\exp(\gamma t), \tag{2.28}$$

with the corresponding density function $f(t) = \lambda\exp(\gamma t)\exp\left(\frac{\lambda}{\gamma}(1 - e^{\gamma t})\right)$.

2.5.4 Maximum likelihood estimation

Estimating the parameters of the survival function $S(t)$ when the function is on a certain parametric form, is done using the maximum likelihood estimator. This approach combines the probability density function of the given distribution along with the survival function. Each survival dataset has a number n of subjects with a corresponding time to event T_i. Following the theory developed in Chapter 3 in the book *Joint Models for Longitudinal and Time-to-Event Data* by Rizopoulos (2012b) [54], we define the log-likelihood function for the i^{th} subject under a parametric survival model, where the parameter vector is denoted $\boldsymbol{\theta}$.

First, let T^* denote the true event time for subject $i = 1, .., N$ and denote the observed survival time by T_i such that $T_i = \min(T^*, C_i)$, that is, the minimum value of the event time T^* and the censoring time C_i. Introduce an event indicator

$$d_i = \begin{cases} 1, & \text{if } T^* \leq C_i \\ 0, & \text{otherwise} \end{cases}.$$

such that the indicator is equal to 1 if the i^{th} subject will experience the event before the end of the study and 0 if the subject will not experience the event at the end of study, that is, the subject is censored [54].

The log-likelihood function is then defined as the following,

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\{f(T_i;\boldsymbol{\theta})^{d_i} S_i(T_i;\boldsymbol{\theta})^{1-d_i}\} = \sum_{i=1}^{n} d_i \log[f(T_i;\boldsymbol{\theta})] + (1-d_i)\log[S_i(T_i;\boldsymbol{\theta})]. \tag{2.29}$$

The i^{th} subject who experienced the event at time T_i will contribute the probability density function $f(T_i;\boldsymbol{\theta})$ to the likelihood and the i^{th} subject who did not experience the event and are censored will contribute the survival function $S_i(T_i;\theta)$ as it is known that this subject has survived up to time $T_i^* > T_i = C_i$.

The log-likelihood can also be expressed in terms of the hazard and the survival function, as we have from Equation (2.10), that $f(t) = h(t)S(t)$.

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\{h(T_i;\boldsymbol{\theta})^{d_i} S_i(t_i;\boldsymbol{\theta})\} = \sum_{i=1}^{n} d_i \log[h(T_i;\boldsymbol{\theta})] + \log[S_i(T_i;\boldsymbol{\theta})]. \tag{2.30}$$

Alternatively, using Eq.(2.9) that $S(t) = \exp(-\int_0^t h(v)dv)$, the likelihood can be written in terms of the hazard function,

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} d_i \log[h(T_i;\boldsymbol{\theta})] - \int_0^{T_i} h(v;\boldsymbol{\theta})dv. \tag{2.31}$$

This means that the maximum likelihood estimate can be obtain using only the hazard function.

The maximum likelihood estimate has the following definitions [30]

Definition 2.1 *The maximum likelihood MLE of a parameter vector $\boldsymbol{\theta}$ given data $\boldsymbol{X}$ is obtained by maximising the log-likelihood*

$$\hat{\boldsymbol{\theta}}_{ML} = arg\ max \log L(\boldsymbol{\theta};\boldsymbol{X})$$

Definition 2.2 *The score vector is defined as the vector of first-order derivatives of the log-likelihood function*

$$S(\boldsymbol{\theta}) = \frac{d \log L(\boldsymbol{\theta};\boldsymbol{X})}{d\boldsymbol{\theta}}$$

The maximum likelihood estimation is computed by solving the score function, $S(\boldsymbol{\theta}) = 0$.

2.5.5 Newton-Raphson algorithm

Maximising the likelihood analytically is not always possible. This is why a numerical approach must be utilised in many practical applications. A numerical method used to maximise the likelihood function is the Newton-Raphson algorithm. The algorithm follows the following three steps [8]:

1. Start with an assumption of initial values $\boldsymbol{\theta_i}$

2. Set a new approximation of $\boldsymbol{\theta}$ as $\boldsymbol{\theta_{i+1}} = \boldsymbol{\theta_i} + \{-\boldsymbol{H}(\boldsymbol{\theta_i})\}^{-1}\boldsymbol{S}(\boldsymbol{\theta_i})$, where $\boldsymbol{S}(\boldsymbol{\theta_i})$ is a vector of score values and $\boldsymbol{H}(\boldsymbol{\theta_i})$ is the matrix containing the second-order partial derivatives.

3. If pre-specified convergence level is reached, $\boldsymbol{\theta_{i+1}}$ is the final approximation of $\boldsymbol{\theta}$, if not repeat step 2 until convergence.

2.6 Cox regression model

The Cox regression model or the Cox proportional hazards model, is used on survival data to, for instance, compare the hazard rate for two groups who received different treatments. The Cox model was developed by Cox in 1972 [7] and has since then been the mostly used regression model for survival data. The model is non-parametric which means that we do not have to define the distribution of T. This section follows the theory developed in Chapter 3 in Rizopoulos (2012b) [54].

For the i^{th} subject, the hazard rate is defined as the following

$$h_i(t|\boldsymbol{w_i}) = h_0(t)\exp(\boldsymbol{\gamma^T w_i}). \tag{2.32}$$

The first term $h_0(t)$ is known as the the baseline hazard which is the hazard rate when the other variables are equal to 0. The vector $\boldsymbol{w_i} = (w_{i1}, ..., w_{ip})$ is the number of p covariates that are related to the hazard of the i^{th} subject and the vector $\boldsymbol{\gamma}$ contains the regression coefficients.

To interpret the regression coefficients we can take the logarithm of (2.32) which will then be

$$\log h_i(t|\boldsymbol{w_i}) = \log h_0(t) + \gamma_1 w_{i1} + \gamma_2 w_{i2} + ... + \gamma_p w_{ip}. \tag{2.33}$$

From this expression we can declare that the regression coefficient, γ_j, for the predictor w_{ij}, is the log hazard change at time point t when there is a one unit increase of w_{ij} and the remaining predictors are held constant. The coefficient $\exp(\gamma_j)$ is equal to the ratio of hazards, when there is a unit change in w_{ij} at time t and the other predictors are constant.

If we observe the i^{th} and the k^{th} subjects, we can define the proportional hazards ratio as

$$\frac{h_i(t|\boldsymbol{w_i})}{h_k(t|\boldsymbol{w}_k)} = \frac{h_0(t)\exp(\boldsymbol{\gamma^T w}_i)}{h_0(t)\exp(\boldsymbol{\gamma^T w}_k)} = \exp\{\boldsymbol{\gamma^T(w_i - w_k)}\}. \tag{2.34}$$

This expression is the hazard ratio of the i^{th} subject with covariate $\boldsymbol{w}_i$ compared to the k^{th} subject with covariate $\boldsymbol{w}_k$.

2.6.1 Partial likelihood

In the paper by Cox [7], he suggested to estimate the parameters using a partial log-likelihood function. The function is a partial likelihood, as it will not include the observed survival or censoring times. Here we write the function for estimating the parameter γ and we begin by stating the likelihood function.

$$L_i(\gamma) = \left[\frac{\exp(\boldsymbol{\gamma}^{\boldsymbol{T}} \boldsymbol{w}_i)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\gamma}^{\boldsymbol{T}} \boldsymbol{w}_l)} \right]^{d_i}, \tag{2.35}$$

where we denote $R(t_i)$ as the set of subjects who are at risk at time t. The log-likelihood function is then defined as [54],

$$\log L_i(\gamma) = \sum_{i=1}^{N} d_i \left\{ \boldsymbol{\gamma}^{\boldsymbol{T}} \boldsymbol{w}_i - \log \sum_{l \in R(t_i)} \exp(\boldsymbol{\gamma}^{\boldsymbol{T}} \boldsymbol{w}_l) \right\} \tag{2.36}$$

2.7 Royston-Parmar survival model

The motivation for the development of the Royston-Parmar model arose when modelling censored survival data with non-proportional hazards, which can arise when the there is a progression of a disease with different rates [50], became arduous using the Cox proportional hazards model. The Royston-Parmar model [58] is a parametric model that can easily handle such data and also better visualise the hazard function. The models proposes to use restricted cubic splines to obtain a smooth and flexible model [58].

2.7.1 Restricted cubic splines

Cubic splines are used in statistical modelling to determine the relationship between the outcome variable and the explanatory variable or variables, when dealing with a nonlinear relationship. Splines can be described as piecewise smooth polynomials, which means that the explanatory variable is divided into a number of intervals with a polynomial function in between that are joined together by knots.

We follow Durrleman and Simon (1989) [15] to define restricted cubic splines. The cubic spline has a third polynomial degree, which are continuous at the knots. The advantage of using cubic splines over splines with a higher polynomial degree is that the cubic splines yield smoother results as the cubic spline fit fewer constants compared to higher degrees polynomials.

The cubic spline function is defined as,

$$s_3(x) = \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{i=1}^{K} \beta_{i3} (x - k_i)_+^3. \tag{2.37}$$

where K is the number of knots with $K+\tau+1$ regression coefficients where $\tau = 3$ is the degree of the spline and

$$(u)_+ = \begin{cases} u, & \text{if } u > 0 \\ 0, & \text{if } u \leq 0. \end{cases}$$

Restricted cubic splines indicate that the splines must be linear at the two end points, that is, it has to be linear before the first knot k_1 and after the last knot k_K. To obtain this linearity, $K-2$ new variables are created for the explanatory variable x. The place of the K knots are at $k_1 < k_2 < \cdots < k_{K-1} < k_K$ and the new variables for x are created by,

$$v_j(x) = (x-k_j)_+^3 - (x-k_{K-1})_+^3 \frac{k_K - k_j}{k_K - k_{K-1}} + (x-k_K)_+^3 \frac{k_{K-1} - k_j}{k_K - k_{K-1}}, \quad j = 1, ..., K-2 \tag{2.38}$$

In addition to these new variables, we assume that in Equation (2.37), $\beta_{02} = \beta_{03} = 0$ to obtain linearity for $x < k_1$ and $\sum_{i=0}^{K} \beta_{i3} = \sum_{i=1}^{K} \beta_{i3} k_i = 0$ for linearity of $x > k_K$. Further, the restricted cubic spline function is defined as,

$$s(x;\gamma) = \beta_{00} + \beta_{01} x + \sum_{i=1}^{K-2} \beta_{i3} v_{i+1}. \tag{2.39}$$

The position and the number of knots needs to be predetermined, but it is not evident how to proceed in order to determine them. Durrleman and Simon (1989) [15], suggests that the chosen number of knots should go to infinity as the sample size goes to infinity so that the use of splines results in a perfect fit of the function. However, 3-5 knots are sufficient to use to obtain a good enough flexibility for an adequate polynomial degree and size of data. For a larger data size, more knots can be used.

2.7.2 Defining the model

The article by Royston and Parmar (2002) [58] defines the flexible parametric models, known as the Royston-Parmar model. The derivation of this model was performed in Chapter 2 in Crowther (2014) [8], in which this section follows.

First we assume that the survival curve follows a Weibull distribution, that we defined in Section 2.5.2,

$$S(t) = \exp(-\lambda t^{\gamma}). \tag{2.40}$$

From Equation (2.8), we have that the cumulative hazard function is defined as $H(t) = -\log S(t)$ so then we can transform onto the log cumulative hazard scale by,

$$\log[H(t)] = \log[-[\log(S(t))] = \log(\lambda) + \gamma \log(t). \tag{2.41}$$

The log cumulative hazard is a linear function in the log time, $\log(t)$. The function is now extended to include a vector $\boldsymbol{w_i}$ of covariates such that,

$$\log[H(t|\boldsymbol{w_i})] = \log(\lambda) + \gamma \log(t) + \boldsymbol{w_i^T}\boldsymbol{\beta} \tag{2.42}$$

The next step is to relax the assumption of a linear function in the log time and introduce restricted cubic splines on $\log(t)$. This to obtain a more flexible model to detect the non-linear relationships.

The baseline function $\log(\lambda) + \gamma \log(t)$ is interchanged to the restricted cubic spline function in Equation (2.39) so the log cumulative hazard is then,

$$\log[H(t|\boldsymbol{w_i})] = \eta_i(t) = s(\log(t)|\boldsymbol{\gamma}, \boldsymbol{K}) + \boldsymbol{w_i^T}\boldsymbol{\beta}, \tag{2.43}$$

where $\boldsymbol{K}$ is a vector of K knots. By Equation (2.8) stating $H(t) = -\log S(t)$, the survival scale is written as,

$$S(t|\boldsymbol{w_i}) = \exp(-\exp(\eta_i(t))). \tag{2.44}$$

To define the hazard function, $h(t) = -\frac{d \log S(t)}{dt}$, we first have to compute the derivative of $\log S(t) = \log\{\exp(-\exp(\eta_i(t)))\} = -\exp(\eta_i(t))$ which is the following,

$$\frac{d \exp(\eta_i(t))}{dt} = \frac{d\eta_i(t)}{dt} \exp(\eta_i(t)) = \frac{d[s(\log(t)|\boldsymbol{\gamma}, \boldsymbol{K})]}{dt} \exp(\eta_i(t))$$

Applying the chain rule to $\frac{d[s(\log(t)|\boldsymbol{\gamma},\boldsymbol{K})]}{dt} = \frac{1}{t}\frac{d[s(\log(t)|\boldsymbol{\gamma},\boldsymbol{K})]}{d\log(t)}$, the hazard function is

$$h(t|\boldsymbol{w}_i) = \frac{1}{t}\frac{d[s(\log(t)|\boldsymbol{\gamma}, \boldsymbol{K})]}{d\log(t)} \exp(\eta_i(t)). \tag{2.45}$$

2.8 Time-varying covariates

It is common in survival analysis that the covariates vary over time. This could be for example biomarkers that are measured a number of times for each individual with a different result for each observation [8].

A biological marker or biomarker is a biological measure, such as blood pressure or cholesterol level, that is used on individuals to predict and determine the health states [43].

There are two types of time-varying covariates: exogenous and endogenous covariates. An exogenous covariate is determined outside the model but has an impact on the model. In a medical study an exogenous covariate could for example be the time of day, which can have an affect on the results of the study of a subject. An endogenous covariate is determined inside the model and could for example be a biomarker or any other time-dependent measurements from the subjects [8]. Due to these differences, the definitions

of the two covariates must be distinguished. This section follows the theory and notation from Chapter 3 in Rizopoulos (2012b) [54].

2.8.1 Exogenous covariates

An exogenous covariate needs the following condition to be fulfilled

$$Pr\{s \leq T_i^* < s + \Delta s | T_i^* \geq s, Y_i(s)\} = Pr\{s \leq T_i^* < s + \Delta s | T_i^* \geq s, Y_i(t)\} \tag{2.46}$$

As before, T_i^* denotes the time of an event and $Y_i(t) = \{y_i(s), 0 \leq s < t\}$ is the resulting covariate history up to time t. This expression holds for all s, t that satisfies $0 < s \leq t$ and also $\Delta s \to 0$.

Equation (2.46) can also be written in the following way, for $s \leq t$.

$$Pr\{Y_i(t)|Y_i(s), T_i^* \geq s\} = Pr\{Y_i(t)|Y_i(s), T_i^* = s\}. \tag{2.47}$$

Based on this definition it can be stated that the covariate vector $\boldsymbol{y_i}(\cdot)$ and the rate of events are connected and also, the events occurring at time s is independent of the events at time point $t > s$.

2.8.2 Endogenous covariates

The main reason for the need to distinguish exogenous and endogenous covariates is that the endogenous covariate is dependent on survival of the subject so the definition of failure in the model should not be death, that is, the subject must be alive. The reason for this is that if the failure was death, the trajectory of the covariate has information of the time of failure. Thus, the survival function with endogenous covariates must fulfill

$$S_i(t|Y_i(t)) = Pr(T_i^* > t|Y_i(t)) = 1. \tag{2.48}$$

This definition states that the survival at time t given the covariate history up to time t is equal to the certain probability that there is an event at time $T^* > t$ which means that the subject is alive at time T^*.

Equation (2.47) differs from (2.48) in a way that if the time of failure of a subject is at time point s, as in (2.47), then the endogenous covariate can not exist at time $t > s$.

2.8.3 Cox regression with exogenous covariates

Dealing with time-varying covariates requires an extension of the Cox regression model defined in Section 2.6, in particular using exogenous covariates. In this model, the presence of events is to be thought of as a realisation of the Poisson process $\{N_i(t), R_i(t)\}$, see Pickands III (1971) [47] for the definition and properties. $N_i(t)$ is the number of events for the i^{th} subject at time t and

$R_i(t)$ is a binary variable with $R(t) = 1$ if the subject is at risk at time t and $R(t) = 0$ if the subject is not at risk at time t. The model is extended such that it is conditional on the covariate history $Y_i(t)$ which results in an added term in the exponential function. The Cox regression model with exogenous time-varying covariates is then defined as the following.

$$h_i(t|Y_i(t), w_i) = h_o(t)R_i(t)\exp\{\boldsymbol{\gamma}^{\boldsymbol{T}}\boldsymbol{w}_i + \alpha\boldsymbol{y}_i(t)\}, \tag{2.49}$$

where $\boldsymbol{y}_i(t)$ is a vector of time-varying covariate. The interpretation of α will then be the following. We assume that there is one time-varying covariate so a one unit increase of $\boldsymbol{y}_i(t)$ at time t leads to that the term $\exp(\alpha)$ will be the relative increase of risk for an event. As $\boldsymbol{y}_i$ varies over time, the hazard ratio obtained from (2.49) will also vary over time.

3 Longitudinal data analysis

Longitudinal data are data retrieved from repeated measurements or samples of the same measure at different time points. In medical research it could be measurement of blood pressure or in economics, the unemployment level [74].

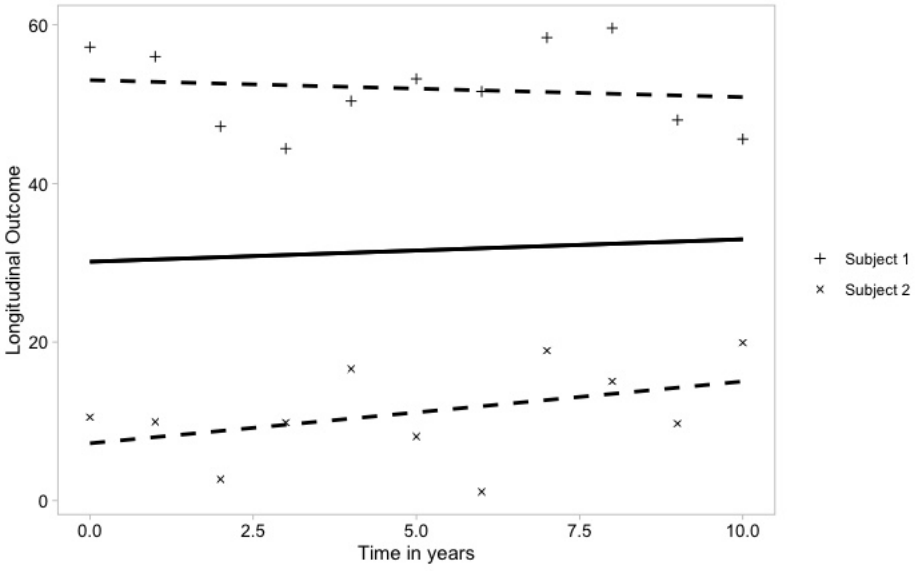
In medical research, a longitudinal study can be made where the subjects are given different treatments for the same type of disease. Then there are two main effects that can be observed. The first one is the cross-sectional effect which give information about the treatment at certain time points and if there is a difference in treatments. The second effect is called the longitudinal effect which tells us if the differences in treatments over a certain period of time [54].

Longitudinal data observations, for example blood pressure, from a subject, are dependent. That is, the data observations have a relationship as they are from the same subject and the same measurement. This means that it is not suitable to use statistical methods that require independent observations. Instead it is suggested to use the linear mixed effects model which we will describe in the next section, following Chapter 2 in Rizopoulos (2012b) [54].

3.1 Linear mixed effect models

Figure 3 illustrates hypothetical longitudinal outcomes for two subjects with the dashed line to represent the linear mean for each of them. The solid line in the middle of the graph represents the mean longitudinal outcome of the population.

Figure 3: Longitudinal outcomes for two subjects with corresponding subject specific linear mean over time. The middle solid line represents the average longitudinal evolution for the population.



This approach of analysing longitudinal data assumes that the subject specific longitudinal data can be analysed by a simple linear regression model, $Y_i = \beta_0 + X_i\beta_1 + \epsilon_i$. The linear mixed effects model is an extension of this model such that it is to be used on dependent observations and allows for fixed effects and random effects, thereby the name, mixed effect models. As the subjects have different slope and intercept parameters the model for response y_{ij} is the following,

$$y_{ij} = \tilde{\beta}_{i0} + \tilde{\beta}_{i1}t_{ij} + \epsilon_{ij}, \tag{3.1}$$

where y_{ij} is the observed response for the i^{th} subject, $i = 1, ..., N$ at time point t_{ij} where $j = 1, .., n_i$ and the error term is denoted $\epsilon_{ij} \sim N(0, \sigma^2)$. The subjects are a random sample from a population so the subject specific regression coefficients $\tilde{\beta}_{i0}$ and $\tilde{\beta}_{i1}$ are assumed to be random samples from the corresponding population of regression coefficients. The regression coefficients are bivariate normally distributed as $N(\boldsymbol{\beta}, \Sigma)$, where $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$. The coefficients $\tilde{\beta}_{i0}$ and $\tilde{\beta}_{i1}$ can then be written as $\tilde{\beta}_{i0} = \beta_0 + b_{i0}$ and $\tilde{\beta}_{i1} = \beta_1 + b_{i1}$ so the model for the response y_{ij} can be expressed as

$$y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t_{ij} + \epsilon_{ij}, \tag{3.2}$$

where $b_i = (b_{i0}, b_{i1})$ are the random effects with the bivariate normal distribution $N(0, \Sigma)$ and the fixed effects are the parameters β_0 and β_1.

If we allow for more random predictors and regression coefficients we obtain the linear mixed effect model.

Let p denote the number of covariates in the model, q number of random effects and n_i is the number of repeated measurements for the i^{th} subject, with a total of N subjects. Let y_{ij} denote the response at time $t_{ij}, j = 1, ..., n_i$ of subject $i, i = 1, ..., N$, then the linear mixed effect model is defined as

$$\begin{cases} \boldsymbol{y_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{\epsilon_i} \\ \boldsymbol{b_i} \sim N(0, \Sigma), \\ \boldsymbol{\epsilon}_i \sim N(0, \sigma^2 I_{n_i}). \end{cases}$$

The first term $\boldsymbol{X_i\beta}$ consists of the n_i x p design matrix $\boldsymbol{X_i}$, which is a matrix of observed random variables of fixed effects with p columns representing each covariate and n_i rows, one for each repeated measurement for subject i. The first column of the design matrix is a vector of 1's. The p x 1 vector $\boldsymbol{\beta}$ represents the fixed effects and is a vector of unknown constants. It is interpreted as that when all other variables are held constant and there is a one unit change in a covariate $x_j, j = 1...p$ then $\beta_j, j = 1...p$ is the change of the average value in y_i.

The second term $\boldsymbol{Z_i b_i}$ consists of the n_i x q design matrix $\boldsymbol{Z_i}$ of observed random variables of random effects with q columns representing each covariate and n_i rows, one for each repeated measurement for subject i.

The q x 1 vector $\boldsymbol{b_i}$ is a vector of the random effects regression coefficients that is normally distributed with mean parameter 0 and variance-covariance matrix Σ.

The error term $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2 I_{n_i})$, is a standard normal distribution where I_{n_i} is the identity matrix of order n_i.

3.1.1 Estimation of parameters

The main advantages of using linear mixed models are that it is possible to estimate both parameters that predicts the mean response change, y_i as described in previous section, and parameters that predicts the response trajectory change over time for each individual. Another advantage is that the number of observations need not be equal for each subject nor do they need to be collected at the same occasion. As stated, the repeated observations of a subject are dependent, we will capture this dependency by a random effect $\boldsymbol{b_i}$. This random effect will represent the marginal correlation of the outcomes for the i^{th} subject. Then we can express the longitudinal responses of the i^{th} subject as independent if we condition on the random effect as following

$$p(\boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}) = \prod_{j=1}^{n_i} p(y_{ij}|\boldsymbol{b_i}; \boldsymbol{\theta}), \tag{3.3}$$

where $\boldsymbol{\theta}$ is the parameter vector.

As for the estimation of parameters of the parametric survival models (Section 2.5.4), the parameters of the linear mixed effects model can also be estimated by the maximum likelihood. First we define the marginal density for the i^{th} subject of the observed response y_i as

$$p(\boldsymbol{y_i}) = \int p(\boldsymbol{y_i}|\boldsymbol{b_i})p(\boldsymbol{b_i})d\boldsymbol{b_i}. \tag{3.4}$$

Given that the conditional distribution $p(\boldsymbol{y_i}|\boldsymbol{b_i})$ and $p(\boldsymbol{b_i})$ are normally distributed, as $\boldsymbol{b}_i \sim N(0, \Sigma)$, and we have that $\boldsymbol{y}_i = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2 I_{n_i})$, Eq. (3.4) will result in a multivariate normal distribution with n_i dimensions. The mean value will then be $\boldsymbol{X_i\beta}$ and the covariance matrix $V_i = Z_i \Sigma Z_i^T + \sigma^2 I_{n_i}$. Then we can define the log-likelihood function, where we will introduce $\boldsymbol{\theta}$ as a vector containing the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta_b})$, where $\boldsymbol{\theta}_b = vech(\Sigma)$ is the vech operator applied to Σ, as

$$\log L_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{y}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int p(\boldsymbol{y}_i|\boldsymbol{b_i}; \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{b_i}; \theta_b) d\boldsymbol{b_i} \tag{3.5}$$

The likelihood function is given by the multivariate normal distribution as,

$$L_i(\boldsymbol{\theta}) = p(\boldsymbol{y_i}; \boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^{n_i}|V_i|}} \exp\left\{ -\frac{(\boldsymbol{y}_i - \boldsymbol{X_i\beta})^T V_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X_i\beta})}{2} \right\}. \tag{3.6}$$

The maximum likelihood estimate of $\boldsymbol{\beta}$, when we assume that the variance-covariance matrix V_i is known, is the following,

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i^T V_i^{-1} \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^T V_i^{-1} \boldsymbol{y_i}. \tag{3.7}$$

This expression is also known as the generalised least squares estimation. For the curious reader, the steps of calculating the maximum likelihood estimate for $\boldsymbol{\beta}$ is found in Appendix A.1. If an estimate of the covariance-variance matrix V_i is available but not the true V_i, Equation (3.7) can be utilised to calculate an estimate for $\boldsymbol{\beta}$, where we replace V_i by the estimate $\hat{V}_i$. The estimate $\hat{V}_i$ can be obtained by maximising the log-likelihood $\log L_i(\boldsymbol{\theta}_b, \sigma^2)$, when $\boldsymbol{\beta}$ is known. The maximum likelihood estimate of V_i will be unbiased, but for small samples, the maximum likelihood estimate of V_i will be biased [54].

3.2 Missing data

When collecting longitudinal data it happens that some subjects in the study do not provide all the measurements which leads to missing observations. In

this section, that follows Chapter 2 in Rizopoulos (2012b) [54], we are going to distinguish between two types of missing longitudinal data based on the pattern of the missing data.

Monotone missingness are patterns that are caused by a dropout, that is, a subject who leaves the study before the completion, or a subject who enters the study after the starting date and thereby has not provided the initial measurements, but stays until the end of study.

Non-monotone missingness, or intermittent missingness, is when a subject provides measurements at first follow-up but then misses to come at next follow up, then comes back to a follow-up and then might be missing more follow ups.

Handle missing longitudinal data can be difficult and may result in inaccurate estimates if the number of data observations are low or if the missing data are poorly handled which can introduce bias leading to inaccurate estimates. Missing data leads to unbalanced data over time as the subjects have provided different number of measurements, but according to Rizopoulos (2012b) [54], unbalanced data are not a concern for linear mixed effects models.

There are three main methods that can be used when handling missing longitudinal data. To define these methods we must first introduce some notation.

In a study, each subject is expected to leave $j = 1, ..., n_i$ measurements such that a vector of measurements is obtained from each subject, so the i^{th} subject has the vector $\boldsymbol{y}_i = (y_{i1}, ..., y_{in_i})$ of measurements. A missing data indicator is introduced to separate the observed measurements from the planned observations, denoted

$$r_{ij} = \begin{cases} 1, & \text{if } y_{ij}\text{is observed} \\ 0, & \text{otherwise} \end{cases}.$$

The response vector $\boldsymbol{y}_i$ will then be divided into two subvectors. The vector y_i^o contains the observed observations, that is when $r_{ij} = 1$, and the vector y_i^m contains the missing data.

When we have monotone missing data, that is missing data due to a dropout of the study, r_i will be $(1, ..., 1, 0, ..., 0)$ which we can write as

$$r_i^d = 1 + \sum_{j=1}^{n_i} r_{ij}. \tag{3.8}$$

That is, r_i^d is one plus the length of observations, in case of drop out, r_i^d will represent the instance of dropout and if there is no drop out it equals $r_i^d = n_i + 1$.

Now we can define the three methods for handling missing data following Chapter 2 in Rizopoulos (2012b) [54]. These are called *missing data mechanisms*, developed by Rubin (1976) [60] and Little and Rubin (2002) [41]

who proposed a conditional probability model that relates the missing data r_i given the response $y_i = (y_i^o, y_i^m)$ and the parameter vector θ_r,

$$p(r_i|y_i^o, y_i^m; \theta_r). \tag{3.9}$$

3.2.1 Missing completely at random (MCAR)

The first missing data mechanism is called missing completely at random (MCAR), that assumes that the probability of missing data are independent of the observations already collected and the measurement values that the missing observations would have produced. Using Equation (3.9), the MCAR longitudinal data are expressed as,

$$p(r_i|y_i^o, y_i^m; \theta_r) = p(r_i; \theta_r). \tag{3.10}$$

The observed data y_i^o are randomly sampled from the complete data, consequently, they share the same distribution. This means that excluding the missing data will still give valid results.

MCAR longitudinal data could for example be a health study where it is decided beforehand the time and number of measurements which implies that the probability to obtain a measurement is independent of the actual measurement [22].

3.2.2 Missing at random (MAR)

The second missing data mechanism is called missing at random (MAR), that assumes that the probability of missing data are dependent on the observed data y_i^o but independent of y_i^m. From Equation (3.9), we can express the MAR longitudinal as,

$$p(r_i|y_i^o, y_i^m; \theta_r) = p(r_i|y_i^o; \theta_r). \tag{3.11}$$

An example of MAR longitudinal data are a medical study where response values that are higher than a predetermined value are removed by the researcher that constructed the study. Hence, the missing data are dependent on the observed response y_i^o. This means that y_i^o is not a random sample from the complete data and therefore they do not have the same distribution. However, if the missing values y_i^m are conditioned on the observed value y_i^o, this will have the same distribution as that of the complete data so predicting missing values can be done by the observed data that is defined in the joint distribution of y_i^o and y_i^m.

3.2.3 Missing not at random (MNAR)

The third and last missing data mechanism is called missing not at random (MNAR). This mechanism assumes that the probability of missing data are

dependent on a subset of responses that would have been observed. Based on Equation (3.9), the MNAR longitudinal data are defined as,

$$p(r_i|y_i^o, y_i^m; \theta_r). \tag{3.12}$$

This expression tells us that the distribution of r_i is dependent on the elements in the subvector y_i^m and y_i^o. The observed data y_i^o are not a random sample from the complete data. Neither is the conditional distribution of y_i^m given y_i^o a sample of the complete data, as was the case in MAR, but y_i^m given y_i^o and $p(r_i|y_i)$ proves to be a random sample of the complete data. An example of MNAR longitudinal data in a medical research is when measuring the pain of patients and some subjects experience such a pain that they need medicine and hence must leave the study and the outcome of those subjects are not recorded.

4 Joint modelling of longitudinal and survival data

Up until now, we have assumed that, as is most common, longitudinal outcomes and survival data are analysed separately by for example using a hazards model for the survival data and a linear mixed effect model for longitudinal outcomes [8]. In Section 2.8.3 we defined the Cox regression model for exogenous covariates but when we have covariates that are endogenous time-varying longitudinal biomarkers that are associated with the survival, this model is no longer useful. This is when the joint model was introduced and the need to utilise joint modelling in clinical trials arises when longitudinal outcomes and survival data are associated to one another.

The advantages of using joint modelling in clinical trials are that the estimates of the survival data and longitudinal outcomes are more efficient when estimating the effects of treatments and additionally, the joint modelling reduce bias of the estimates [33].

4.1 Joint model

The notation and theory in this section is that of Chapter 4 Rizopoulos (2012b) [54]. In this section we define the joint model by stating the survival and longitudinal submodels. As in Section 2.5.4, we introduce an event indicator

$$d_i = \begin{cases} 1, & \text{if } T^* \leq C_i \\ 0, & \text{otherwise} \end{cases},$$

where T^* is the true event time for subject $i = 1, .., N$ and the observed survival time $T_i = \min(T^*, C_i)$. We continue to denote the longitudinal response as $y_{ij} = \{y_i(t_{ij}), j = 1, ..., n_i\}$, where $y_i(t_{ij})$ represents the j^{th} observation of longitudinal response, or endogenous time-varying covariate,

for the i^{th} subject at time point t_{ij}. There are n_i observations for each subject.

4.1.1 Survival submodel

The proportional hazard survival submodel is defined by

$$h_i(t|M_i(t), \boldsymbol{w_i}) = h_0(t) \exp\{\boldsymbol{\gamma^T w_i} + \alpha m_i(t)\}. \tag{4.1}$$

At first sight, the model appears to be similar to the extended Cox regression model from Section 2.8.3, but there are a few differences. In the survival submodel, we define $M_i(t) = \{m_i(s), 0 \leq s < t\}$ as the history of the true unobserved longitudinal process up to time t and $m_i(t)$ as the true unobserved value of the longitudinal outcome at time point t. Further details about $m_i(t)$ is defined in the next section for the longitudinal submodel. The parameter $h_0(t)$ is, the same as before, the baseline risk or hazard function and $\boldsymbol{w_i}$ is the vector of baseline covariates related to the hazard of the i^{th} subject. The regression coefficients are found in parameter $\boldsymbol{\gamma}$ and are interpreted such that, for a one unit change in w_{ij} at time t, $\exp(\gamma_j)$ is the ratio of hazards. The parameter α is defined as the effect that the longitudinal outcome has on the risk of the event and is interpreted in the following way; for a one unit increase in $m_i(t)$ at time t, $\exp(\alpha)$ is the relative increase in the risk for the event. The proportional hazard model (4.1) assumes a dependency on $m_i(t)$, that the risk of an event at time t is dependent on the present value of $m_i(t)$. We define the survival function, from equation (2.9), such that it is dependent on $M_i(t)$,

$$S_i(t|M_i(t), \boldsymbol{w_i}) = \exp\left(-\int_0^t h_0(u) \exp\{\boldsymbol{\gamma}^T \boldsymbol{w_i} + \alpha m_i(u)\} du\right). \tag{4.2}$$

In the Cox regression model in Section 2.6, we did not specify any distribution assumptions to model the baseline risk function $h_0(\cdot)$ for the survival times as that implies a risk of choosing the wrong distribution. In the joint modelling framework this baseline needs to be specified as otherwise we might obtain underestimations of the standard errors [54]. There are multiple ways to specify this baseline. One approach is to use a parametric distribution such as the Weibull distribution defined in Section 2.5.2, for the risk function. Other non-parametric approaches are the use of splines to estimate the risk function, proposed methods are to use linear splines, a B-spline or cubic splines. The approach that we will focus on in this thesis will be the piecewise-constant approach, which is defined by Rizopoulos (2012b) [54] in the following way,

$$h_0(t) = \sum_{q=1}^{Q} \xi_q I(v_{q-1} < t \leq v_q). \tag{4.3}$$

Here $0 = v_0 < v_1 < \cdots < v_Q$ is the time scale divided into smaller parts where the last part v_Q is larger than the longest observation time. The hazard value ξ_q is defined in the interval $(v_{q-1}, v_q]$. The more knots, the more flexible baseline hazards. When there are no ties and each interval $(v_{q-1}, v_q]$ only contains one true event time, which is true in the limiting case, the model will correspond to having $h_0(\cdot)$ unspecified and then estimate it with non-parametric maximum likelihood.

The last approach for specify the baseline risk function is the use of cubic splines. This model is based on the spline coefficients $\kappa = (\kappa_0, \kappa_1, ..., \kappa_m)$ and the so-called B-spline function $B(\cdot)$ in the following way

$$\log h_0(t) = \kappa_0 + \sum_{d=1}^{m} \kappa_d B_d(t, q). \tag{4.4}$$

The degree of the spline function is given by q and the number of interior knots is defined by m^*, such that $m = m^* + q - 1$. The number of knots should to be chosen depending on the number of events, usually one chose to include between $1/10$ and $1/20$ of the number of events. These knots are then placed based on the percentile of the observed or true event times.

4.1.2 Longitudinal submodel

In Section 3.1 we defined the linear mixed effect model that estimates the mean observed longitudinal response change denoted y_{ij}. The longitudinal data are observed intermittently, with error, at a number of time points t_{ij} for each subject. Our aim is to measure what the effect of longitudinal data has on the risk of an event. To do this, we estimate $m_i(t)$ which we defined in the previous section as the true unobserved value of the longitudinal outcome at time point t. Then we can define the longitudinal submodel which is based on the linear mixed effects model in Section 3.1, but we will now extend this model, as given in Rizopoulos (2012b) [54] where we assume that $y_i(t) = y_i$.

$$y_i(t) = m_i(t) + \epsilon_i(t), \qquad \epsilon_i(t) \sim N(0, \sigma^2) \tag{4.5}$$

$$m_i(t) = \boldsymbol{x_i}^T(t)\beta + \boldsymbol{z_i}(t)^T \boldsymbol{b_i}, \qquad \boldsymbol{b_i} \sim N(0, \Sigma) \tag{4.6}$$

As before, when we defined the linear mixed effects model, we have the time-varying measurements, $\boldsymbol{x_i}(t)$ which is the vector of the fixed effects $\boldsymbol{\beta}$ and $\boldsymbol{z_i}(t)$ that is the vector for the random effects $\boldsymbol{b_i}$. We also assume that the normal distributed error terms $\epsilon_i(t)$ are independent of the random effects and independent between each other, that is $(\epsilon_i(t), \epsilon_i(u)) = 0$, $t \neq u$.

The measurement errors are now accounted for as $y_i(t)$, the observed longitudinal outcome, is equal to $m_i(t)$, the true unobserved value of the longitudinal outcome plus an error term $\epsilon_i(t)$.

4.2 Alternative associations within the joint model

Crowther (2014) [8] formulated, in his thesis on joint longitudinal and survival models, a few alternative association structures of the joint model. In particular, he studies different ways of combining the survival and longitudinal submodels. This section explores the different ways of defining the joint model as given in Chapter 6 in Crowther (2014) [8].

4.2.1 Interaction effects

The joint model, as described in Section 4.1, assumes a current value parameterisation. This means that the model assumes an association between the true unobserved longitudinal profile $m_i(t)$ and the risk of the event at time point t, for all subjects in the data. Sometimes, this assumption does not hold and we need to consider different associations for subgroups of subjects. To allow this, we need to extend the model in Equation (4.1) by inserting interaction terms between the baseline covariates $\boldsymbol{w}_{in}$ and the true unobserved longitudinal profile $m_i(t)$. The model will then be the following,

$$h(t|M_i(t), \boldsymbol{w}_{i1}, \boldsymbol{w_{i2}}) = h_0(t) \exp\{\boldsymbol{\gamma}^T \boldsymbol{w_{i1}} + \boldsymbol{\alpha}^T \boldsymbol{w_{i2}} m_i(t)\}. \tag{4.7}$$

The difference between this expression and Eq. (4.1) is that we have added the vector w_{i2} containing the interaction covariates and which is linked to $m_i(t)$.

4.2.2 Time-dependent slope

The second association structure Crowther introduced is the time-dependent slope. This association can be used when we are interested in how the rate of change of a biomarker affects the risk of an event. In particular, this structure can be used when we want to know the effect of a biomarker over time, if it is a decreasing or increasing trend with a certain level of biomarker. The time-dependent slope, or equally, the rate of change is defined as,

$$h(t|M_i(t), \boldsymbol{w_i}) = h_0(t) \exp\{\boldsymbol{\gamma}^T \boldsymbol{w_i} + \alpha_1 m_i(t) + \alpha_2 m_i'(t)\}, \tag{4.8}$$

where

$$m_i'(t) = \frac{dm_i(t)}{dt} = \frac{d\{\boldsymbol{x_i}^T(t)\boldsymbol{\beta} + \boldsymbol{z_i}^T(t)\boldsymbol{b_i}\}}{dt} \tag{4.9}$$

4.2.3 Random effects parameterisation

The last alternative structure is the random effects parameterisation. This structure assumes a time-dependent association but only the random effects of the survival submodel are included. The model is defined below as,

$$h(t|M_i(t), \boldsymbol{w_i}) = h_0(t) \exp\{\boldsymbol{\gamma}^T \boldsymbol{w_i} + \boldsymbol{\alpha}^T(\boldsymbol{\beta} + \boldsymbol{b_i})\}. \tag{4.10}$$

In this model, $\boldsymbol{\beta}$ is the mean population average of the random effect and $\boldsymbol{b}$ is the deviation of the slope for each subject. We can also choose to only include $\boldsymbol{b}$ to obtain the following model,

$$h(t|M_i(t), \boldsymbol{w_i}) = h_0(t) \exp\{\boldsymbol{\gamma}^T \boldsymbol{w_i} + \boldsymbol{\alpha}^T \boldsymbol{b_i}\}. \tag{4.11}$$

The vector $\boldsymbol{\alpha}$ contains the association parameters between the random effect and the hazard, or risk, of event. This model is useful when dealing with a longitudinal submodel with a random intercept and a random slope, such that the random effects will determine the deviations for each subject based on the average intercept and slope. In particular, the model provides information about subjects that have a longitudinal level that is lower/higher at baseline or subjects that have an increase/decrease in the longitudinal trajectories have a higher risk to undergo the event [54].

Important to note is that the association parameters $\boldsymbol{\alpha}$ in models (4.10) and (4.11) have different interpretations. Let

$$m_i(t) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t. \tag{4.12}$$

be the longitudinal submodel with a random intercept and slope. Set $t = 0$ and insert $m_i(t)$ into model (4.10) to obtain

$$h(t|M_i(t), \boldsymbol{w_i}) = h_0(t) \exp\{\boldsymbol{\gamma}^T \boldsymbol{w_i} + \alpha_1(\beta_0 + b_{0i})\}. \tag{4.13}$$

The association parameter $\exp(\alpha_1)$ represents the hazard ratio when there is a one unit increase of the intercept. On the contrary, inserting $m_i(t)$ into model (4.11) gives

$$h(t|M_i(t), \boldsymbol{w_i}) = h_0(t) \exp\{\boldsymbol{\gamma}^T \boldsymbol{w_i} + \alpha_2 b_{0i}\}, \tag{4.14}$$

where the association parameter $\exp(\alpha_2)$ is dependent on the average subject specific deviation.

Alternatively, using the time-dependent slope in Equation (4.8), assuming $\alpha_1 = 0$ will result in the following,

$$\begin{aligned} h(t|M_i(t), \boldsymbol{w_i}) &= h_0(t) \exp\left\{\boldsymbol{\gamma}^T \boldsymbol{w_i} + \alpha_3 \frac{dm_i(t)}{dt}\right\} \\ &= h_0(t) \exp\{\boldsymbol{\gamma}^T \boldsymbol{w_i} + \alpha_3(\beta_1 + b_{1i}), \end{aligned} \tag{4.15}$$

Now, the association parameter $\exp(\alpha_3)$ is the hazard ratio when there is a one unit increase of the subject specific slope.

4.3 Estimating the joint model

Through the years, joint models have been estimated in various ways. The estimation of the model parameters were originally done by a two-stage method proposed by Self and Pawitan (1992) [64], where the first step was to estimate the random effects by least-squares. In the second step, the obtained estimates were used to assign values of $m_i(t)$ which were then used to calculate the partial likelihood of the Cox model as in Section 2.6.1. In this approach, Self and Pawitan (1992) [64] changed the term $\exp\{\alpha m_i(t)\}$ by $\exp\{1 + \alpha m_i(t)\}$ in the survival submodel (4.1). This in order to obtain linear random effects $\boldsymbol{b_i}$. However, Dafni and Tsiatis (1998) [10], Tsiatis and Davidian (2001) [71] and Sweetening and Thompson (2011) [67] have all proved by simulation that these approaches produce a high bias. Instead it has been suggested to use the maximum likelihood approach that will remove the bias.

4.4 Joint likelihood approach

In this section we define the maximum likelihood approach used for joint models following Chapter 4 in Rizopoulos (2012b) [54]. The joint distribution is defined as $\{T_i, d_i, y_i\}$ which represent the observed outcomes, survival time T_i, the event indicator d_i and the longitudinal response y_i. Define the vector $\boldsymbol{b_i}$ to contain the time-independent random effects that represents the longitudinal process and the survival process. The parameter vector $\boldsymbol{\theta} = \{\boldsymbol{\theta}_t, \boldsymbol{\theta}_y, \boldsymbol{\theta}_b\}$ contains the parameter $\boldsymbol{\theta}_t$ which is the event time outcome, $\boldsymbol{\theta}_y$ is the parameters for longitudinal outcomes and $\boldsymbol{\theta}_b$ is the parameter vector that are unique in the random effects covariance matrix.

In this model we will assume that the random effects describes the association between the longitudinal outcomes and the survival outcomes as well as the correlation between the repeated measurements of the longitudinal process.

The joint likelihood for the full joint model can then be defined as

$$p(T_i, d_i, \boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}) = p(T_i, d_i|\boldsymbol{b_i}; \boldsymbol{\theta})p(\boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}), \tag{4.16}$$

where

$$p(\boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}) = \prod_{j=1}^{n_i} p\{y_i(t_{ij})|\boldsymbol{b_i}; \boldsymbol{\theta}\}. \tag{4.17}$$

The n_i x 1 vector $\boldsymbol{y_i}$ contains the longitudinal responses for the i^{th} subject and j is the index of the longitudinal measurements. Furthermore, we assume that the censoring is independent of event time, given observed history, and also the process that decides the time points of the longitudinal

measurements is independent of the time of event. This means that censoring of a subject is dependent on the observed history up to time t.

Then the log-likelihood for the i^{th} subject is the following by Rizopoulos (2012b) [54],

$$\begin{aligned}\log p(T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta}) &= \log \int p(T_i, d_i, \boldsymbol{y_i}, \boldsymbol{b_i}; \boldsymbol{\theta}) d\boldsymbol{b_i} \\ &= \log \int p(T_i, d_i | \boldsymbol{b_i}; \boldsymbol{\theta_t}, \boldsymbol{\beta}) \left[\prod_{j=1}^{n_i} p\{y_i(t_{ij}) | \boldsymbol{b_i}; \boldsymbol{\theta_y}\} \right] p(\boldsymbol{b_i}; \boldsymbol{\theta_b}) d\boldsymbol{b_i}.\end{aligned} \tag{4.18}$$

The conditional survival density $p(T_i, d_i | \boldsymbol{b_i}; \boldsymbol{\theta_t}, \boldsymbol{\beta})$ is given by

$$\begin{aligned}p(T_i, d_i | \boldsymbol{b_i}; \boldsymbol{\theta_t}, \boldsymbol{\beta}) &= h_i(T_i | M_i(T_i), \boldsymbol{\theta_t}, \boldsymbol{\beta})^{d_i} \quad \text{x} \quad S_i(T_i | M_i(T_i); \boldsymbol{\theta_t}, \boldsymbol{\beta}) \\ &= [h_0(T_i) \exp\{\boldsymbol{\gamma}^T \boldsymbol{w_i} + \alpha m_i(T_i)\}]^{d_i} \\ &\text{x} \quad \exp\left(- \int_0^{T_i} h_0(u) \exp\{\boldsymbol{\gamma}^T \boldsymbol{w_i} + \alpha m_i(u)\} du \right),\end{aligned} \tag{4.19}$$

The conditional normality density of the longitudinal response $p\{y_i(t_{ij}) | \boldsymbol{b_i}; \boldsymbol{\theta_y}\}$ is given by

$$p\{y_i(t_{ij}) | \boldsymbol{b_i}; \boldsymbol{\theta_y}\} = \frac{1}{(2\pi\sigma^2)^{n_i/2}} \exp\left\{ - \frac{||\boldsymbol{y}_i - \boldsymbol{X_i}\boldsymbol{\beta} - \boldsymbol{Z_i}\boldsymbol{b_i}||^2}{2\sigma^2} \right\}. \tag{4.20}$$

And, finally, the normal distribution of the random effects $p(\boldsymbol{b_i}; \boldsymbol{\theta_b})$ is

$$p(\boldsymbol{b_i}; \boldsymbol{\theta_b}) = \frac{1}{(2\pi)^{q/2}} \frac{1}{\sqrt{|\Sigma|}} \exp\left\{ - \frac{\boldsymbol{b_i}^T \Sigma^{-1} \boldsymbol{b_i}}{2} \right\}, \tag{4.21}$$

where q is the dimension of the vector with random effects, $\boldsymbol{b_i}$ and $||x|| = \{\sum_i x_i^2\}^{1/2}$ is the Euclidean vector norm.

Different approaches have been suggested in how to compute the likelihood function for joint models. The most common method has been to estimate the maximum likelihood by using a method called the EM - algorithm (Expectation-Maximisation) Rizopoulos (2012b) [54], where the idea is that it is easier to maximise the log-likelihood that correspond to the complete data. The basic procedure of the algorithm begins with the E-step where missing data are added in order to obtain a complete data, then the log-likelihood function is replaced by a substitute function. The algorithm is then continued with the M-step where the substitute function is maximised. Even though this method has been preferred to use, there is a drawback, which is that the linear convergence rate of the algorithm give

rise to a slow convergence near the maximum. Another method that is used is the Newton-Raphson algorithm, defined in Section 2.5.5, to maximise the log-likelihood. In this thesis, we will focus on the approach by Rizopoulos (2012b) [54] which suggests to calculate the score vector that is to be solved by numerical integration. Following the calculations in Rizopoulos (2012b) [54], the score function of $\log p(T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta})$ is the following,

$$\begin{aligned}
S(\boldsymbol{\theta}) &= \sum_i \frac{\partial}{\partial \boldsymbol{\theta}^T} \log \int p(T_i, d_i|\boldsymbol{b_i}; \boldsymbol{\theta}) p(\boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}) p(\boldsymbol{b_i}; \boldsymbol{\theta}) d\boldsymbol{b_i} \\
&= \sum_i \frac{1}{p(T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}^T} \int p(T_i, d_i|\boldsymbol{b_i}; \boldsymbol{\theta}) p(\boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}) p(\boldsymbol{b_i}; \boldsymbol{\theta}) d\boldsymbol{b_i} \\
&= \sum_i \frac{1}{p(T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta})} \int \frac{\partial}{\partial \boldsymbol{\theta}^T} \{p(T_i, d_i|\boldsymbol{b_i}; \boldsymbol{\theta}) p(\boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}) p(\boldsymbol{b_i}; \boldsymbol{\theta})\} d\boldsymbol{b_i} \\
&= \sum_i \int \left[\frac{\partial}{\partial \boldsymbol{\theta}^T} \log \{p(T_i, d_i|\boldsymbol{b_i}; \boldsymbol{\theta}) p(\boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}) p(\boldsymbol{b_i}; \boldsymbol{\theta})\} \right] \\
&\quad \text{x} \quad \frac{p(T_i, d_i|\boldsymbol{b_i}; \boldsymbol{\theta}) p(\boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}) p(\boldsymbol{b_i}; \boldsymbol{\theta})}{p(T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta})} d\boldsymbol{b_i} \\
&= \sum_i \int A(\boldsymbol{\theta}, \boldsymbol{b_i}) p(\boldsymbol{b_i}|T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta}) d\boldsymbol{b_i}.
\end{aligned} \tag{4.22}$$

In the last equality we denote the complete data score vector, $A(\cdot)$, as

$$A(\boldsymbol{\theta}, \boldsymbol{b_i}) = \frac{\partial \{\log p(T_i, d_i|\boldsymbol{b_i}; \boldsymbol{\theta}) + \log p(\boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta}) + \log p(\boldsymbol{b_i}; \boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}^T}. \tag{4.23}$$

The integrals in Equation (4.2) and (4.22) are multidimensional and analytically intractable to compute which means that numerical integration techniques are needed in order to solve the integral. In the following sections we are going to use the numerical integration technique called Gauss-Hermite quadrature, in order to approximate the log-likelihood and score vector.

4.4.1 Numerical integration with Gauss-Hermite quadrature

The idea of this approach to solve the log-likelihood and score vector is to first of all fit the longitudinal outcomes using the mixed effects model. The result of this fit will yield information about the subject specific random effects given the longitudinal response, in particular, the location and scale of the conditional distribution of the random effects. Then, using this information, the subject specific integrands of the log-likelihood and score vector can be re-scaled. Following Rizopoulos (2012a) [53], this approach is motivated using the standard and adaptive Gauss-Hermite quadrature rules.

The integral in the score vector (4.22) can be approximated with the standard Gauss-Hermite rule by a weighted sum of integrand evaluations. The weighted sums are evaluated at the abscissas of the random effects. The abscissas is defined as the number that determines the position of the random effects [29]. A requirement in this approximation is that the abscissas needs to be prespecified. Using the standard Gauss-Hermite rule, the integral in the score vector (4.22) is approximated as follows [53],

$$E\{A(\boldsymbol{\theta}, \boldsymbol{b_i})|T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta}\} = \int A(\boldsymbol{\theta}, \boldsymbol{b_i}) p(\boldsymbol{b_i}|T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta}) d\boldsymbol{b_i}$$
$$\approx 2^{q/2} \sum_{t_1 \cdots t_q} \vartheta_t A(\boldsymbol{\theta}, \boldsymbol{b_t}\sqrt{2}) p(\boldsymbol{b_t}\sqrt{2}|T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta}) \exp(||\boldsymbol{b_t}||^2), \tag{4.24}$$

where $||\boldsymbol{b_t}|| = \{\sum_j b_{tj}^2\}^{1/2}$ is the Euclidean vector norm, the sum $\sum_{t_1 \cdots t_q}$ is another way of writing $\sum_{t_i=1}^{K} \cdots \sum_{t_q=1}^{K}$ where K is the number of quadrature points. The vector $\boldsymbol{b_t}^T = (b_{t_1}, ..., b_{t_q})$ are the abscissas of the random effects with corresponding weights ϑ_t. As the number of quadrature points, K, increases, the approximation will be improved. This is true if the integrand can be defined as $\exp(-\boldsymbol{b}^T\boldsymbol{b})l(\boldsymbol{b})$, where $l(\boldsymbol{b})$ is a polynomial degree of $2K - 1$ or less, then the location of the weights and abscissas will results in an exact solution to the integrand. However, there are some drawbacks of this approximation. First of all, the integrand is evaluated over the product of the abscissas for each random effect so as q increases, the computations will get heavy. Second, if the main mass of the integrand $g(\boldsymbol{b}) = A(\boldsymbol{\theta}, \boldsymbol{b})p(\boldsymbol{b}|T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta})$ has a different spread than the weight function $\exp(-\boldsymbol{b}^2)$ or far from zero, then the location of the quadrature points with respect to the main mass $g(\boldsymbol{b})$ can give a poor approximation as the abscissas will then not coincide with $g(\boldsymbol{b})$ when the Gauss-Hermite rule is applied. To improve on these drawbacks, the adaptive Gauss-Hermite has been developed. In each iteration, this approximation centers and scales the integrand following Rizopoulos (2012b) [54],

$$E\{A(\boldsymbol{\theta}, \boldsymbol{b_i})|T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta}\}$$
$$\approx 2^{q/2}|\hat{\boldsymbol{B}}_i|^{-1} \sum_{t_1 \cdots t_q} \vartheta_t A(\boldsymbol{\theta}, \hat{r}_t) p(\hat{r}_t|T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta}) \exp(||\boldsymbol{b_t}||^2). \tag{4.25}$$

The expression $\hat{r}_t = \hat{\boldsymbol{b}}_i + \sqrt{2}\hat{\boldsymbol{B}}_i^{-1}\boldsymbol{b_t}$, where $\hat{\boldsymbol{b}}_i = \text{argmax}_b\{\log p(T_i, d_i, \boldsymbol{y_i}, \boldsymbol{b}; \boldsymbol{\theta})\}$ and $\hat{\boldsymbol{B}}_i$ is the Choleski factor, decomposition of a matrix, of the matrix $\hat{H}_i$, where

$$\hat{H}_i = \frac{-\partial^2 \log p(T_i, d_i, \boldsymbol{y_i}, \boldsymbol{b}; \boldsymbol{\theta})}{\partial \boldsymbol{b} \partial \boldsymbol{b}^T}, \tag{4.26}$$

evaluated in $\boldsymbol{b} = \hat{\boldsymbol{b}}_i$.

The adaptive Gauss-Hermite rule gives the optimal approximation as now the integrand will have an approximate behaviour that of the density of the normal distribution $N(0, 2^{-1}I)$. The weight function will also be proportional to this normal density. However, the location of the mode $\hat{\boldsymbol{b}}_i$ and the second-order derivative of $\hat{H}_i$ are computationally heavy. To decrease this heavy computations, Rizopoulos (2012a) [53] suggests to investigate the conditional distribution of the random effects $p(\boldsymbol{b_i}|T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta})$, which is the second expression of the score vector (4.22), on the log-scale. In particular, determine the mode $\hat{\boldsymbol{b}}_i$ and the second-order derivative of $\hat{H}_i$. The density of the conditional distribution written on the log scale is proportional to the following [53],

$$\log p(\boldsymbol{b_i}|T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta})$$

$$\propto \sum_{j=1}^{n_i} \log p(\{y_i(t_{ij})|\boldsymbol{b_i}; \boldsymbol{\theta_y}\}) + \log p(\boldsymbol{b_i}; \boldsymbol{\theta_b}) + \log p(T_i, d_i|\boldsymbol{b_i}; \boldsymbol{\theta_t}, \boldsymbol{\beta}). \tag{4.27}$$

As n_i increases, the first term will equal the log density of the liner mixed model from Equation (4.17) such that $\log p(\boldsymbol{y_i}|\boldsymbol{b_i}; \boldsymbol{\theta_y}) = \sum_j \log p\{y_i(t_{ij})|\boldsymbol{b_i}; \boldsymbol{\theta_y}\}$ which will have a similar shape to that of the multivariate normal distribution. As $n_i \to \infty$ [53]

$$p(\boldsymbol{b_i}|T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta}) \xrightarrow{\text{P}} N(\tilde{\boldsymbol{b}}_i, \tilde{H}_i^{-1}), \tag{4.28}$$

where $\tilde{\boldsymbol{b}}_i = \text{argmax}_b\{\log p(\boldsymbol{y_i}|\boldsymbol{b}; \boldsymbol{\theta_y})\}$ and the matrix,

$$\tilde{H}_i = \frac{-\partial^2 \log p(\boldsymbol{y_i}|\boldsymbol{b}; \boldsymbol{\theta_y})}{\partial \boldsymbol{b} \partial \boldsymbol{b}^T}, \tag{4.29}$$

evaluated in $\boldsymbol{b} = \tilde{\boldsymbol{b}}_i$.

Thus, as n_i increase, only the information form the mixed effects model for the longitudinal outcome is enough to use to be able to re-center and re-scale the subject specific integrands. Therefore the adaptive Gauss-Hermite rule in Equation (4.25) where a standard transformation was used can be further developed by fitting the linear mixed effects model with

$\tilde{\boldsymbol{b}}_i = \text{argmax}_b\{\log p(\boldsymbol{y_i}, \boldsymbol{b}; \tilde{\boldsymbol{\theta}}_{\boldsymbol{y}})\}$ and the matrix,

$$\tilde{H}_i = \frac{-\partial^2 \log p(\boldsymbol{y_i}, \boldsymbol{b}; \tilde{\boldsymbol{\theta}}_{\boldsymbol{y}})}{\partial \boldsymbol{b} \partial \boldsymbol{b}^T}, \tag{4.30}$$

evaluated in $\boldsymbol{b} = \tilde{\boldsymbol{b}}_i$.

Then the following transformation can be used [54],

$$E\{A(\boldsymbol{\theta}, \boldsymbol{b_i})|T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta}\} \approx 2^{q/2}|\tilde{\boldsymbol{B}}_i|^{-1} \sum_{t_1 \cdots t_q} \vartheta_t A(\boldsymbol{\theta}, \tilde{r}_t) p(\tilde{r}_t|T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta}) \exp(||\boldsymbol{b}_t||^2). \tag{4.31}$$

The expression $\tilde{r}_t = \tilde{\boldsymbol{b}}_i + \sqrt{2}\tilde{\boldsymbol{B}}_i^{-1}\boldsymbol{b_t}$, and $\tilde{\boldsymbol{B}}_i$ is the Choleski factor of the matrix $\tilde{H}_i$, and the maximum likelihood estimates from the linear model are denoted $\tilde{\boldsymbol{y}}$. The difference between the adaptive Gauss-Hermite (4.25) and of this transformation (4.31) is that in the latter, the adaptive Gauss-Hermite rule is only implemented one time in the start and the quadrature points are not updated. This gives an advantage as the quadrature points are fewer than in the standard Gauss-Hermite (4.24) and the relocation of the quadrature points in the adaptive Gauss-Hermite (4.25) at each iteration is eliminated.

4.4.2 Estimating the random effects

The estimates of the random effects $\boldsymbol{b}_i$ are obtained using the **ranef()** function from the **JM** package [55]. Readers interested in the details of this estimation are referred to read Chapter 4.5 in Rizopoulos (2012b) [54].

4.5 Predicted survival probabilities

In this section we will define the procedure on the ability to predict the expected survival of a new entered subject or the survival probability of a subject at a certain time point during follow up. To be able to predict the survival probability, all information available for the subject such as biomarker values and baseline information has to be used [54]. This estimation is presented by Rizopoulos (2010) [52] which this section follows. The estimation is based on the joint model that is fitted on a random sample denoted $D_n = \{T_i, d_i, \boldsymbol{y_i}; i = 1, ..., n\}$ and the newly arrived subject has a set of longitudinal outcomes $Y_i(t) = \{y_i(s); 0 \leq s \leq t\}$ where $y_i(t)$ is an endogenous time-dependent covariate described in Section 2.8.2. We assume that $y_i(t)$ is associated with the survival, that is, if a subject has longitudinal observations up to time point t, it means that the subject has survived up to time t. The following survival function calculates the probability of surviving up to time $u > t$, given that the subject has survived at time t and the true event time is denoted by T^*. The conditional probability is then given by the following [52]

$$\pi_i(u|t) = P(T_i^* \geq u|T_i^* > t, Y_i(t), D_n; \boldsymbol{\theta}^*), \tag{4.32}$$

where D_n is the dataset that was used to fit the joint model and $\boldsymbol{\theta}^*$ is the true parameter values. When we obtain new information from a subject at time $t' < t$ the predictions can be updated so that we obtain $\pi_i(u|t')$ for $u > t$ and continue in this way as more information is obtained. This procedure follows a time dynamic manner which we will illustrate using real data in Section 6.

A first-order estimate of $\pi_i(u|t)$ is presented in Rizopoulos (2012b) [54] as

$$\tilde{\pi}_i(u|t) = \frac{S_i\{u|M_i(u, \hat{\boldsymbol{b}}_{\boldsymbol{i}}^{(t)}, \hat{\boldsymbol{\theta}}); \hat{\boldsymbol{\theta}}\}}{S_i\{t|M_i(t, \hat{\boldsymbol{b}}_{\boldsymbol{i}}^{(t)}, \hat{\boldsymbol{\theta}}); \hat{\boldsymbol{\theta}}\}} + O([n_i(t)]^{-1}). \tag{4.33}$$

The maximum likelihood estimate is denoted by $\hat{\boldsymbol{\theta}}$, which can be calculated using procedure in Section 2.5.4, and $\hat{\boldsymbol{b}}_{\boldsymbol{i}}^{(t)}$ is the node of the conditional distribution $\log p(\boldsymbol{b_i}|T_i^* > t, Y_i(t); \hat{\boldsymbol{\theta}})$ and finally $n_i(t)$ is the number of longitudinal responses obtained from subject i up until time t.

Even though this estimator works in practice, as indicated in Rizopoulos (2011) [51], it is difficult to calculate standard errors and confidence intervals for $\pi_i(u|t)$. Therefore, Rizopoulos (2011) [51] suggested using Monte Carlo sampling in order to take the variability of the maximum likelihood estimates into account. The procedure and details of the Monte Carlo sampling scheme is presented in Rizopoulos (2010) [52].

The simulated $\pi_i^{(l)}(u|t)$ is then used to derive the the median and mean of $\pi_i(u|t)$, for l = 1,...,L, as the following,

$$\hat{\pi}_i(u|t) = \text{median}\{\pi_i^{(l)}(u|t)\} \tag{4.34}$$

and

$$\hat{\pi}_i(u|t) = L^{-1} \sum_{l=1}^{L} \pi_i^{(l)}(u|t), \tag{4.35}$$

as well as deriving standard errors from the sample standard deviation from Monte Carlo samples and confidence intervals. Note that the estimates in (4.34) give more accurate results compared to (4.33) as the estimates in (4.34) approximate the integrals in a more accurate manner by the simulation [54].

4.6 Residuals

To examine the model assumptions, residual plots are created and analysed. Residual plots of longitudinal data and survival data separately have been widely studied, were residual examples of linear mixed models are presented by Nobre and Singer (2007) [45] and Verbeke and Molenberghs (2000) [73] and residuals based on survival outcomes are found in Therneau and Grambsch (2000) [70] and Harrell (2001) [27]. Problems may arise using these residuals to examine model assumptions based on joint models when the

longitudinal data have nonrandom dropout [52] as then the observed data does not correspond to a random sample of the population [56].

4.6.1 Residuals for longitudinal part

In this section we present two types of residuals that are commonly used in the linear-effects model following the definitions from Chapter 6 in Rizopoulos (2012b) [54]. The first type is the subject specific residual that confirm the assumption of the linear mixed effects model from Section 3.1,

$$\begin{cases} \boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{b_i} + \boldsymbol{\epsilon_i} \\ \boldsymbol{b_i} \sim N(0, \Sigma), \\ \boldsymbol{\epsilon}_i \sim N(0, \sigma^2). \end{cases}$$

The subject specific residuals are defined in the following way [54],

$$r_i^{ys}(t) = \{\boldsymbol{y}_i(t) - \boldsymbol{x}_i^T(t)\hat{\boldsymbol{\beta}} - \boldsymbol{z}_i^T(t)\hat{\boldsymbol{b}}_i\} \tag{4.36}$$

To obtain a standardised version of this residual, we divide the expression by the standard deviation $\hat{\sigma}$,

$$r_i^{yss}(t) = \{\boldsymbol{y}_i(t) - \boldsymbol{x}_i^T(t)\hat{\boldsymbol{\beta}} - \boldsymbol{z}_i^T(t)\hat{\boldsymbol{b}}_i\}/\hat{\sigma}. \tag{4.37}$$

The estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ are maximum likelihood estimates and the estimates for the random effects is denoted by $\hat{\boldsymbol{b}}_i$. The subject specific residuals predict the conditional error term $\epsilon_i(t)$ which is normally distributed, so these residuals verify the normal distribution assumption.

The marginal residuals is calculated by the marginal model of $\boldsymbol{y}_i$ which is defined as the following,

$$\begin{cases} \boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{\epsilon_i^*} \\ \boldsymbol{\epsilon}_i^* \sim N(0, \boldsymbol{Z}_i\Sigma\boldsymbol{Z}_i^T + \sigma^2 I_{n_i}). \end{cases}$$

We can then define the marginal residuals as [54]

$$r_i^{ym} = \boldsymbol{y}_i - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}. \tag{4.38}$$

For standard version of this residual we multiply by the term $\hat{\boldsymbol{V}}_i = \boldsymbol{Z}_i\hat{\Sigma}\boldsymbol{Z}_i + \sigma^2 I_n$ which is the estimated covariance matrix of $\boldsymbol{y}_i$.

$$r_i^{ysm} = \hat{\boldsymbol{V}}_i^{-1/2}(\boldsymbol{y}_i - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}). \tag{4.39}$$

The marginal residuals are utilised to verify that the covariance for each subject is on the form $\boldsymbol{V}_i$ and also to predict the marginal errors defined as $\boldsymbol{y}_i - \boldsymbol{X_i}\hat{\boldsymbol{\beta}} = \boldsymbol{Z_i}\boldsymbol{b_i} + \boldsymbol{\epsilon_i}$.

4.6.2 Residuals for the survival part

In this section we will explore a common residual for the survival part in the joint model, called a martingale residual. The martingale residual is defined for the i^{th} subject as the following [54],

$$r_i^{tm}(t) = N_i(t) - \int_0^t R_i(s)h_i(s|\hat{M}_i(s); \hat{\boldsymbol{\theta}})ds. \tag{4.40}$$

As already defined in Section 2.8.3, $N_i(t)$ is a process counting the number of events that the i^{th} subject has experienced at time t and $R_i(t)$ is a binary variable that is equal to 1 if the subject is at risk at time point t and 0 otherwise. Following the definition of the survival submodel in Equation (4.1), we can write the expression as the following,

$$r_i^{tm}(t) = N_i(t) - \int_0^t R_i(s)\hat{h}_0(s)\exp\{\hat{\boldsymbol{\gamma}}^{\boldsymbol{T}}\boldsymbol{w_i} + \hat{\alpha}\hat{m}_i(s)\}ds, \tag{4.41}$$

where $\hat{m}_i(t) = \boldsymbol{x_i}^T(t)\hat{\boldsymbol{\beta}} + \boldsymbol{z}_i^T(t)\hat{\boldsymbol{b}}_{\boldsymbol{i}}$ and $\hat{h}_0(s)$ is the estimated baseline risk function. This residual calculates for, the i^{th} subject, the difference between the number of observed events and the expected number of events estimated from the fitted model. This means that based on this residual we can identify if some subjects are poorly fitted by the model [54].

4.6.3 Residuals with nonrandom dropout

The computation of residuals for joint longitudinal and survival data with nonrandom dropout is proposed by Rizopoulos, Verbeke, and Molenberghs (2010) [56] which the theory in this section follows.

First of all, we present the joint model with nonrandom dropouts to prove that the observed data will not correspond to a random sample of the population. Denote the observed longitudinal measurements, in accordance with Section 3.2, for the i^{th} subject before the event time by $y_i^o = \{y_i(t_{ij}) : t_{ij} < T_i, j = 1, ..., n_i\}$ and denote the missing values of longitudinal measurements, that is, the measurements that would have been observed before the end of the study if the event had not occurred, by $y_i^m = \{y_i(t_{ij}) : t_{ij} < T_i, j = 1, ..., n_i'\}$.

Based on these definitions, the dropout mechanism can be derived for the joint model. The mechanism is a conditional distribution of the time until dropout, T^*, given the longitudinal outcomes (y_i^o, y_i^m) and the parameter vector $\boldsymbol{\theta}$, it is expressed as,

$$p(T_i^*|y_i^o, y_i^m; \boldsymbol{\theta}) = \int p(T_i^*|\boldsymbol{b_i}; \boldsymbol{\theta})p(\boldsymbol{b_i}|y_i^o, y_i^m; \boldsymbol{\theta})d\boldsymbol{b_i}. \tag{4.42}$$

Note that the time until dropout, T_i^* is dependent on y_i^m by the second term which represents the conditional distribution of the random effects,

$p(\boldsymbol{b_i}|y_i^o, y_i^m; \boldsymbol{\theta})$. Thus, the observed data which the residuals are calculated from, are not a random sample of the population, hence the model assumptions obtained from calculations of residuals, such as constant variance and a zero mean value, should not be expected to be fulfilled [52].

To obtain correct residuals for the joint models, Rizopoulos, Verbeke, and Molenberghs (2010) [56] suggests a different approach. They propose to increase the observed data by adding random longitudinal outcomes that corresponds to longitudinal outcomes that would have been been observed if the subject did not drop out of the study. That is, we assume that a joint model has been fit to the data and the maximum likelihood estimates are calculated so that we have an estimated parameter vector $\hat{\boldsymbol{\theta}}$ and a covariance matrix denoted $v\hat{a}r(\hat{\boldsymbol{\theta}})$. Denote the prespecified time points, of which the longitudinal measurements are assumed to be observed, as $t_0, t_1, ..., t_{max}$, where the last prespecified time is less than T_i for the i^{th} subject. This method assumes a conditional distribution of y_i^m from repeated samplings, given the observed data. Assuming the joint model in Equation (4.18) and the dropout mechanism (4.42), the density for the conditional distribution of y_i^m can be written as [56],

$$p(y_i^m|y_i^o, T_i, d_i) = \int p(y_i^m|y_i^o, T_i, d_i; \boldsymbol{\theta}) p(\boldsymbol{\theta}|y_i^o, T_i, d_i) d\boldsymbol{\theta}. \tag{4.43}$$

The conditional distribution in the first expression in Equation (4.43) can be rewritten using Equations (4.16), (4.17) and (4.42) as [56]

$$\begin{aligned} p(y_i^m|y_i^o, T_i, d_i; \boldsymbol{\theta}) &= \int p(y_i^m|\boldsymbol{b_i}, y_i^o, T_i, d_i; \boldsymbol{\theta}) p(\boldsymbol{b_i}|y_i^o, T_i, d_i, ; \boldsymbol{\theta}) d\boldsymbol{b_i} \\ &= \int p(y_i^m|\boldsymbol{b_i}; \boldsymbol{\theta}) p(\boldsymbol{b_i}|y_i^o, T_i, d_i, ; \boldsymbol{\theta}) d\boldsymbol{b_i}. \end{aligned} \tag{4.44}$$

The second expression in Equation (4.43) is the conditional distribution of the parameters $\boldsymbol{\theta}$ given the observed data. This distribution can be approximated, assuming n is large, by $N\{\hat{\boldsymbol{\theta}}, v\hat{a}r(\hat{\boldsymbol{\theta}})\}$, that is, $\{\boldsymbol{\theta}|y_i^o, T_i, d_i\} \sim N\{\hat{\boldsymbol{\theta}}, v\hat{a}r(\hat{\boldsymbol{\theta}})\}$. Assuming this approximation together with (4.43) and (4.44) gives a simulation scheme presented in Chapter 6.3 in Rizopoulos (2012b) [54] that simulates missing longitudinal responses $y_i^{m(l)}(t_{ij})$ that, together with the observed responses y_i^o, are used to calculate the residuals. The advantage of using the simulated missing responses is that the residuals will contain the same properties as in the complete data which means that correct assumptions about the model can be made even though there are some dropouts.

4.7 Joint models with flexible parameters

When the survival data and baseline hazard functions are more complex, a more flexible approach to formulate the survival submodel is needed in order to detect hazards with complex functions. Crowther et al., (2012a) [9] proposed to use the parametric Royston-Parmar model that uses restricted cubic splines described in Section 2.7, as a survival submodel from Equation (4.1), together with the longitudinal submodel in Section 4.1.2.

4.7.1 Defining the model

Following Chapter 7 in Crowther (2014) [8], we start by defining the Royston-Parmar survival submodel as,

$$\log H_i(t|M_i(t), \boldsymbol{w_i}) = \log H_0(t) + \boldsymbol{\gamma}^T \boldsymbol{w_i} + \alpha m_i(t), \tag{4.45}$$

where $\log H_i()$ is the log cumulative hazard scale, $M_i(t) = \{m_i(s), 0 \leq s < t\}$ is the history of the true unobserved longitudinal process up to time t and $m_i(t)$ is the true unobserved value of the longitudinal outcome at time point t. The parameter $H_0(t)$ is the cumulative baseline risk or hazard function, $\boldsymbol{w_i}$ is the vector of baseline covariates related to the hazard of the i^{th} subject and $\boldsymbol{\gamma}$ is the log cumulative hazard ratios.

The log baseline cumulative hazard H_0 is now written in terms of a restricted cubic spline function as in Equation (2.43), such that,

$$\log H_i(t|M_i(t), \boldsymbol{w_i}) = \eta_i(t) = s(\log(t)|\boldsymbol{\gamma}, \boldsymbol{K}) + \boldsymbol{\gamma}^T \boldsymbol{w_i} + \alpha m_i(t), \tag{4.46}$$

with the survival scale written as

$$S_i(t|M_i(t), \boldsymbol{w_i}) = \exp(-\exp(\eta_i(t))) \tag{4.47}$$

and the hazard function

$$\begin{aligned} h_i(t|M_i(t), \boldsymbol{w_i}) &= \frac{d\eta_i(t)}{dt} \exp(\eta_i(t)) \\ &= \frac{d[s(\log(t)|\boldsymbol{\gamma}, \boldsymbol{K}) + \boldsymbol{\gamma}^T \boldsymbol{w_i} + \alpha m_i(t)]}{dt} \exp(\eta_i(t)) \\ &= \left\{ \frac{1}{t} \frac{d[s(\log(t)|\boldsymbol{\gamma}, \boldsymbol{K})]}{d[log(t)]} + \alpha \frac{dm_i(t)}{dt} \right\} \exp(\eta_i(t)). \end{aligned} \tag{4.48}$$

It is important to note that the association parameter α and $m_i(t)$ occur both in the derivative and in $\eta_i(t)$ which means that for a one-unit increase in the longitudinal outcome $m_i(t)$ at time t, α is the log cumulative hazard ratio. This interpretation is not the same as for a one-unit increase in $m_i(t)$ at time t, α is the log hazard ratio. That is, the cumulative hazard ratio and the hazard ratio must be distinguished.

4.7.2 Joint likelihood function

The joint likelihood for the joint model with flexible parameters is similar to the joint model, where the log-likelihood for the i^{th} subject is presented in Crowther (2014) [8] as ,

$$\begin{aligned}\log p(T_i, d_i, \boldsymbol{y_i}; \boldsymbol{\theta}) &= \log \int p(T_i, d_i, \boldsymbol{y_i}, \boldsymbol{b_i}; \boldsymbol{\theta}) d\boldsymbol{b_i} \\ &= \log \int p(T_i, d_i | \boldsymbol{b_i}; \boldsymbol{\theta_t}) \left[\prod_{j=1}^{n_i} p\{y_i(t_{ij}) | \boldsymbol{b_i}; \boldsymbol{\theta_y}\} \right] p(\boldsymbol{b_i}; \boldsymbol{\theta_b}) d\boldsymbol{b_i}.\end{aligned} \tag{4.49}$$

The conditional normality density of the longitudinal response $p\{y_i(t_{ij})|\boldsymbol{b_i}; \boldsymbol{\theta_y}\}$ and the normal distribution of the random effects random effects $p(\boldsymbol{b_i}; \boldsymbol{\theta_b})$ are defined the same as in Section 4.4 in Equations (4.64) and (4.65), but the now the conditional survival density $p(T_i, d_i|\boldsymbol{b_i}; \boldsymbol{\theta_t})$ is given by [8]

$$\begin{aligned}p(T_i, d_i | \boldsymbol{b_i}; \boldsymbol{\theta}) = &\left[\left\{ \frac{1}{T_i} \frac{d[s(\log(T_i)|\boldsymbol{\gamma}, \boldsymbol{K})]}{d[log(T_i)]} + \alpha \frac{dm_i(T_i)}{dT_i} \right\} \exp(\eta_i(T_i)) \right]^{d_i} \\ \text{x} \quad &\exp[-\exp(\eta_i(T_i))].\end{aligned} \tag{4.50}$$

The advantage of this survival density compared to the survival density in Equation (4.19) is that it is easier to compute the cumulative hazard functions due the the absence of an integral which means dealing with nested numerical integration.

4.8 Joint models using finite mixture models

As an alternative to the joint models with flexible parameters using the Royston-Paramer model, that modelled the baseline hazard on the log cumulative hazard scale in Section 4.7, Crowther (2014) [8] proposes in Chapter 8, a parametric survival submodel where the parametric distributions have finite mixtures and the baseline hazard is modelled on the log hazard scale. In this joint model approach, Crowther assume a survival submodel based on a finite mixture of parametric distributions. The motivation for using this approach is that the use of finite mixture models increase the flexibility and improve the estimates.

4.8.1 Defining the model

The longitudinal submodel is defined as in Equations (4.5) and (4.6), that is,

$$y_i(t) = m_i(t) + \epsilon_i(t), \qquad \epsilon_i(t) \sim N(0, \sigma^2)$$

$$m_i(t) = \boldsymbol{x_i}^T(t)\beta + \boldsymbol{z_i}(t)^T\boldsymbol{b_i}, \qquad \boldsymbol{b_i} \sim N(0, \Sigma)$$

To obtain flexibility in this model, we can allow for the design matrices $\boldsymbol{x_i}$ and $\boldsymbol{z_i}$ to obtain restricted cubic spline function or polynomials.

To define the parametric submodel a two-component mixture proportional hazards model is derived.

The baseline survival function $S_0(t)$ is defined as,

$$S_0(t) = pS_{01}(t) + (1-p)S_{02}(t), \tag{4.51}$$

where $S_{01}(t)$ and $S_{02}(t)$ represents the two component survival function and $0 \leq p \leq 1$ is the mixture parameter. From Equation (2.7), the baseline hazard function is the following,

$$h_0(t) = -\frac{d \log S_0(t)}{dt} \tag{4.52}$$

with the proportional hazards function to be

$$h(t) = h_0(t) \exp(\boldsymbol{X_i\beta}), \tag{4.53}$$

where $\boldsymbol{X_i\beta}$ is a linear predictor with no intercept.

The survival functions $S_{01}(t)$ and $S_{02}(t)$ can take any distribution form. Following Crowther (2014) [8] we will use the Weibull distribution with the baseline hazards function, using the survival function defined in Equation (2.27), written as

$$S_0(t) = p \exp(-\lambda_1 t^{\gamma_1}) + (1-p) \exp(-\lambda_2 t^{\gamma^2}). \tag{4.54}$$

The scale parameters are λ_1, λ_2 and the shape parameters are γ_1, γ_2.

The cumulative hazard function is, Equation (2.8) defined as
$H(t) = -\log S(t)$, which gives

$$H_0(t) = -\log S_0(t) = -\log[p \exp(-\lambda_1 t^{\gamma_1}) + (1-p) \exp(-\lambda_2 t^{\gamma^2})]. \tag{4.55}$$

The baseline hazards function can then be derived by $h_0(t) = -\frac{d \log S_0(t)}{dt}$, using Equation (4.55)

$$h_0(t) = \frac{\lambda_1\gamma_1 t^{\gamma_1 - 1} p \exp(-\lambda_1 t^{\gamma_1}) + \lambda_2\gamma_2 t^{\gamma_2 - 1}(1-p) \exp(-\lambda_2 t^{\gamma_2})}{p \exp(-\lambda_1 t^{\gamma_1}) + (1-p) \exp(-\lambda_2 t^{\gamma^2})}, \tag{4.56}$$

where we use the derivative rule [46], $\frac{d}{dx} log(x) = \frac{1}{x}$.

Inserting this baseline hazard function of the Weibull distribution into the proportional hazards function, Equation (4.53) gives

$$h(t) = \frac{\lambda_1\gamma_1 t^{\gamma_1-1} p\exp(-\lambda_1 t^{\gamma_1}) + \lambda_2\gamma_2 t^{\gamma_2-1}(1-p)\exp(-\lambda_2 t^{\gamma_2})}{p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma^2})}\exp(\boldsymbol{X}_s(t)\boldsymbol{\beta}), \tag{4.57}$$

with the matrix of time-independent/dependent covariates $\boldsymbol{X}_s(t)$ and $\boldsymbol{\beta}$ as the log hazard ratios associated with the covariates.

The two-component mixture proportional hazards model can be used with the joint models in different ways. We are now going to state three examples, stated by Crowther (2014) [8], of where the two-component mixture model is used with the joint models.

Consider the survival submodel from Equation (4.1),

$$h_i(t|M_i(t), \boldsymbol{w_i}) = h_0(t)\exp\{\boldsymbol{\gamma}^T\boldsymbol{w_i} + \alpha m_i(t)\},$$

where $m_i(t)$ is the longitudinal response that determines the association between the survival model and longitudinal model at time t. Now interchange $h_0(t)$ with the two-component baseline hazards function, Equation (4.56), we obtain,

$$\begin{aligned} h(t|M_i(t), \boldsymbol{w_i}) = {} & \frac{\lambda_1\gamma_1 t^{\gamma_1-1} p\exp(-\lambda_1 t^{\gamma_1}) + \lambda_2\gamma_2 t^{\gamma_2-1}(1-p)\exp(-\lambda_2 t^{\gamma_2})}{p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma^2})} \\ \text{x} \quad & \exp\{\boldsymbol{\gamma}^T\boldsymbol{w_i} + \alpha_1 m_i(t)\}. \end{aligned} \tag{4.58}$$

The vector γ contains the log hazard ratios that corresponds to the baseline covariates in $\boldsymbol{w_i}$ and α_1 is a parameter that determines the relationship between the survival and longitudinal components.

The next example is if we want to study how the change of the longitudinal data, that is the biomarker trajectory, affects the survival. The survival submodel is then written as,

$$\begin{aligned} h(t|M_i(t), \boldsymbol{w_i}) = {} & \frac{\lambda_1\gamma_1 t^{\gamma_1-1} p\exp(-\lambda_1 t^{\gamma_1}) + \lambda_2\gamma_2 t^{\gamma_2-1}(1-p)\exp(-\lambda_2 t^{\gamma_2})}{p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma^2})} \\ \text{x} \quad & \exp\{\boldsymbol{\gamma}^T\boldsymbol{w_i} + \alpha_2 m_i'(t)\}, \end{aligned} \tag{4.59}$$

with the first derivative of the longitudinal submodel $m_i'(t) = \frac{d}{dt}m_i(t)$ and α_2 estimates the relationship between the change of biomarker trajectory and the survival.

In the last example we include random effects into the survival submodel, hence it can be written as,

$$h(t|M_i(t), \boldsymbol{w_i}) = \frac{\lambda_1\gamma_1 t^{\gamma_1-1} p\exp(-\lambda_1 t^{\gamma_1}) + \lambda_2\gamma_2 t^{\gamma_2-1}(1-p)\exp(-\lambda_2 t^{\gamma_2})}{p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma^2})} \\ \text{x} \quad \exp\{\boldsymbol{\gamma}^T \boldsymbol{w_i} + \alpha_3(\beta_0 + b_{0i})\}, \tag{4.60}$$

where β_0 denotes the intercept such that for a one unit increase in the biomarker baseline value for subject i at $t = 0$, α_3 denotes the log hazard ratio.

4.8.2 Joint likelihood function

We define the likelihood function for a continuous biomarker as [8],

$$\prod_{i=1}^{n}\left[\int_{-\infty}^{\infty}\left(\prod_{j=1}^{m_i} f\{y_i(t_{ij})|\boldsymbol{b_i};\boldsymbol{\theta}\}\right) f(\boldsymbol{b_i}|\boldsymbol{\theta}) f(T_i, d_i|\boldsymbol{b_i};\boldsymbol{\theta}) d\boldsymbol{b_i}\right], \tag{4.61}$$

where the conditional survival density $f(T_i, d_i|\boldsymbol{b_i}, \boldsymbol{\theta})$ is analogous to the definition in Section 4.4,

$$f(T_i, d_i|\boldsymbol{b_i}, \boldsymbol{\theta}) = h(T_i, d_i|\boldsymbol{b_i}, \boldsymbol{\theta})^{d_i} \quad \text{x} \quad S(T_i, d_i|\boldsymbol{b_i}, \boldsymbol{\theta}), \tag{4.62}$$

where $h(T_i, d_i|\boldsymbol{b_i}, \boldsymbol{\theta})^{d_i}$ is defined as in Equation (4.58) and

$$S(T_i, d_i|\boldsymbol{b_i}, \boldsymbol{\theta}) = \exp\left(-\int_0^{T_i} h(u, d_i|\boldsymbol{b_i}, \boldsymbol{\theta}) du\right). \tag{4.63}$$

The conditional normality density of the longitudinal response $f\{y_i(t_{ij})|\boldsymbol{b_i}, \boldsymbol{\theta}\}$ is given by

$$f\{y_i(t_{ij})|\boldsymbol{b_i}, \boldsymbol{\theta}\} = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left\{-\frac{[y_i(t_{ij}) - m_i(t_{ij})]^2}{2\sigma_e^2}\right\}. \tag{4.64}$$

And, finally, the normal distribution of the random effects $f(\boldsymbol{b_i}|\boldsymbol{\theta})$ is

$$f(\boldsymbol{b_i};\boldsymbol{\theta}) = \frac{1}{(2\pi|\Sigma|)^{q/2}} \exp\left\{-\frac{\boldsymbol{b_i}^T \Sigma^{-1} \boldsymbol{b_i}}{2}\right\}, \tag{4.65}$$

where q is the dimension of the vector with random effects, $\boldsymbol{b_i}$.

The full likelihood in Equation (4.61) can be estimated using the adaptive Gauss-Hermite from Section 4.4.1, according to Crowther (2014) [8].

5 Data analysis

In this section we are going to present the data used in this thesis as well as the procedure to implement the joint model on these data. In Section 5.1 the cohort where the data are taken from is described, following by Section 5.2 in which the variables in the data are defined. In Section 5.3 we describe the implementation details of the joint model using the R - package **JM**.

5.1 The AMORIS cohort

The data used in this thesis are taken from the AMORIS cohort at Karolinska Institutet in Stockholm, Sweden. AMORIS is an abbreviation of Apolipoprotein related mortality risk. The observations in the cohort were collected between 1985 and 1996, where the 812 073 subjects, 51 % women and 49 % men [18], provided blood and urine samples which were then analysed at the Central Automation Laboratory, (CALAB) in Stockholm [16], resulting in about 35 million values. This large amount of subjects and laboratory values allows the data to represent the population in Stockholm county in between this year span as the total population then was about 1.6 million people in Stockholm County [18]. Up until 2012, the follow-up has been updated with information about mortality, cancer and hospitalisation. Not only does the AMORIS cohort contain laboratory values, but for about 50% of the subject it also contains information about smoking, physical activity, blood pressure and BMI [17]. Half of the subjects included in the cohort were healthy, so the laboratory values were either collected from yearly health check-ups, 26% of the subjects, or occupational health care, 24%. The remaining 50 % were so called outpatients, that is patients who are in need of medical care but do not have to be hospitalised [18].

The AMORIS cohort differs from the general population of Stockholm County based on year 1990 in a way that the mean age of the first given observation in AMORIS was 42.6 years and the mean age in Stockholm was 38.4 years. Also the age distribution of the AMORIS cohort differs from the true age distribution in Stockholm. In AMORIS, 8 % of the subjects were 20 years or younger, 39 % were between 20 and 39 years, 38% between 40 and 59 years, 13 % between 60 and 79 years and finally 2 % of the subjects were 80 years or older. The corresponding percentages in Stockholm County were, 24 % who were 20 years or younger were, 31 % between 20 and 39 years old, 26% between 40 and 59 years, 16 % between 60 and 79 years and 3 % of the population were 80 years or older. [18]

5.2 Description of data

The AMORIS dataset that is used in this thesis consists of 266 037 observations and 29 variables which are the following;

- **Personal ID-number** Each subject is assigned an individual identification number, this to distinguish the subjects.
- **Sex** The sex of the subjects where 1 denotes a man and 2 a woman.
- **Date of birth** The date of birth of each subject, on the format "YYYY-MM-DD".
- **Sampling date** The date of examinations for each subject.
- **Sequence number** The order number of sampling. For example, Subject with ID-number 7 had four samples at different time points, so the sequence number of the first sampling date will be 1, the number for the second is denoted 2 and so on.
- **Age** The age of the subject at the sampling date.
- **BMI** Body mass index, but is missing for the majority of subjects.
- **First emigration after sampling** If the subject moves to another country they leave the study and the date of emigration is recorded.
- **Date of death** For the individuals who have died, the date of death is recorded. If a subject has emigrated, the date of death is a missing value as no further information of these individuals are recorded.
- **Underlying cause of death** For those subjects who have died, the cause of death are uniquely coded. In this dataset, there are 1625 different causes of death, for example, subject no. 604 had cause of death *J449* which means that this person died of the lung disease chronic obstructive pulmonary disease [44].
- **Last date for FoB/RTB** FoB (Folk- och bostadsräkningen) stands for the people and and housing census and RTB (Registret över totalbefolkningen) is the register of the total population [65]. At the end of each year, all Swedish residents are registered, that is, this variable stands for the last year that the individual was registered in Sweden. For example, subject no. 103 died in year 2001, so the last year this individual was registered was in "2000-12-31". This dataset contains data up until 2012, so those who have not died have a last date for FoB/RTB equal to "2011-12-31".
- **Fasting status** Before the examination, each subject were asked if they had eaten on the day or the night before the examination.
- **S-Apolipoprotein A (g/L)** Apolipoprotein A is a protein that is a part of the HDL cholesterol. This cholesterol is good for the body as it decreases the fat storage in the blood vessels by transporting the excess

cholesterol in the cells to the liver [75]. A high value of Apolipoprotein A indicates a low risk for heart diseases and a low value of under 1.25 g/L for women and 1.15 g/L for men is associated with a high risk of cardiovascular disease [61].

- **S-Apolipoprotein B (g/L)** Apolipoprotein B is a carrier protein that is included in the transport of the LDL-cholesterol [76]. LDL contains high values of cholesterol so when it is transported, it easily loses cholesterol along the way that gets stuck in the vessel walls, this causes cardiovascular disease [31]. An Apolipoprotein B value higher than 0.9 g/L indicates a higher risk of developing a cardiovascular disease [62].

- **S-Cholesterol (mmol/L)** S-cholesterol is a measure of the total cholesterol, that is both the "good" cholesterol, HDL, and the "bad" cholesterol, LDL. A total cholesterol value of under 5.2 mmol/L (1 g/L = 0.129116 mmol/L) is considered to be a good value for a healthy person [34].

- **fs-Triglycerider (mmol/L)** Triglycerides are, together with the cholesterol, the fat in the blood. Triglycerides are an important source of energy and are produced in the liver and by the intake of food such as diary products, meat and fats. A healthy person should have triglyceride levels under 1.7 mmol/L. High values increase the risk of cardiovascular diseases and factors that cause high levels could be a diet that mainly consist of sugars, meat and diary togehter with poor physical activity, or the reason for high values cold be genetic [28].

- **fs-Glucose (mmol/L)** When we eat bread, pasta, fruits and vegetables, the carbohydrates in these foods are converted to glucose. Glucose circulates in the blood system and is the source of energy to the cells, the brain and the nervous system. A hormone called insulin that is produced in the pancreas, transports the glucose to the cells. A person who has diabetes has a low production of insulin so the cells does not obtain the energy, hence the glucose levels in the blood are higher [77]. A non-diabetic healthy person should have a glucose level between 4.0 mmol/L - 5.9 mmol/L before meals and less than 7.9 mmol/L after meal. For a diabetic, the levels before a meal are between 4 and 7 mmol/L and between 5 and 9 mmol/L after a meal. A diabetic person with high levels of glucose over a long period of time have an increase risk of developing heart disease, kidney disease or stroke [12].

- **S-Haptoglobin (g/L)** The red blood cells transport oxygen from the lungs to the rest of the body. By time, the red blood cells are destroyed and release a protein called hemoglobin. The haptaglobin binds with the hemoglobin and transports it to the liver where it is removed from

the body. If the red blood cells are destructed faster than they are produced, the haptaglobin levels will decrease. Low levels indicate a developement of hemolytic anemia. This condition can result in dizziness, fatigue or increased heart rate [57]. The haptaglobin levels for a healthy adult should be between 0.24 g/L and 1.9 g/L [63].

- **S-CRP (mg/L)** Another protein found in the blod is the C-reactive protein (CRP) which is part of the immune system that is produced from the liver. The CRP indicates if the body is infected, when a disease or inflammation is developing, the CRP value increases. A highly sensitive CRP measures the risk of cardiovascular diseases. A CRP level that is higher than 3 mg/L indicates a high risk for developing cardiovascular disease. A level higher than 10 mg/L indicates that there is an ongoing infection in the body such as a cold, the individual is a smoker, has diabetes or is obese. Therefore, interpretation about cardiovascular diseases must be made with caution when the CRP value is higher than 10 mg/L [78].

- **LDL-C-Jr (mmol/L)**

  Low-density lipoprotein-cholesterol, LDL-C is known as the "bad" cholesterol as it leaves behind cholesterol in the blood vessels that eventually clogs the blood vessels which leads to cardiovascular diseases. For a healthy person, the LDL value should be below 3.4 mmol/L [34].

- **HDL-C-Jr (mmol/L)**

  High density lipoproteins-cholesterol, HDL-C is known as the "good" cholesterol. The cholesterol decreases the amount of cholesterol in the blood vessels by transporting the excess to the liver where it is removed from the body [75]. A low value of HDL means that the blood vessels have an excess of cholesterol, which causes them to clog and this leads to cardiovascular diseases. A healthy man should have a HDL vaule between 1.1-1.8 mmol/L and a healthy woman should have a level between 1.2-2.0 mmol/L [4].

- **S-Albumin (g/L)**

  Albumin is a protein, produced in the liver that transports vitamins, hormones, enzymes and medicines through the body. A low value of albumin is an indication of a liver or kidney disease because as the kidney begins to fail, the albumin will leave the body trough the urine. For a healthy person without any liver or kidney disease has an albumin level between 34 g/L and 54 g/L [2].

- **First date for AAA** The aorta is the blood vessel that goes from the heart down to the chest and stomach where it then divides in two and goes down the leg. The aorta is the main blood vessel that provides

blood to the whole body. The walls of the aorta can be weakened with age, which causes the walls to bulge out, called aneurysm, so that the aorta swells up. This condition is called abdominal aortic aneurysm, (AAA) [37]. The size of the AAA can vary, if it is between 3 cm and 5.4 cm, regularly screenings must be made to control the size. If the AAA is larger than 5.5 cm a surgery must be performed, as there is a high risk that the aorta will burst. If the aorta bursts it can lead to internal bleeding which is very severe. This variable states the date that the subject was diagnosed with AAA.

- **First diagnosis of AAA**

  The diagnosis for AAA is in this dataset coded differently depending on the ICD, indicates International Classification of Disease. The three editions of ICD that are relevant for this dataset are the eighth edition, ICD-8, which was used until the end of 1986, the ninth edition, ICD-9, which was used between 1987 and 1996, and lastly the 10th edition denoted ICD-10 which has been used since 1997. The code for AAA in ICD-9 is 441D and 441E and in ICD-10, it is coded I713, I714, I715 or I716. However, in this data there is only occurence of I714 in ICD-10. The code 44iD in ICD-9 can also stand for a rupture AAA. [39].

- **First date for CVD**

  The date where the subject was diagnosed with a cardiovascular disease, if not diagnosed, the date is a missing value.

- **First diagnosis for CVD**

  The type of CVD the subject was diagnosed with. For example the code I619 stands for a intracerebral haemorrhage [14] which is caused when a blood vessel in the brain bursts and blood is leaking into the brain, this can lead to severe damage to the brain [36].

- **First date for IHD**

  The date when some subjects were diagnosed with ischemic heart disease, IHD. An IHD is when the arteries that transport blood to the heart are narrowed io that the heart will not get enough blood and oxygen. If the arteries are completely clogged and the heart does not get any blood at all, the heart muscle will die and it will result in a myocardial infarction or heart attack [35].

- **First diagnosis for IHD** There are different types of IHD diagnosis that all have a unique code. For example, I21.9 denotes an acute myocardial infarction and I21.4B is an acute subendocardial infarction [13].

- **Socioeconomic index**

  The individuals in the data are given a socioeconomic index, (SEI). The index used was defined in the book published by SCB, the Central Bureau of Statistics in Sweden [66]. The index in the data are based on the occupation that was registered in FoB in 1990 so the index indicates the occupations of the individuals in the data in November 1990. The socioeconomic index is the following:

  - **11** Unskilled workers, producing goods
  - **12** Unskilled workers, producing services
  - **21** Skilled workers, producing goods
  - **22** Skilled workers producing services
  - **33** Lower official I (level of education; less than 2 years)
  - **36** Lower official I (level of education; 2 but not 3 years)
  - **46** Middle-level official
  - **56** Senior official
  - **57** Leading positions
  - **60** Freelancers with an academic education
  - **79** Entrepreneurs not including farmers
  - **89** Farmers

5.3 Joint models in R - The JM package

In this section, we will go through the procedure to fit the joint model for survival and longitudinal data on the AMORIS dataset using the statistical computing software *RStudio* [59]. To implement the joint model on the data, the R package **JM** developed by Rizopoulos (2018) [55] will be utilised.

The first step is to specify and fit the linear mixed effects model for the longitudinal data using the function **lme()** from the **nlme** package in R [48]. The main argument in this function is to specify the random effects and fixed effects [54].

The survival data are then fitted by a Cox model with the function **coxph()** from the **survival** package in R [69]. The main argument in this function is the formula where we specify the relationship between the observed event times and covariates [54].

The results from the linear mixed effects model and the Cox model are then applied to the **jointModel()** function from the **JM** package. The structure of the fitted joint model will be the same as if the linear mixed effects model and Cox model were fitted separately, but with the joint model function, the survival submodel will contain the estimated true longitudinal outcome $m_i(t)$ in the linear predictor, as described in Section 4.1.1.

The **JM** package includes various options for the model specification, that is, how the survival submodel should be fitted and which type of numerical integration method to be used. For this analysis, the method called *"piecewise-PH-aGH"* will be used which follows the theory of the adaptive Gauss-Hermite rule described in Section 4.4.1. Other methods can also be used to fit the joint model by, for example, applying a spline approximation or a Weibull assumption to the baseline risk function. Readers interested in the other possible methods are referred to *Joint Models for Longitudinal and Time-to-Event Data With Applications in R* [54] by Rizopoulos (2012b) or to his article *JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data* [52].

The motivation to use the adaptive Gauss-Hermite quadrature rule instead of the simple Gauss-Hermite quadrature rule is based from the result of the article *Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive Gaussian quadrature rule* by Rizopoulos (2012a) [53]. In this article Rizopoulos investigate the computational difference between the two procedures where he concluded that the adaptive Gauss-Hermite quadrature was 18 times faster to compute than the simple Gauss-Hermite and both of the methods produced the same amount of bias.

To predict the probability of survival for the subjects who are still alive at the end of the study we use the function **survfitJM()** from the **JM** package. This function takes the joint model as argument to predict the conditional survival probabilities $\pi_i(u|t)$ that we defined in Section 4.5. The function follows the Monte Carlo simulation scheme as presented in Rizopoulos (2010) [52] where we use the default of 200 Monte Carlo samples.

In addition to the predicted survival probabilities we can also analyse how the survival probability changes over time as more longitudinal measurements are collected from the subjects. In Section 4.5 we mentioned that the predicted survival probabilities follows a time dynamic procedure. As more observations from the subjects are obtained, the survival probabilities are updated and these are called dynamic predictions of the conditional survival probabilities. To illustrate this theory, we construct a for-loop in R which updates $\pi_i(u|t)$ for a subject after each additional longitudinal measurement of the longitudinal measurements, following Chapter 7 in Rizopoulos (2012b) [54].

In Section 4.6.1 we defined how to calculate the residuals for the longitudinal and survival part of the joint model. In this section we are going to illustrate these residuals of the joint model, following Chapter 6 in Rizopoulos (2012b) [54]. If we simply plot the joint model in R we obtain three residual plots representing the residual values versus the fitted values, a Q-Q plot of the subject specific residuals, the marginal survival and the marginal cumulative hazard which is calculated as $H(t) = -\log S(t)$. To plot the marginal survival, the following expression is used [54],

$$S(t) = \int S_i(t|\boldsymbol{b_i}; \hat{\boldsymbol{\theta}}) p(\boldsymbol{b_i}; \hat{\boldsymbol{\theta}}) d\boldsymbol{b_i} \approx \frac{\sum_i S_i(t|\hat{\boldsymbol{b_i}}; \hat{\boldsymbol{\theta}})}{n}. \tag{5.1}$$

The residuals from the longitudinal part, that is, the subject specific and marginal residuals defined in Section 4.6.1, are obtained using the functions **residuals()** and **fitted()** from base *R*. The standardised marginal residuals r_i^{ysm} and the marginal fitted values $\boldsymbol{X_i}\hat{\boldsymbol{\beta}}$ are plotted by specifying the type to *stand-Marginal* and *Marginal* respectively. Similarly, the standardised subject specific residuals r_i^{yss} and the subject specific fitted values are plotted by specifying the type to *stand-Subject* and *Subject* respectively.

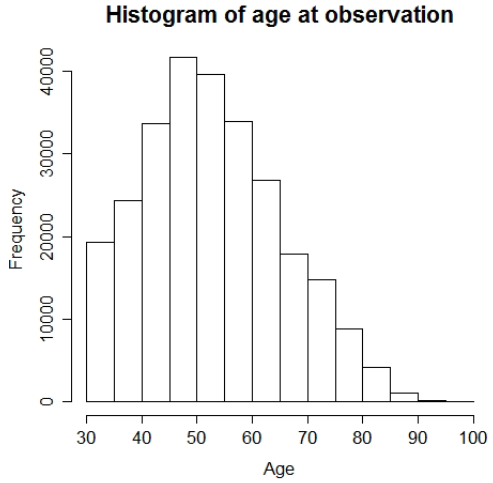
The martingale residual defined in Section 4.6.2 for the survival part of the joint model is also fitted using the function **residuals()** but we specify the type of residual to *martingale*.

6 Results of joint model on AMORIS data

The dataset that we are going to use to fit the joint model is a subset from the AMORIS cohort with 266 037 observations and 29 variables, which were defined in Section 5.2. To illustrate the joint model approach on the AMORIS data, we will perform a simple study where the variable **Age** will be the predictor variable.

The AMORIS data contains numerous variables with longitudinal measurements and predictor variables. As the dataset contains a large number of observations, we will choose to implement the joint model on a sample of these data. This to be able to perform a simple, but illustrative, example on how the theory of joint models is used on real data that is not computationally heavy. To create a sample of the data, we will first separate men and women and choose to work with the dataset with observations from the men population. Following the results in Figure 4 we observe that the age span of the male dataset is between 30 and 95 and that the subjects provided measurements mainly between the ages of 40 and 65. We then choose to include the measurements obtained between the age of 40 and 50 in our data analysis as this is an age span with many observations.

Figure 4: Distribution of the age of the subjects at observation.



The selection of longitudinal biomarkers that we are going to analyse are Apolipoprotein A, Apolipoprotein B, total cholesterol and triglycerides, so only the subjects who have measurements from all of these biomarkers will be included in the data-subset. Before the examination, each subject were asked if they had eaten on the day or the night before the examination and we choose to only include the subjects who had not eaten the night before as this might affect the biomarker measurements. This simplification of the data results in a dataset of 33 930 observations from 23 768 subjects. The number of subjects that experienced the event, that is the subjects who have died at the end of the study, is 2444 individuals, which corresponds to 89.7% censoring.

Before we get into the joint model, we will fit the extended Cox model from Section 2.8.3. We decide to include this model as well to compare the results of the α parameter that yields the association between the risk of death and a one unit increase of the biomarkers, between the Cox and the joint model. The extended Cox model takes the longitudinal biomarkers separately as an exogenous time-dependent covariate and we will control for the age variable. The survival model is the following,

$$h_i(t) = h_0(t) \exp\{\gamma \mathbf{Age} + \alpha y_i(t)\}, \tag{6.1}$$

where $y_i(t)$ is the observed value of the biomarker at time t.

In the longitudinal part of the joint model, we will fit linear mixed effects models on each of the longitudinal biomarkers, Apolipoprotein A, Apolipoprotein B, total cholesterol and triglycerides. The predictor variable in this model will be the age of each subject at observation time. Using the definition of a longitudinal submodel in Equations (4.5) and (4.6), the linear mixed model is defined as

$$
\begin{aligned}
y_i(t) &= m_i(t) + \epsilon_i(t) \\
&= \beta_0 + \beta_1 \mathbf{Age} + b_{i0} + b_{i1}\mathbf{Age} + \epsilon_i(t)
\end{aligned}
\tag{6.2}
$$

In the fixed effect part, the main effect of age is included and in the random effect design matrix we have an intercept and age term.

For the survival part of the joint model we will analyse if the subjects are alive or dead at the end of the study. The subjects entered the study when they were between 40 and 50 years old which they were somewhere in the years 1985 and 1996. The end of the study is the 31^{st} December 2011 as this is the last date that we have information if they are alive or not. The survival submodel is defined as,

$$
h_i(t) = h_0(t) \exp\{\gamma \mathbf{Age} + \alpha m_i(t)\} \tag{6.3}
$$

As we use a piecewise adaptive Gauss-Hermite quadrature rule to model the data, the baseline risk function, $h_0(t)$ is assumed to be piecewise constant. The number of knots that we will use are six which are placed at equal intervals of the observed event times [54]. The association parameter α determines the association between the true value of the longitudinal biomarkers and the risk for an event at time t [53].

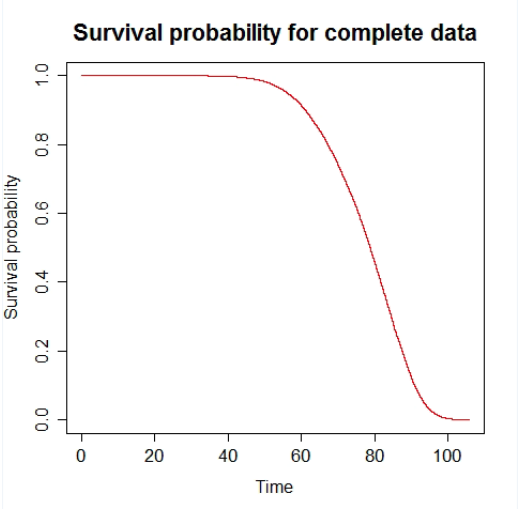
To fit the data to the **jointModel()** function in the **JM** package, we first have to specify and fit the linear mixed effects model for the longitudinal data and the Cox models for the survival data and then apply the results to the joint model function.

In the remaining of this section, we will present the results obtained from analysing the AMORIS data of men who have provided longitudinal measurements of Apolipoprotein A, Apolipoprotein B, total cholesterol and triglycerides between the age 40 and 50. The results will include analyses of the extended Cox model, the joint model with age as predictor variable, predicted and dynamic survival probabilities from a selection of subjects and finally some diagnostic plots in the form of residuals. But first, we provide some illustrations of the dataset.

6.1 Illustration of data

To illustrate the dataset that we are going to analyse, we plot the survival probabilities. In Figure 5 the survival probabilities for the whole dataset, both men and women, are illustrated and we can observe that the probability of survival when the subjects are aged 0 to about 40 years is 1. Then there is a small decrease up until age 60 where there the probability of survival steeply decreases until the age of 100 when the probability is equal to 0.

Figure 5: Survival probability for whole data.



The survival probabilities for the dataset that we are going to use are illustrated in Figure 6. At first, we observe a similar pattern as in the previous figure, that the survival probability is equal to 1 until about the age of 40. Then the survival probability decreases to about 0.7 at the age of 80.

Figure 6: Survival probability for men aged 40-50.

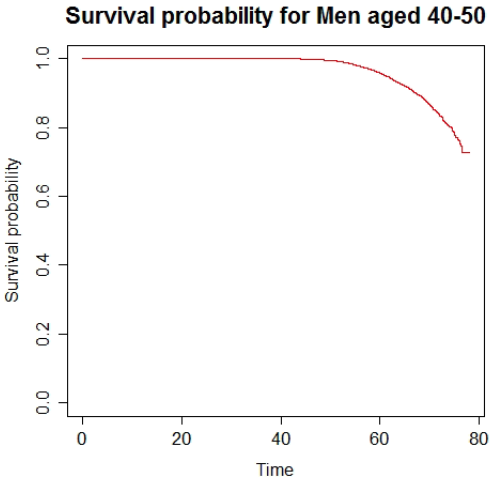


Table 1 presents some statistics of the age at death of the 2444 individuals who died by the end of the study. The mean age at death was 61 years old, the median 61.4 and the subject who lived the longest was 76.6 years old, and the one who died the youngest was 41.8 years old.

Table 1: Statistics of age at death of the subjects who died by the end of the study.

	Mean	Median	Max	Min
Age at death	61	61.4	76.6	41.8

In Table 2 we have some statistics of the age of the subjects who were alive at the end of the study. The mean and median age was 66 years, the maximum age was 77.9 years and the minimum age was 41 years.

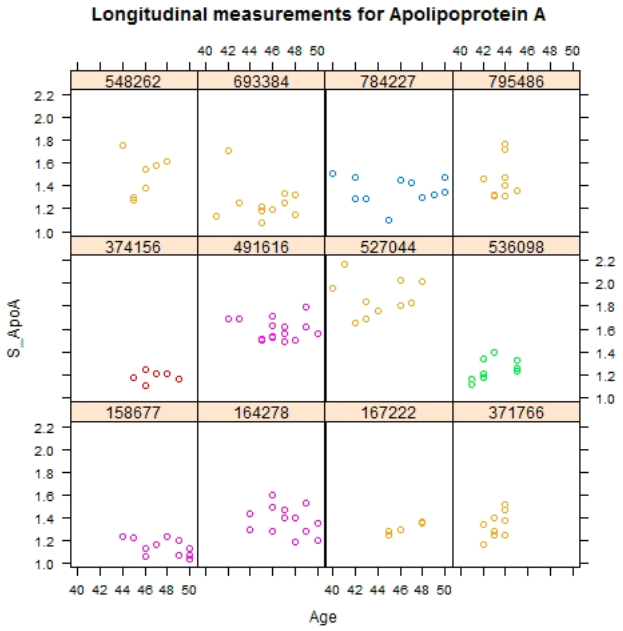
Table 2: Statistics of age of the subjects who were alive at the end of the study.

	Mean	Median	Max	Min
Age at death	66	66	77.9	41

6.2 Apolipoprotein A

Figure 7 presents the longitudinal measurements of Apolipoprotein A for 12 subjects who all, but the last one in the right corner, were censored, that is they were alive at the end of the study. These subjects were selected because they were the top 12 subjects with the most number of follow-ups, this enables us to easier observe a pattern of the measurements over time. From the description of data in Section 5.2, we know that a Apolipoprotein A value of over 1.15 g/L for men indicates a low risk for heart disease and a value under this threshold is associated with a high risk of developing cardiovascular disease [61]. From the figure we can observe that almost all subjects have measurements that are higher than 1.15 g/L and according to the data, only subject number 693 384 had a heart disease. None of the subjects in the middle column nor the subject in the right bottom corner developed a cardiovascular disease, and studying the graphs in the figure, we can observe that none of these subjects had values lower than 1.15 g/L.

Figure 7: Longitudinal measurements of Apolipoprotein A for some subjects.



Before we fit and present the results from the joint model, we fit the extended Cox model from Equation (6.1) with the resulting parameter estimates presented in Table 3.

Table 3: Parameter estimates from extended Cox model fit on Apolipoprotein A.

	coef	exp(coef)	se(coef)	z	p
Age	-0.03	0.97	0.01	-5.60	0.00
S-ApoA	-0.12	0.89	0.08	-1.42	0.16

The estimate for Apolipoprotein A in Table 3 indicates that there is a $\exp(-\alpha) = 1.13$ fold increase , with a 95% confidence interval of (0.95, 1.32), in the risk of death for a one unit decrease in Apolipoprotein A.

The joint model for longitudinal and survival data as defined in Equations (6.2) and (6.3) was fitted to the data with Apolipoprotein A as the longitudinal observation. The result of this fit is presented in Table 4.

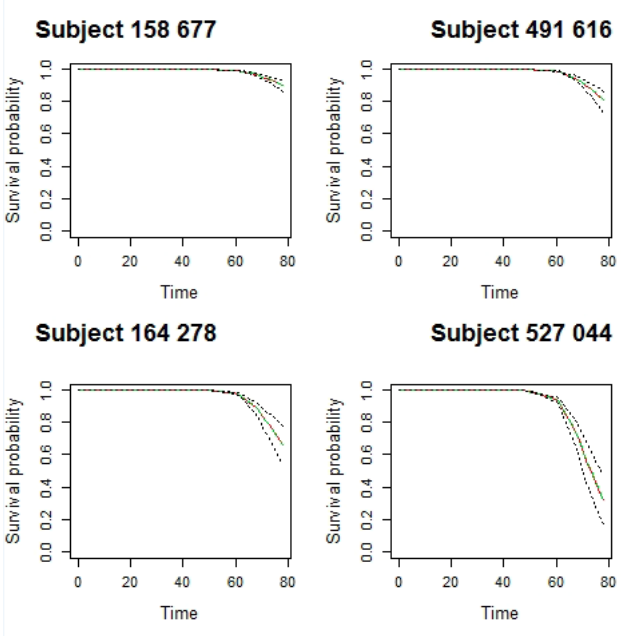
Table 4: Parameter estimates for Apolipoprotein A, standard errors and p-values under the joint modeling analysis. Σ_{ij} denote the ij-element of the covariance matrix for the random effects.

	Event Process				Longitudinal Process		
	Value	Std.Err	p-value		Value	Std.Err	p-value
Age	-0.0357	0.0070	< 0.0001	(Intercept)	0.9606	0.0157	< 0.0001
Assoct	2.9436	0.1333	< 0.0001	Age	0.0088	0.0003	< 0.0001
$\log(\xi_1)$	-9.4061	0.3657		$\log(\sigma)$	-1.9631	0.0072	
$\log(\xi_2)$	-7.7585	0.3950					
$\log(\xi_3)$	-7.7306	0.4025		Σ_{11}	0.0297	0.0005	
$\log(\xi_4)$	-7.4600	0.4083		Σ_{12}	-0.0002	0.0000	
$\log(\xi_5)$	-7.5776	0.4167		Σ_{22}	0.0000	0.0000	
$\log(\xi_6)$	-7.1440	0.4222					
$\log(\xi_7)$	-7.0706	0.4343					

From the parameter estimates in Table 4 we observe that a one unit increase of Apolipoprotein A will result in a $\exp(2.94) = 18.98$ fold increase, 95% confidence interval (14.62, 24.65), risk of death for a subject. The estimates ξ_i, $i = 1, ..., 7$ are the parameters from the piecewise-constant baseline risk function in Equation (4.3). The standard error of the covariance matrix Σ_{12} in the longitudinal process is 0.0000042 and the value of Σ_{22} is 0.0000064 with standard error 0.0000001. The estimated parameters in the longitudinal process are not values that can be interpreted, so we continue to predict the survival probabilities.

Using the function **survfitJM()** from the **JM** package we can predict the survival probabilities for subjects who have not yet died at the end of the study. From Table 1 we know that the maximum age of the subjects who died at the end of the study was 76.6 years old. This means that for the prediction of survival probability, we will estimate the probability that the subjects will be older than 76.6 years old. We select to study the conditional survival probability for four subjects in the data. The resulting survival probabilities are illustrated in Figure 8

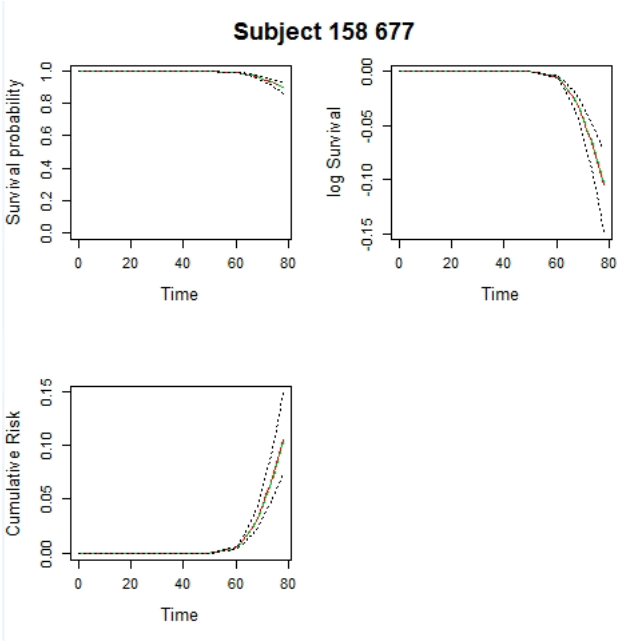
Figure 8: Predicted survival probabilities from joint model fit on Apolipoprotein A.



The red dashed line in Figure 8 represents the median estimator from Equation (4.34) and the solid green line is the mean estimator from Equation (4.35). The black dotted lines is a 95% confidence interval. Studying the four predicted survival probabilities of four different subjects we can observe that subject 158 677 has the highest survival probability of about 0.9 at the age of 80. As we move along to the other three subjects, the predicted survival probability decreases gradually to a survival probability of about 0.4 for subject 527 044. If we look back at the longitudinal outcomes for Apolipoprotein A in Figure 7 for these four subjects, we can observe that subject 158 677, which had the highest survival probability, had very low Apolipoprotein A and subject 527 044 with the lowest survival probability had the highest Apolipoprotein A values out of these four subjects.

Based on the predicted survival probabilities we can transform these estimates to obtain subject specific log survival and cumulative risk. In Figure 9, the predicted survival probability, the log survival and the cumulative risk are illustrated for subject 158 677.

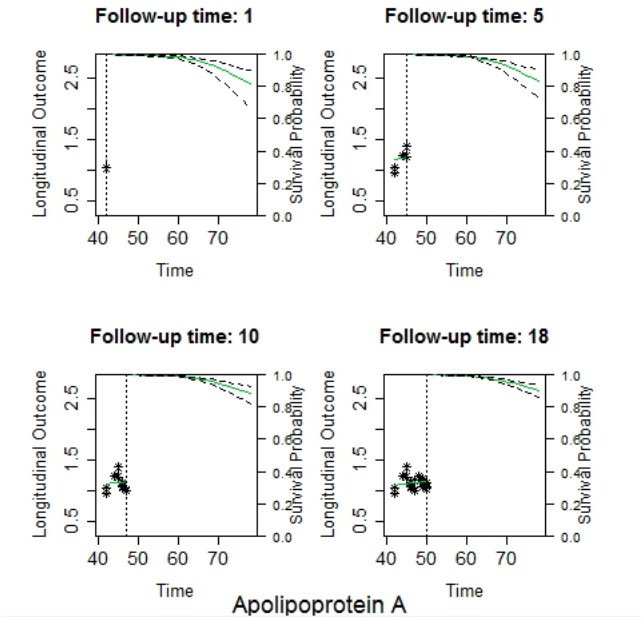
Figure 9: Transformations of predicted survival probabilities for subject 158 677 based on 200 Monte Carlo samples from joint model for Apolipoprotein A.



The first plot is the predicted survival from Figure 9, the second is the survival on the log scale and the last plot is the cumulative risk for subject 158 677. As before, the dashed red line is the median estimate, the solid green line is the mean estimate and the dotted black line is the 95% confidence interval. The cumulative risk is non existent up until the age of 50 when there is a small increase, which after the age of 60 rapidly increases.

To illustrate the dynamic subject specific survival probabilities we construct a for-loop in R which updates $\pi_i(u|t)$ for patient 158 677 after each additional longitudinal measurement of Apolipoprotein A. The results are displayed in Figure 10.

Figure 10: Dynamic survival probabilities for subject 158 677 during follow-up from Apolipoprotein A measurements.

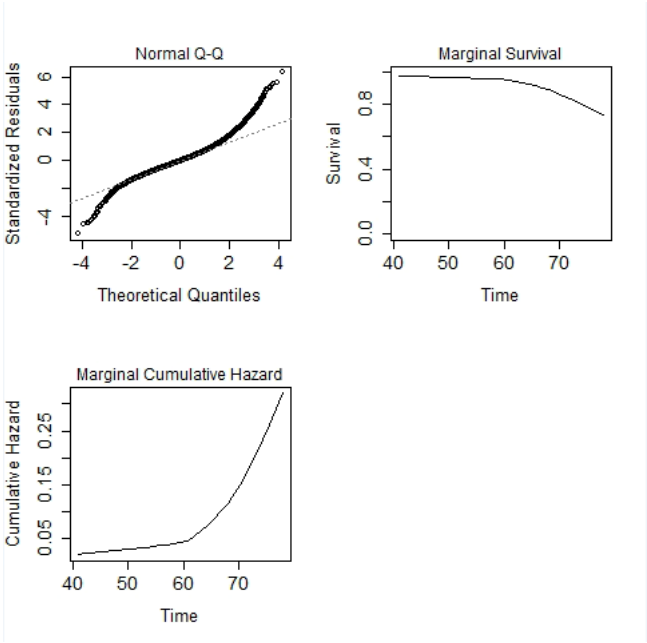


The vertical dotted lines in Figure 10 is the time of the last measurement of Apolipoprotein A at each follow-up. The left side of the vertical line presents the fitted longitudinal trajectory of Apolipoprotein A and the value of these measurements. To the right of the vertical line we have the predicted survival probability at the time of follow-up where the solid green line is the median estimator and the dashed lines is the 95% confidence interval. The first plot in Figure 10 presents the predicted survival after the first follow-up, the second after the 5^{th} follow-up where we have 5 longitudinal measurements available to predict the survival probability. Then follows the predicted survival probability after the 10^{th} follow-up and lastly after the 18^{th}. This subject had a total of 19 longitudinal measurements of Apolipoprotein A in this time span. Studying these four plots we can observe that the more longitudinal measurements of Apolipoprotein A that are collected, the higher is the survival probability for subject 158 677.

Figure 11 below illustrates three diagnostic plots for the fitted joint model obtained from using the function **plot()** on the joint model. The black dotted points in the normal QQ-plot declares that the data follows the dotted line in the middle of the plots, but at the two ends the points deviate from the line. This means that this data have more extreme values than if we would have expected it to be of a normal distribution [23]. The marginal survival is calculated as $\frac{\sum_i S_i(t|\hat{\boldsymbol{b}};\hat{\boldsymbol{\theta}})}{n}$ and the cumulative hazard as $H(t) = -\log S(t)$ as we described in Section 5.3. Note that the marginal survival follows the

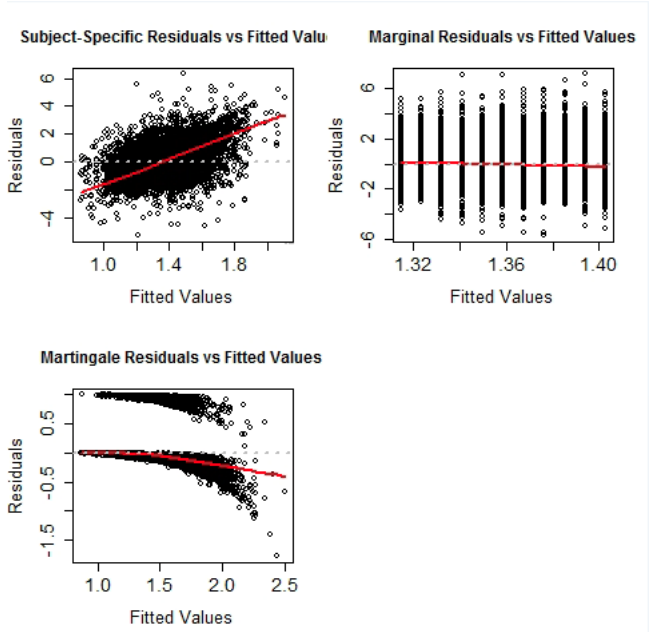
pattern from the subject specific survivals in Figure 8.

Figure 11: Standard diagnostic plots of the joint model with Apolipoprotein A as longitudinal data.



In Section 4.6.1 we defined two residuals for the longitudinal data, namely the subject specific residuals and the marginal residuals. In Figure 12 these two residuals are illustrated as well as the martingale residual was defined in Section 4.6.2 as a residual for the survival data. The y-axis on the top left plot illustrates the subject specific residuals for the longitudinal part with the fitted values on the x-axis. The shape of the points in this plot shows that as the fitted values increase, the variation of the residuals increase as well. As the predictor variable is age, this suggests that there is a variability between the fitted values and the predictor variable age. The second plot illustrates the marginal residuals vs the fitted values. The red line is a fitted loess curve which is a vertical line following zero. This suggests that the covariance for each subject is on the form $\boldsymbol{V_i}$ from Section 4.6.1. The final plot illustrates the martingale residuals and fitted values where we can observe that for larger values of Apolipoprotein A the fitted loess curve in red deviates from zero,

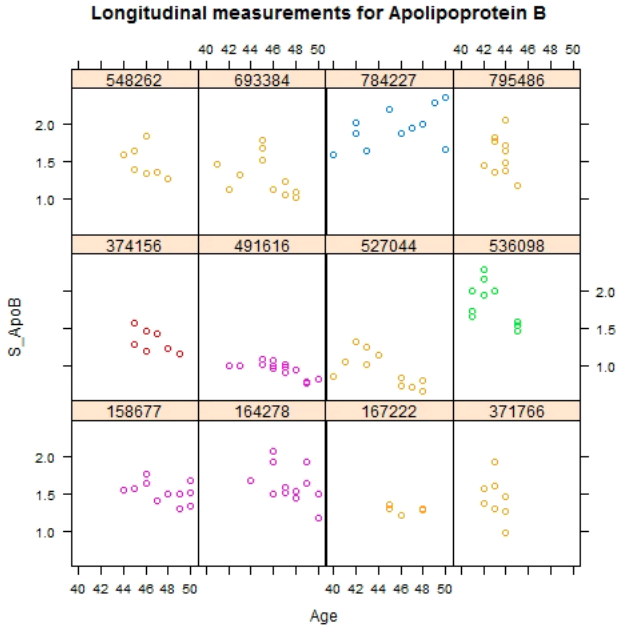
Figure 12: Diagnostic plots of the joint model with Apolipoprotein A as longitudinal data.



6.3 Apolipoprotein B

In Figure 13 we have a collection of plots representing longitudinal measurements of Apolipoprotein B for 12 different subjects, the same subjects as in Figure 7. From Section 5.2 we know that a value of Apolipoprotein B higher than 0.9 g/L indicates a higher risk of developing a cardiovascular disease [62]. We can observe that all the subjects in the first and last row have values over this threshold. Only the two subjects in the middle, subject 491 616 and 527 044 have values lower than 0.9 g/L. From the data, none of the subjects from the middle row, that is subjects 374 156, 491 616, 527 044, 536 098 nor subject 371 766 had developed a cardiovascular disease.

Figure 13: Longitudinal measurements of Apolipoprotein B for some subjects.



We begin the analysis by fitting the data to the extended Cox model to later compare the results obtained from the joint model.

Table 5: Parameter estimates from extended Cox model fit on Apolipoprotein B.

	coef	exp(coef)	se(coef)	z	p
Age	-0.04	0.96	0.01	-5.92	0.00
S-ApoB	0.30	1.35	0.05	6.61	0.00

The result of the extended Cox model in Table 5 indicates that the estimate for Apolipoprotein B is 0.30 which means that there is a $\exp(0.30) = 1.35$ fold increased risk, with a 95% confidence interval (1.24, 1.48), that the i^{th} subject dies when there is a one unit increase in Apolipoprotein B.

The joint model with Apolipoprotein B as longitudinal data and age as predictor was fitted with the resulting parameter estimates presented in Table 6.

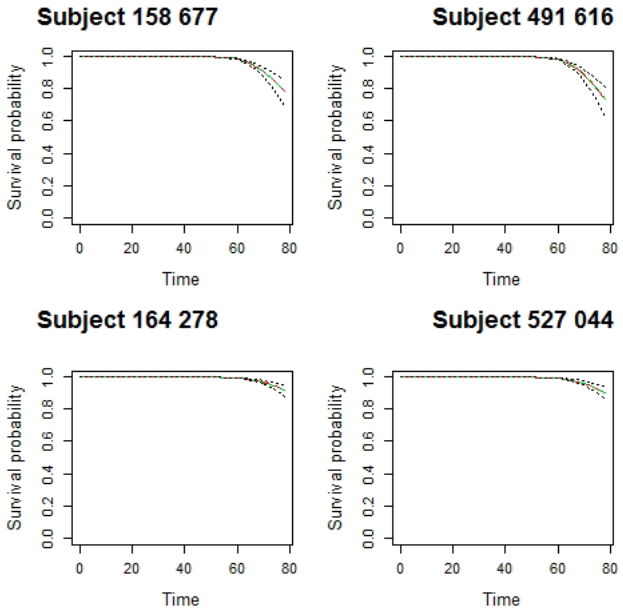
Table 6: Parameter estimates for Apolipoprotein B, standard errors and p-values under the joint modeling analysis. Σ_{ij} denote the ij-element of the covariance matrix for the random effects.

	Event Process				Longitudinal Process		
	Value	Std.Err	p-value		Value	Std.Err	p-value
Age	-0.0431	0.0071	< 0.0001	(Intercept)	0.6336	0.0250	< 0.0001
Assoct	1.9165	0.0643	< 0.0001	Age	0.0164	0.0005	< 0.0001
$\log(\xi_1)$	-7.7093	0.3250		$\log(\sigma)$	-1.4331	0.0064	
$\log(\xi_2)$	-6.1812	0.3431					
$\log(\xi_3)$	-6.2027	0.3502		Σ_{11}	0.0758	0.0012	
$\log(\xi_4)$	-5.9360	0.3558		Σ_{12}	-0.0005	0.0000	
$\log(\xi_5)$	-6.0570	0.3642		Σ_{22}	0.0000	0.0000	
$\log(\xi_6)$	-5.6302	0.3687					
$\log(\xi_7)$	-5.5307	0.3781					

The parameter estimates for Apolipoprotein B from the joint model fit in Table 6 states that for one unit increase of Apolipoprotein B, there is a $\exp(1.92) = 6.80$ fold , with a 95% confidence interval (5.99, 7.71), risk of death. This value is considerably higher compared to the risk from the extended Cox model in Table 5. The standard error of covariance matrix Σ_{12} is 0.00001 and the value of covariance matrix Σ_{22} is 0.00002 with standard error equal to 0.0000002.

Figure 14 illustrates survival probabilities for four different subjects who were all alive at the end of the study. The green solid line represents the mean estimator from Equation (4.35), the red dashed line that coincide with the green line is the median estimator defined in Equation (4.34) with a 95% confidence interval.

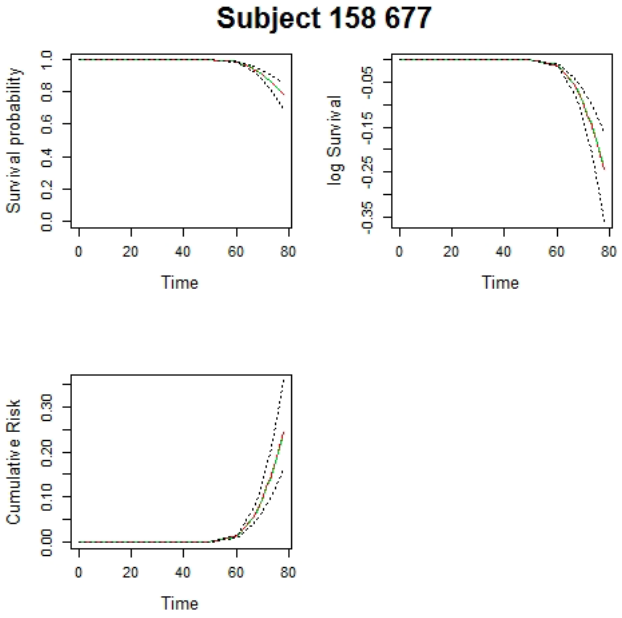
Figure 14: Predicted survival probabilities from joint model fit on Apolipoprotein B.



Studying the four predicted survival probabilities of the four different subjects we observe that the two top plots have a predicted survival probability at age 76.6 of approximately 0.7-0.8. The bottom two plots have a predicted probability of 0.9 of surviving to age 76.6. If we compare the survival probabilities with the longitudinal measurements of Apolipoprotein B for these four subjects in Figure 13 we can conclude that subject 158 677 and 527 044 had low values of Apolipoprotein B and survival probabilities of approximately 0.8 and 0.9 respectively. Subjects 491 616 and 164 278 had high values of Apolipoprotein B with survival probabilities of approximately 0.7 and 0.9 respectively.

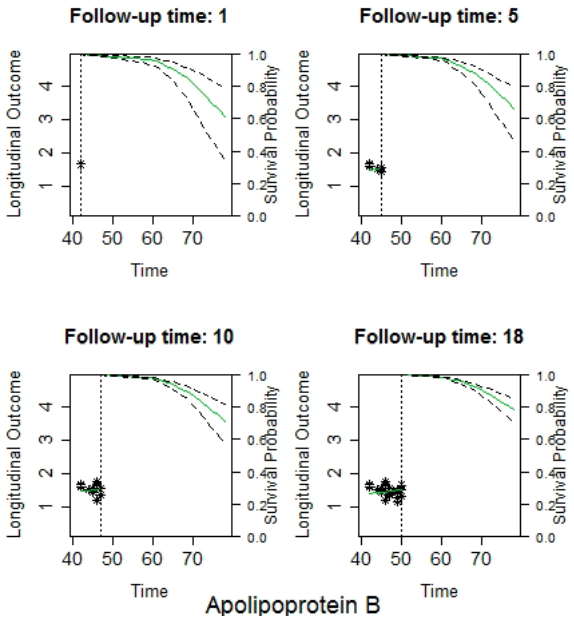
From the predicted survival probabilities we can transform obtain subject specific log survival and cumulative risk based on 200 Monte Carlo samples. Figure 15 illustrates the predicted survival probability, the log survival and the cumulative risk for subject 158 677. The same pattern is observed for the cumulative risk as the corresponding plot for Apolipoprotein A. In that plot, the cumulative risk reached a maximum value of 0.10 while in Figure 15 we observe that the cumulative risk reaches a higher value of 0.25.

Figure 15: Transformations of predicted survival probabilities for subject 158 677 based on 200 Monte Carlo samples from joint model for Apolipoprotein B.



As was described for Figure 10 in the previous section 6.2, we plot the dynamic survival probabilities for subject 158 677 during four follow-ups to observe how the survival probability changes when more information about this subject is known. Figure 16 presents the survival probability after follow-up from Apolipoprotein B measurements.

Figure 16: Dynamic survival probabilities for subject 158 677 during follow-up from Apolipoprotein B measurements.



From the first plot in Figure 16 we observe that the survival probability is 0.6 and the Apolipoprotein B measurement is almost 2 g/L which is higher than the threshold value of 0.9 g/L. After follow-up 5 and 10, the probability of survival has increased and the measurements have become lower than the first measurements. In the last plot, after the 18^{th} out of 19 follow-ups in this age span, the survival probability has increased to 0.8 with Apolipoprotein B measurements between 1 and 2 g/L.

The Q-Q plot from the diagnostic plots in Figure 17 shows that the data follows a normal distribution but at the end tails it deviates from the line, suggesting that we have more extreme values than the normal distribution assumption. The marginal survival and marginal cumulative hazard follows the same curves as the survival curve and cumulative hazard curve in Figure 15.

Figure 17: Standard diagnostic plots of the joint model with Apolipoprotein B as longitudinal data.

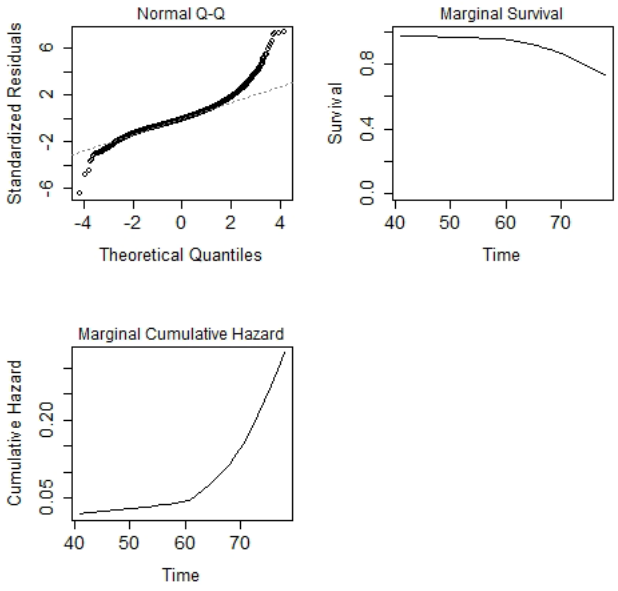
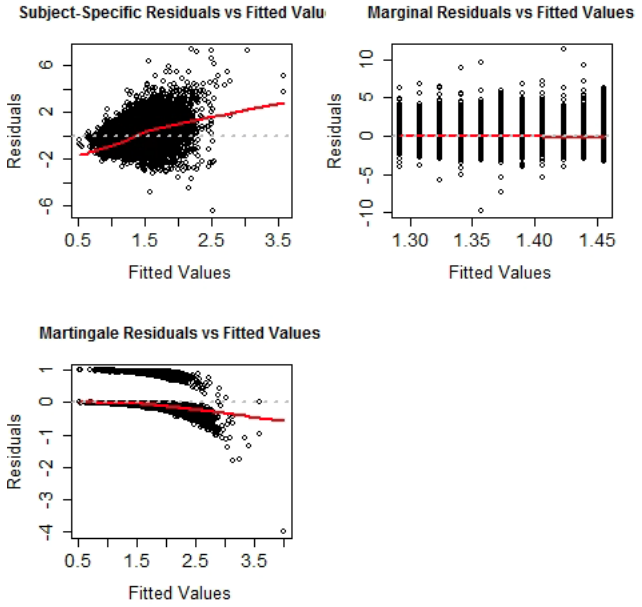


Figure 18 illustrates the residual plots of the subject specific and marginal residuals for the longitudinal part and the martingale residuals for the survival part. From the shape of the points in this plot we observe clearly that the variation increases as the fitted values increase. Thus, as the predictor variable is age, this suggests that the measurement for Apolipoprotein B varies depending on the age at measurement. The loess curve in the marginal residual curve follows a horizontal line at zero which suggests a covariance on the form $\boldsymbol{V_i}$ and from the martingale residual plot we can observe that for larger values of Apolipoprotein B, the residual deviates from zero.

Figure 18: Diagnostic plots of the joint model with Apolipoprotein B as longitudinal data



6.4 Total cholesterol

In this section we repeat the analysis from Section 6.2 and 6.3 but with total cholesterol as longitudinal measurements. To begin with, we study some longitudinal measurements of total cholesterol for 12 different subjects shown in Figure 19, the same selection of subjects as in the previous sections. In Section 5.2 the definition of total cholesterol was stated to be a measurement of both the good and bad cholesterol. A total cholesterol value of under 5.2 mmol/L is characterised with a healthy person [34]. From the first row of plots in Figure 19 we observe that almost all measurements of total cholesterol for these four subject exceeds 5.2 mmol/L. In the second row we can observe that subject 491 616 has the majority of measurements below this threshold and subject 527 044 has some values under 5.2 mmol/L and the remaining a bit above. The last row represents four subjects where three of them have high total cholesterol measurements and the last two have some measurements below the threshold and the remaining just above.

Figure 19: Longitudinal measurements of total cholesterol for some subjects.

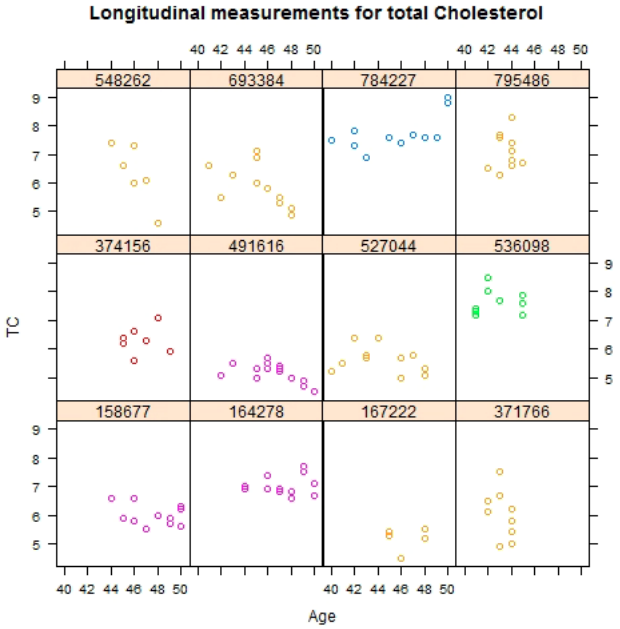


Table 7 represents the results from the extended Cox model with age and total cholesterol as predictors. From the result we can depict that for a one unit increase of total cholesterol, there is a $\exp(0.08) = 1.08$ fold, with a 95% confidence interval (1.05, 1.11), increased risk of death.

Table 7: Parameter estimates from extended Cox model fit on total cholesterol.

	coef	exp(coef)	se(coef)	z	p
Age	-0.04	0.96	0.01	-6.00	0.00
TC	0.08	1.08	0.01	5.34	0.00

We proceed to study the results from the joint model where we used the total cholesterol as longitudinal biomarker.

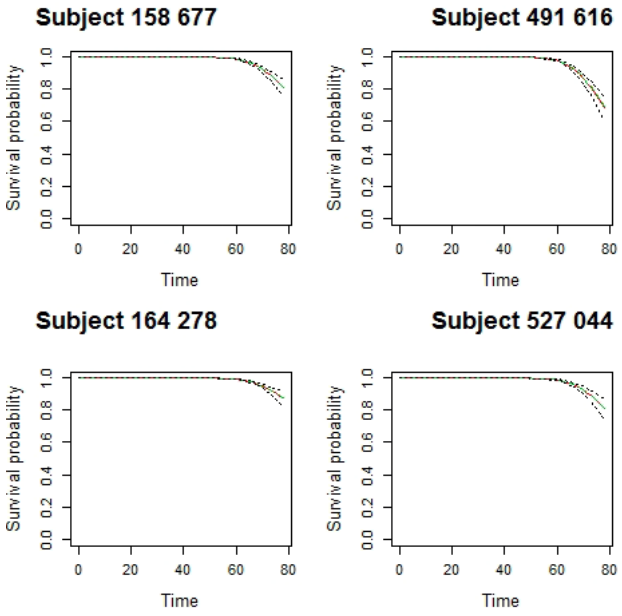
	Event Process				Longitudinal Process		
	Value	Std.Err	p-value		Value	Std.Err	p-value
Age	-0.0408	0.0071	< 0.0001	(Intercept)	3.4533	0.0760	< 0.0001
Assoct	0.6314	0.0210	< 0.0001	Age	0.0589	0.0017	< 0.0001
$\log(\xi_1)$	-9.0475	0.3517		$\log(\sigma)$	-0.3925	0.0067	
$\log(\xi_2)$	-7.5609	0.3743					
$\log(\xi_3)$	-7.6078	0.3816		Σ_{11}	1.1140	0.0144	
$\log(\xi_4)$	-7.3225	0.3871		Σ_{12}	-0.0052	0.0001	
$\log(\xi_5)$	-7.4289	0.3951		Σ_{22}	0.0001	0.0000	
$\log(\xi_6)$	-6.9947	0.3999					
$\log(\xi_7)$	-6.8758	0.4097					

Table 8: Parameter estimates for total cholesterol, standard errors and p-values under the joint modeling analysis. Σ_{ij} denote the ij-element of the covariance matrix for the random effects.

From Table 8 we have that the parameter estimate of the parameter α, describing the association between the longitudinal data and the risk of an event, is equal to 0.63. This means that there is a $\exp(0.63) = 1.88$ fold, with a 95% confidence interval of (1.80, 1.96), increase that the i^{th} subject experience the event, with the event being death, if there is a one unit increase of total cholesterol. Contrary to the analysis of Apolipoprotein A and B, this value is not far from the one obtained from the extended Cox model. Lastly, the standard error of the covariance matrix Σ_{22} is 0.0000006.

The predicted survival probabilities for four subjects who did not experience the event at the end of the study is illustrated in Figure 20.

Figure 20: Predicted survival probabilities from joint model fit on total cholesterol.



The top left plot illustrates that the predicted survival probability for subject 158 677 is approximately 0.8. The following top plot indicates that subject 491 616 has a lower survival probability of 0.7. The predicted value for the first subject of the two bottom plots is between 0.8 and 0.9 and the last plot shows a survival probability of 0.8. Comparing these results with the corresponding plots for Apolipoprotein A, the predicted survivals when we model on total cholesterol are all slightly higher. The greatest difference in survival probability is found for subject 527 044 who in Figure 20 has a predicted value of 0.8 whilst when we model for Apolipoprotein A the probability is 0.3. Comparing the probabilities the results when we model on Apolipoprotein B, the survivals are approximately the same.

In Figure 21 the predicted survival for subject 158 677 is transformed into the log survival and the cumulative risk. In the plot for cumulative risk, we observe that the risk is at 0 up until the subject reaches the age of 60, then there is a steep increase of the risk to a value of 0.20, a value between the cumulative risks from the previous sections.

Figure 21: Transformations of predicted survival probabilities for subject 158 677 based on 200 Monte Carlo samples from joint model for total cholesterol.

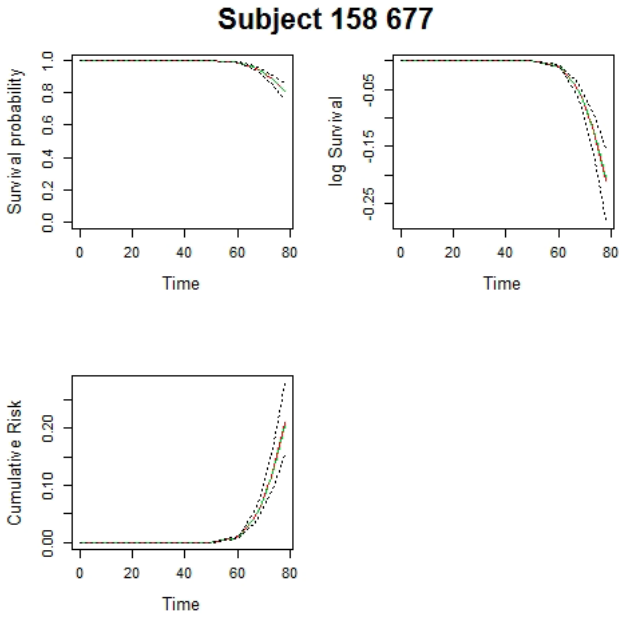
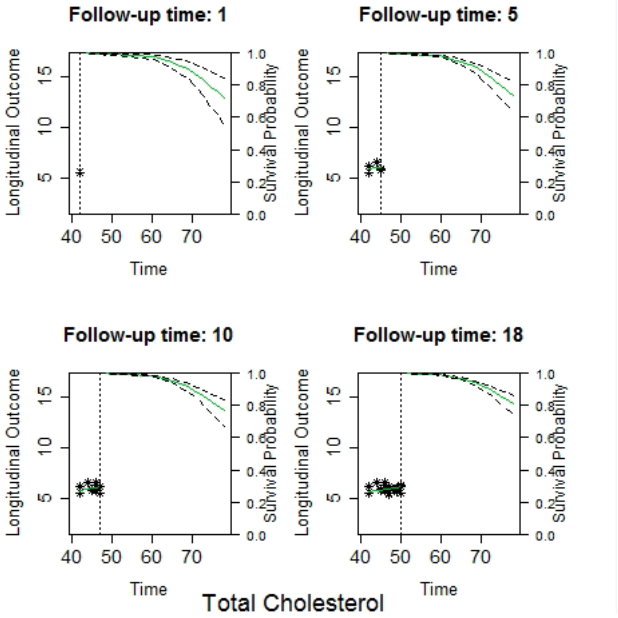


Figure 22 illustrates the dynamic survival probabilities for subject 158 677 after four different follow-ups.

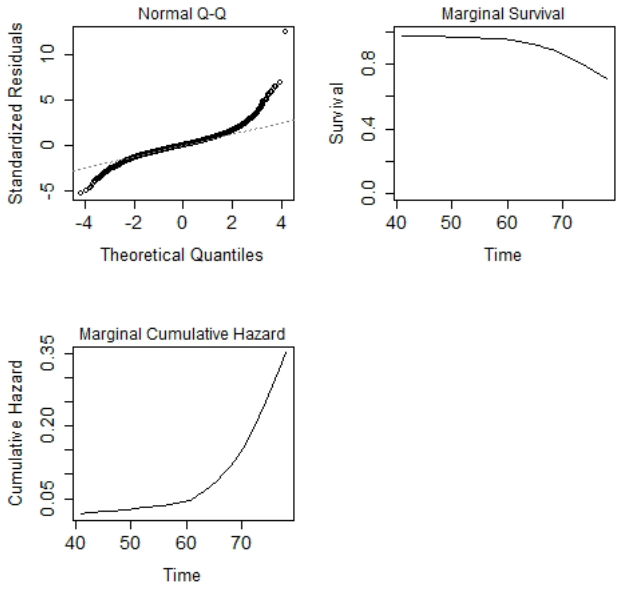
Figure 22: Dynamic survival probabilities for subject 158 677 during follow-up from total cholesterol measurements.



Similarly to the corresponding plots in previous sections, the survival probability in Figure 22 is low after the first follow-up and then increases as more longitudinal measurements are obtained from the subject and hence more information is known to better predict the survival probability. For this specific subject, the measurements for total cholesterol are roughly the same for all follow-ups and all the measurements are slightly higher than the threshold value of 5.2 mmol/L.

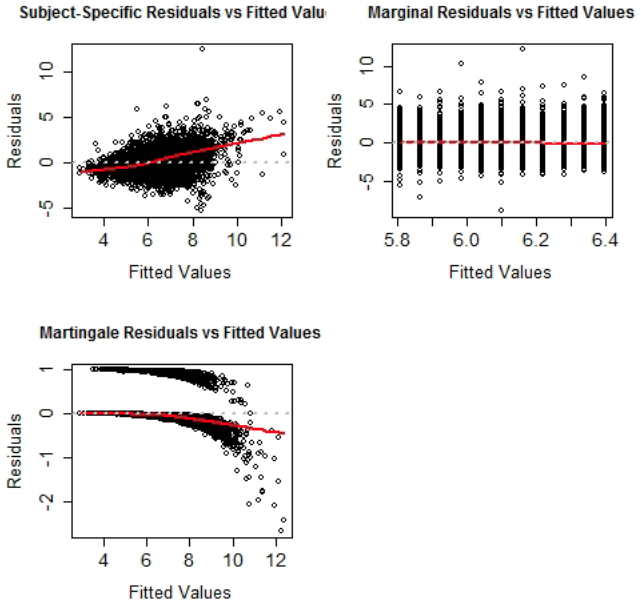
The descriptive plots in Figure 23 suggests that the data is more extreme than the normal distribution assumption as the two endpoints varies from the straight line. The survival and cumulative hazard curves in Figure 21 resembles the marginal survival and marginal cumulative hazard in the plots below.

Figure 23: Standard diagnostic plots of the joint model with total cholesterol as longitudinal data.



As for the residual plots of Apolipoprotein B, in the first residual plot for total cholesterol in Figure 24, the variability of the residual increases as the fitted values increase suggesting that the variability of the measurements for total cholesterol for the different ages, between 40 and 50, in the fitted joint model is not the same. The horizontal fitted loess curve in the marginal residual plot indicates that the covariance for each subject is on the same form. The fitted loess curve follows the horizontally zero line at first but for the largest fitted values, it deviates a little from zero.

Figure 24: Diagnostic plots of the joint model with total cholesterol as longitudinal data.



6.5 Triglycerides

Now we have reached the last analysis where we fit the joint model with triglycerides as longitudinal biomarker. Triglycerides are the fat in the blood and a healthy person should have a value under 1.7 mmol/L [28]. Figure 25 illustrates triglycerides measurements from different follow-ups together with the age of the subjects at the time of follow-up. The selection of subjects is the same as for the previous analysis. From the first row of subjects we observe that all of them have some measurements below the threshold value of 1.7 mmol/L and that the spread of values is quite wide and varies between approximately 0.5 and 4.5. The measurements for the first three subjects in the middle row are all below 1.7 mmol/L but the last subject has values that are all higher. Finally, in the last row, the first two subjects have all measurements below the threshold value while the last two have higher values of triglycerides and hence have higher risk of developing cardiovascular diseases. From data we know that all subjects expect the ones in the middle row and subject 371 766, had developed cardiovascular disease by the end of the study.

Figure 25: Longitudinal measurements of triglycerides for some subjects.

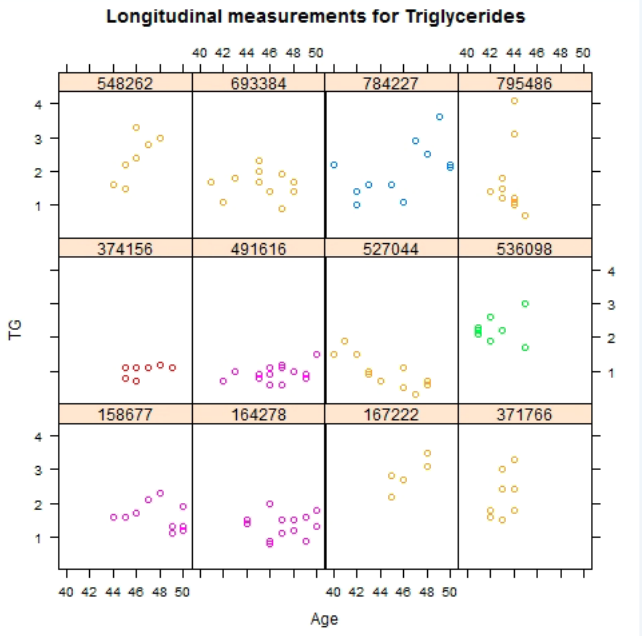


Table 9 presents the parameter estimates obtained from the extended Cox model where we modelled the survival with age and triglycerides as predictors. From this table we can declare that for a one unit increase of triglycerides, there is a $\exp(0.12) = 1.12$ fold, with a 95% confidence interval of (1.10, 1.14), increase to experience the event.

Table 9: Parameter estimates from extended Cox model fit on triglycerides.

	coef	exp(coef)	se(coef)	z	p
Age	-0.04	0.96	0.01	-5.85	0.00
TG	0.12	1.12	0.01	13.48	0.00

The results from the joint model fit with age as predictor variable are presented in Table 10.

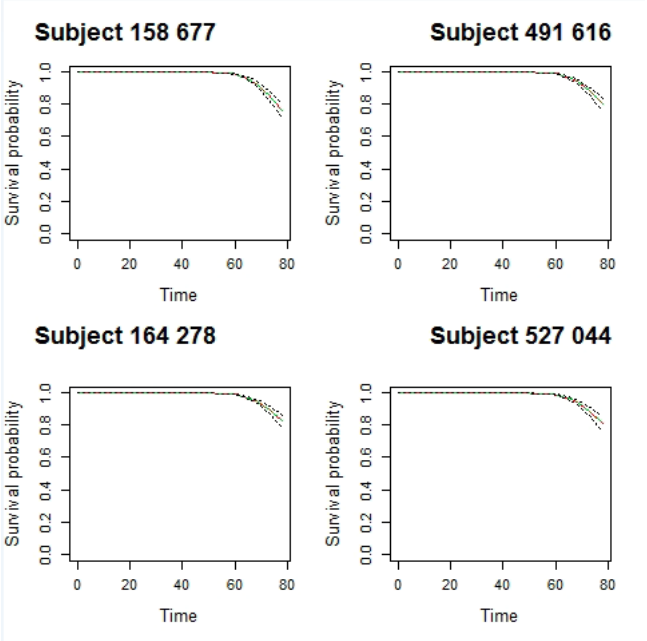
Table 10: Parameter estimates for triglycerides, standard errors and p-values under the joint modeling analysis. Σ_{ij} denote the ij-element of the covariance matrix for the random effects.

	Event Process				Longitudinal Process		
	Value	Std.Err	p-value		Value	Std.Err	p-value
Age	-0.0189	0.0070	0.0073	(Intercept)	-0.7775	0.1000	< 0.0001
Assoct	0.4352	0.0116	< 0.0001	Age	0.0548	0.0022	< 0.0001
$\log(\xi_1)$	-6.9510	0.3182		$\log(\sigma)$	-0.0643	0.0060	
$\log(\xi_2)$	-5.2201	0.3299					
$\log(\xi_3)$	-5.1873	0.3363		Σ_{11}	1.1232	0.0171	
$\log(\xi_4)$	-4.8984	0.3413		Σ_{12}	-0.0029	0.0001	
$\log(\xi_5)$	-5.0039	0.3492		Σ_{22}	0.0000	0.0000	
$\log(\xi_6)$	-4.5349	0.3521					
$\log(\xi_7)$	-4.3849	0.3604					

The estimate for the association parameter in from the joint model fit presented in Table 10 indicates a $\exp(0.44) = 1.54$ fold, with a 95% confidence interval of (1.51, 1.58), increased risk of death for a one unit increase in triglycerides. This estimate gave a similar increase of risk compared to the result from the extended Cox regression model. The value of the covariance matrix Σ_{22} is 0.00004 with standard error 0.0000004.

Figure 26 illustrates the predicted survival probabilities for four different subjects who were alive at the end of the study. We can observe that the probabilities are approximately the same, at a value of 0.8, for all the subjects.

Figure 26: Predicted survival probabilities from joint model fit on triglycerides.



The predicted survival probabilities are in the next plots, Figure 27 transformed into log survival and cumulative risk. Similarly to what we previously have observed, the cumulative risk is at 0 until the age of 60 when there is a large increase in the risk of death.

Figure 27: Transformations of predicted survival probabilities for subject 158 677 based on 200 Monte Carlo samples from joint model for triglycerides.

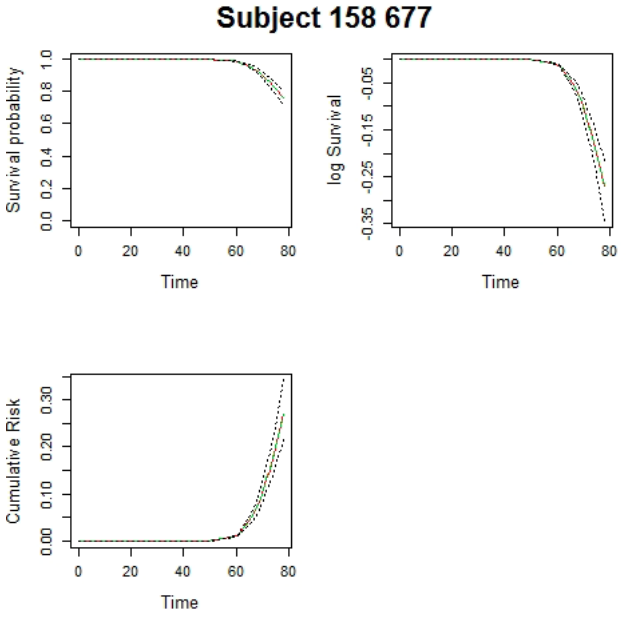


Figure 28 illustrates how the survival probability for subject 158 677 evolves over time as more information about the level of triglycerides are added. After the first follow-up, the survival probability is almost 0.6 with the triglyceride value over 1.7 mmol/L. At the fifth follow-up, the triglyceride levels are lower and the survival probability has increases to about 0.7. Then finally after the 18^{th} follow-up, the survival probability has reached almost 0.8.

Figure 28: Dynamic survival probabilities for subject 158 677 during follow-up from triglycerides measurements.

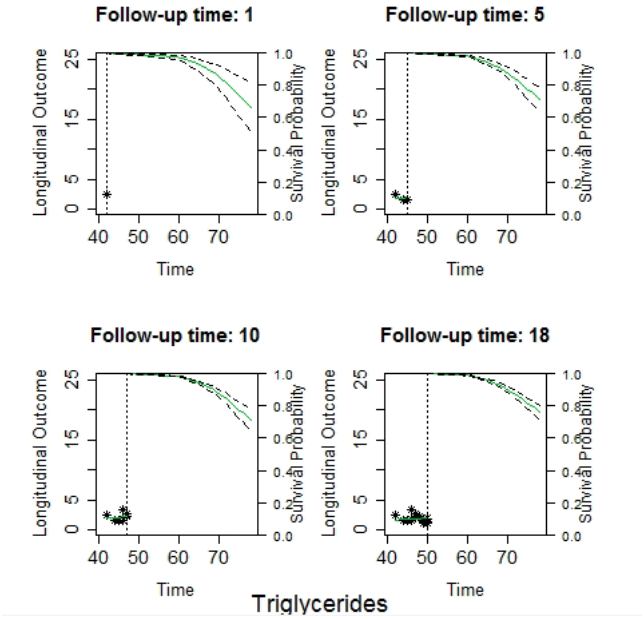


Figure 29 illustrates some diagnostic plots of the fitted model. The Q-Q plot illustrates an even more deviation from the normal distribution than the previous Q-Q plots indicated. This means that the joint model with triglycerides as longitudinal data have more extreme values than the previous joint models. The marginal survival was fitted by $\frac{\sum_i S_i(t|\hat{\boldsymbol{b}};\hat{\boldsymbol{\theta}})}{n}$ and the marginal cumulative hazard as $H(t) = -\log S(t)$.

Figure 29: Standard diagnostic plots of the joint model with triglycerides as longitudinal data.

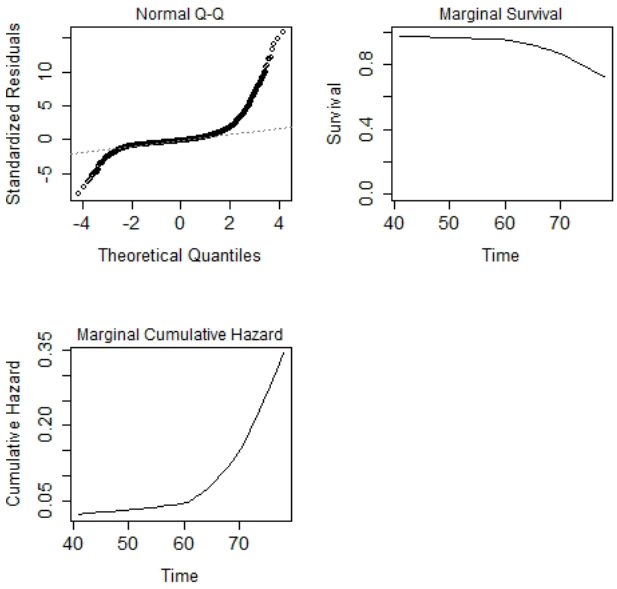
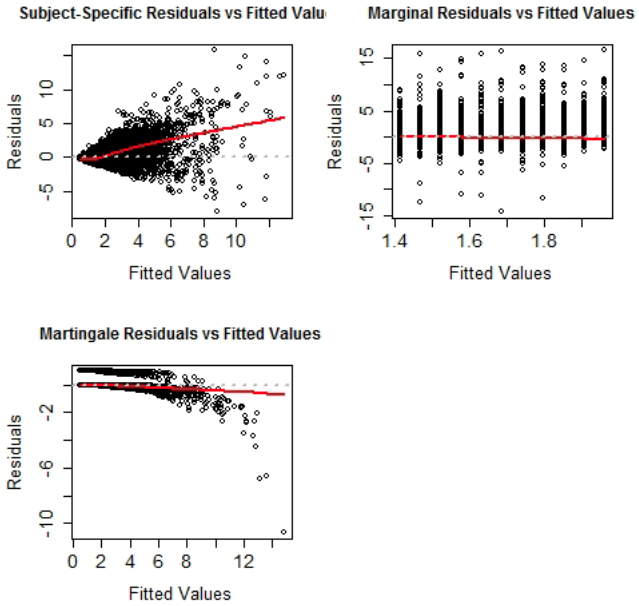


Figure 30: Diagnostic plots of the joint model with triglycerides as longitudinal data.



The last plots in this section illustrates in Figure 30 diagnostic plots for

the longitudinal and survival parts of the joint model. The black points in the subject specific residuals plots show clearly that as the value of the fitted value increases, the variability of the fitted variables increase. The marginal residuals in the second plot presents a horizontally fitted loess line suggesting that the subjects have the same variance of the form $\boldsymbol{V_i}$, see Section 4.6.1. The fitted loess curve in the martingale residual plot indicates that the subjects all fit well to the model.

7 Conclusion and Discussion

In this section the results from the data analyses are summarised. Some of the limitations that appeared during the process of writing this thesis are discussed together with some of my personal remarks on the results. Finally, in Section 7.3, some suggestions for future research on this topic are proposed.

7.1 Conclusion

The joint model combines the survival analysis, that is, the time until an event, together with the longitudinal observations to make predictions of survival probability. Further, the joint model provides a value representing the increased/decreased risk of death at a one unit increase of a specific longitudinal measurement. The joint model also allows producing residual plots.

In this thesis, we have learned and used the theory of a joint longitudinal and survival model to predict subject specific survival probabilities with longitudinal measurements in the form of Apolipoprotein A, Apolipoprotein B, cholesterol and triglycerides. In addition, we also produced residual plots and compared the results of the increased/decreased risk of death at a one unit increase of the longitudinal observations from the joint model, with the results from the extended Cox model. The data used in this thesis are a subset from the AMORIS cohort where we selected to include all men aged 40-50 at observation who had provided measurements of the four longitudinal measurements: Apolipoprotein A, Apolipoprotein B, cholesterol and triglycerides. This resulted in a data containing 23 768 subjects with a total of 33 930 observations.

The parameter estimate for Apolipoprotein A in the joint model indicated that a one unit increase of Apolipoprotein A results in a 18.98 fold increase in the risk of death. A rather high increase compared to the estimate from the extended Cox model that indicated a 1.13 fold increase of the risk of death.

The parameter estimate for Apolipoprotein B in the joint model indicated that a one unit increase of Apolipoprotein B results in a 6.80 fold increase

in the risk of death, whereas the estimate from the extended Cox model indicated a 1.35 fold increase of the risk of death.

Moving on to the results from the parameter results for the joint model with total cholesterol as longitudinal measurements we obtained that a one unit increase of total cholesterol results in a 1.88 fold increase in the risk of death. The estimate from the extended Cox model indicated a 1.08 fold increase of the risk of death.

Finally, the parameter estimates from the joint model where triglycerides were the longitudinal measurements indicated a 1.54 fold increase of the risk of death at one unit increase of triglycerides. The corresponding value from the extended Cox model indicated a 1.12 fold increase.

7.2 Discussion

In Table 4 the parameter estimates for the joint model on Apolipoprotein A are presented. In the table we found that the association parameter α was equal to 2.9 which means that the a one unit increase of Apolipoprotein A is associated with a $\exp(2.94) = 18.98$ fold risk of death. However, these results do not seem to be reliable as first of all, a higher value of Apolipoprotein A is associated with a smaller risk of developing a heart disease or cardiovascular disease according to the definition of Apolipoprotein A. Second, I also performed a fit model on the AMORIS data, but for the age span 30-60 years for Apolipoprotein A and the result of this fit was that the estimated associated parameter was $\exp(-\alpha) = 1.23$ which implies a 1.23 fold risk of death for a one unit increase of Apolipoprotein A. This result appears to be much more reasonable and the result concurs better with the result from the extended Cox model in Table 3 which indicated a 1.05 fold increase of the risk of death if a one unit decrease of Apolipoprotein A.

In the study of the predicted survival probabilities for Apolipoprotein A we observed that the subject with the lowest Apolipoprotein A values, subject 158 677, had the highest survival probability of approx 0.9 and the subject with the highest Apolipoprotein A values, subject 527 044, had the lowest survival probability of approx 0.4. This is a contradiction to the theory of Apolipoprotein A as a higher value indicates a low risk for heart disease, which can lead to death. As both the joint model fit and the survival probability fit contradicts to the theory of Apolipoprotein A, we can conclude that based on this data of men aged 40-50, the true association between Apolipoprotein A and the survival probability is not justified.

The parameter estimate results from the joint model of Apolipoprotein B seems more reasonable as we got the result that for a one unit increase of Apolipoprotein B, the risk of death is 6.80 fold increased, and from the definition of Apolipoprotein B, a higher value indicates a higher risk of developing severe diseases. The same goes for the joint model parameter estimates results for total cholesterol and triglycerides.

Comparing the predicted survival probabilities for the four selected subjects from the four joint models with different longitudinal data, we can note that for Apolipoprotein A, subject 158 677 has the highest survival probability of about 0.9. The probability decreases to 0.8 for Apolipoprotein B and stays on that level for the joint models of total cholesterol and triglycerides. For patient 491 616, the survival probability is 0.8 for Apolipoprotein A, then decreases to about 0.7 for Apolipoprotein B and total cholesterol, to then increase to 0.8 for triglycerides. Subject 164 278 commences with a probability of survival at 0.6 for Apolipoprotein A, but for Apolipoprotein B, triglycerides and total cholesterol, the value is 0.8. In the result of the predicted survival probabilities for Apolipoprotein A, subject 527 044 has the lowest probability of 0.4, which is contradictory looking at this subjects longitudinal measurements in Figure 7, where we observe that the measurements are far over the threshold value. Moreover, the result from the predicted survivals of Apolipoprotein B for this subject indicates a survival probability of over 0.9, to then decrease to 0.8 for total cholesterol and triglycerides. The longitudinal measurements of Apolipoprotein B for subject 527 044 are all low with most of them below the threshold value which can be the reason for the very high survival probability of over 0.9.

Studying Figures 7, 13, 19 and 25 that illustrates the longitudinal measurements of the four different biomarkers, we can observe nonlinear relationships for the selection of subject specific measurements. Rizopoulos (2012b) [54] discussed in his book in Chapter 4.3.7 that nonlinear relationships can cause convergence problems which he suggested can be solved by using, for example, a spline function in the longitudinal submodel. This is an important note for future research on this topic, however, as convergence was not an issue with the dataset used for the analysis in this thesis, this method was not utilised.

7.3 Future work

From the results of this analysis we can conclude that using joint longitudinal and survival submodels is an effective model to analyse and predict survival probabilities. This as we can both predict subject specific survival probabilities based on longitudinal outcomes as well as studying the dynamic survival predictions over time.

Further and better studies can be made using the AMORIS data to implement joint longitudinal and survival models. For example, to obtain a better picture of the data and include as many observations as possible, the whole data can be used to fit the model, or choose an age span between 40 and 65 in which we noted from Figure 4 is the age where most of the observations were made. More predictor variables can also be utilised, such as socioeconomic index and the inflammation indicator S-CRP, both described in Section 5.2. As was discussed in Section 7.2, the longitudinal biomarkers

follow a nonlinear pattern. Further studies can be made to investigate if the results can be improved, by for example, adding a spline function to the longitudinal submodel or adding arguments in the longitudinal function or the joint model function as described in Chapter 4.3.7 in Rizopoulos (2012b) [54].

In this thesis, we chose the event time to be the time of death for the survival part of the joint model. From the AMORIS data we can also analyse the event of the disease AAA and predict the survival time after first diagnosis of AAA or the event of cardiovascular disease.

References

[1] Aalen O.O, Borgan Ø & Gjessing H.K. (2008) Survival and Event History Analysis. Springer

[2] Albumin (Blood) - Health Encyclopedia - University of Rochester Medical Center. (2020). Available at: https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=albumin_blood. Retrieved 29 September 2020.

[3] Airy G. B. (1861). On the Algebraical and Numerical Theory of Errors of Observation and the Combination of Observations. London: Macmillan

[4] Barter P. (1994). Abnormal Laboratory Results: HDL cholesterol testing: implications for clinical management. Australian Prescriber, 17(4), 99-102. doi: 10.18773/austprescr.1994.101

[5] Camilleri L. (2019). History of survival analysis. Times of Malta. Available at: https://timesofmalta.com. Retrieved 14 October 2020.

[6] Collett D. (2003). Modelling Survival Data in Medical Research. Chapman & Hall.

[7] Cox D. (1972). Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B 34, 187 – 220.

[8] Crowther M.J. (2014) Development and application of methodology for the parametric analysis of complex survival and joint longitudinal-survival data in biomedical research. [Doctoral dissertation, University of Leicester]. Available at: https://www.mjcrowther.co.uk/pdf/2014_CROWTHER_MJ_PhD.pdf

[9] Crowther M.J, Abrams K.R, & Lambert P.C. (2012a) Flexible parametric joint modelling of longitudinal and survival data. Stat Med, 31(30):4456–4471.

[10] Dafni U. & Tsiatis A. (1998). Evaluating surrogate markers of clinical outcome measured with error. Biometrics 54, 1445 – 1462.

[11] De Gruttola V & Tu X.M. (1994). Modeling progression of CD4-lymphocyte count and its relationship to survival time. Biometrics 50:1003-1014.

[12] Diabetes.co.uk - the global diabetes community. (2019). Blood sugar level ranges. Available at: https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html. Retrieved 29 September 2020.

[13] Diagnoskoder (ICD-10). Available at: http://icd.internetmedicin.se/diagnos/I212. Retrieved 29 September 2020.

[14] Diagnoskoder (ICD-10). Available at: http://icd.internetmedicin.se/diagnos/I619. Retrieved 29 September 2020.

[15] Durrleman S. & Simon R. (1989). Flexible Regression Models with Cubic Splines. Stat Med, 8(5):551–561.

[16] Edstorp J. (2019). How it all started. Available at: https://ki.se/en/imm/how-it-all-started. Retrieved 7 October 2020.

[17] Edstorp J. (2019). Strengths and limitations. Available at: https://ki.se/en/imm/strengths-and-limitations. Retrieved 7 October 2020.

[18] Edstorp J. (2019). Who are they?. Available at: https://ki.se/en/imm/who-are-they. Retrieved 7 October 2020.

[19] Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh 52, 399–433

[20] Fisher, R. A. (1925). Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd.

[21] Fitzmaurice G & Molenberghs G. (2009). Advances in Longitudinal Data Analysis: An Historical Perspective. doi: 10.1201/9781420011579.pt1.

[22] Fitzmaurice G., Laird N. & Ware J. (2004). Applied Longitudinal Analysis. Wiley, Hoboken.

[23] Ford C. (2015). Understanding Q-Q Plots | University of Virginia Library Research Data Services + Sciences. Available at: https://data.library.virginia.edu/understanding-q-q-plots/. Retrieved 1 November 2020.

[24] Fowler H., Fowler F., & Sykes J. (1987). The concise Oxford dictionary of current English (7th ed., p. 524). London: Guild Publishing.

[25] Freedman D. (2008). Survival Analysis. The American Statistician, 62(2), 110-119. doi: 10.1198/000313008x298439

[26] Goodyear M., Malhotra N. & Seager S. (2018). Life-tables and their demographic applications. Available at: https://www.healthknowledge.org.uk/public-health-textbook/health-information/3a-populations/life-tables-demographic-applications. Retrieved 14 October 2020.

[27] Harrell F (2001). Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer-Verlag, New York

[28] Heart UK - The cholesterol charity. Triglycerides. Available at: https://www.heartuk.org.uk/cholesterol/triglycerides. Retrieved 28 September 2020.

[29] Hedegaard R & Weisstein E.W. "Abscissa." From MathWorld–A Wolfram Web Resource. Available at: https://mathworld.wolfram.com/Abscissa.html. Retrieved 3 November 2020.

[30] Held L. & Bové D.S, (2014) Applied Statistical Inference: Likelihood and Bayes, Springer.

[31] Hjärt-Lungfonden. (2018). Högt kolesterol. Available at: https://www.hjart-lungfonden.se/halsa/riskfaktorer/hogt-kolesterol/. Retrieved 28 September 2020.

[32] Hughson G. (2017). CD4 cell counts. Available at: https://www.aidsmap.com/about-hiv/cd4-cell-counts. Retrieved 20 October 2020.

[33] Ibrahim J.G, Chu H & Chen LM.(2010) Basic concepts and methods for joint models of longitudinal and survival data. Journal of Clinical Oncology 2010; 28:2796–2801

[34] InformedHealth.org. (2017) Cologne, Germany: Institute for Quality and Efficiency in Health Care (IQWiG); 2006-. High cholesterol: Overview. 2013 Aug 14 [Updated 2017 Sep 7]. Available at: https://www.ncbi.nlm.nih.gov/books/NBK279318/. Retrieved 28 September 2020.

[35] Institute of Medicine (US) Committee on Social Security Cardiovascular Disability Criteria. Cardiovascular Disability: Updating the Social Security Listings. Washington (DC): National Academies Press (US); 2010. 7, Ischemic Heart Disease. Available at: https://www.ncbi.nlm.nih.gov/books/NBK209964/. Retrieved 29 Septemeber 2020.

[36] Intracerebral Hemorrhage | Internet Stroke Center. Available at: http://www.strokecenter.org/patients/about-stroke/intracerebral-hemorrhage/. Retrieved 29 September 2020.

[37] Kandola A. (2020). Abdominal aortic aneurysm: Screening, treatment, and symptoms. Availabel at: https://www.medicalnewstoday.com/articles/abdominal-aortic-aneurysm. Retrieved 29 September 2020.

[38] Klein J.P. & Moeschberger M.L. (2003). Survival Analysis: Techniques for Censored and Truncated Data; 2nd edition, Springer USA

[39] Landenhed M, Engström G, Gottsäter A, Caulfield M, Hedblad B, Newton-Cheh C, Melander O, Smith J. (2015). Risk Profiles for Aortic Dissection and Ruptured or Surgically Treated Aneurysms: A Prospective Cohort Study. Journal of the American Heart Association. 4. 10.1161/JAHA.114.001513.

[40] Lee T., Zeng L., Thompson D. & Dean C. (2011). Comparison of imputation methods for interval censored time-to-event data in joint modelling of tree growth and mortality. The Canadian Journal of Statistics / La Revue Canadienne De Statistique, 39(3), 438-457. Available at: http://www.jstor.org/stable/41304476. Retrieved October 16, 2020.

[41] Little R. & Rubin D. (2002). Statistical Analysis with Missing Data, 2nd edition. Wiley, New York.

[42] Maharani A. (2019). Socio-economic inequalities in C-reactive protein levels: Evidence from longitudinal studies in England and Indonesia. Brain, behavior, and immunity, 82, 122-128.

[43] Mandal A. (2019). What is a Biomarker?. News Medical. Available at: https://www.news-medical.net/health/What-is-a-Biomarker.aspx. Retrieved 3 September 2020.

[44] National Center for Health Statistics. Section I - Instructions for classifying the underlying cause of death, 2017. Available at: https://www.cdc.gov/nchs/data/dvs/2a_2017.pdf. Retrieved 28 September 2020

[45] Nobre J & Singer J (2007). Residuals Analysis for Linear Mixed Models. Biometrical Journal, 6, 863–875.

[46] Person A. & Böiers L-C. (2010). Analys i en variabel (3rd ed., p. 203). Lund: Studentlitteratur.

[47] Pickands J. (1971). The Two-Dimensional Poisson Process and Extremal Processes. Journal of Applied Probability, 8(4), 745-756. doi:10.2307/3212238

[48] Pinheiro J, Bates D, DebRoy S & Sarkar D, R Core Team (2019). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-142. Available at: https://R-project.org/package=nlme. Retrieved 11 November 2020.

[49] Piulachs X., Alemany Leira R., Guillén M., & Rizopoulos, D. (2017). Joint models for longitudinal counts and left-truncated time-to event data with applications to health insurance. SORT: statistics and operations research transactions, 41 (2) July-December 2017, 347-372. doi: 10.2436/20.8080.02.63

[50] Ristl R, Ballarini N, Goette H, Schueler A, Posch M & Koenig F.(2020). nph Examples: Delayed treatment effects, treatment switches and heterogeneous patient populations: how to design and analyse RCTs in oncology. Available at: https://cran.r-project.org/web/packages/nph/vignettes/examples.html. Retrieved 2 November 2020.

[51] Rizopoulos D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. Biometrics, 67(3):819–829.

[52] Rizopoulos D. (2010). JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. Journal of Statistical Software, 35(9), 1 - 33. doi:http://dx.doi.org/10.18637/jss.v035.i09

[53] Rizopoulos D. (2012a). Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive Gaussian quadrature rule. Computational Statistics Data Analysis 56, 491 – 501.

[54] Rizopoulos D. (2012b). Joint Models for Longitudinal and Time-to-Event Data With Applications in R. Chapman & Hall.

[55] Rizopoulos D. (2018). JM: Joint Modeling of Longitudinal and Survival Data. R package version 1.4-8. Available at: https://CRAN.R-project.org/package=JM. Retrieved 2 November 2020.

[56] Rizopoulos D, Verbeke G & Molenberghs G (2010). Multiple-Imputation-Based Residuals and Diagnostic Plots for Joint Models of Longitudinal and Survival Outcomes. Biometrics, **66**, 20–29. doi: 10.1111/j.1541-0420.2009.01273.x

[57] Rooth E. (2019). Haptoglobin Test: Purpose, Procedure, and Results. [Updated 2019-01-24] Available at:https://www.healthline.com/health/haptoglobin. Retrieved 28 September 2020

[58] Royston P. & Parmar M.K.B.(2002). Flexible Parametric Proportional Hazards and Proportional Odds Models for Censored Survival Data, with Application to Prognostic Modelling and Estimation of Treatment Effects. Stat Med, 21(15):2175–2197, doi: 10.1002/sim.1203

[59] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA http://www.rstudio.com/.

[60] Rubin D. (1976). Inference and missing data. Biometrika 63, 581 – 592

[61] Sahlgrenska Universitetssjukhuset. (2019). ApoA1. [Updated 2019-06-02] Available at: https://www.sahlgrenska.se/for-dig-som-ar/vardgivare/laboratoriemedicin/analyslista/apoa1/. Retrieved 28 September 2020.

[62] Sahlgrenska Universitetssjukhuset. (2019). ApoB. [Updated 2019-06-02] Available at: https://www.sahlgrenska.se/for-dig-som-ar/vardgivare/laboratoriemedicin/analyslista/apob/17000.html. Retrieved 28 September 2020.

[63] Sahlgrenska Universitetssjukhuset. (2020). Heptaglobin. [Updated 2020-05-11] Available at: https://www.sahlgrenska.se/for-dig-som-ar/vardgivare/laboratoriemedicin/analyslista/haptoglobin/. Retrieved 28 September 2020.

[64] Self S. & Pawitan Y. (1992). AIDS Epidemiology: Methodological Issues, chapter Modeling a marker of disease progression and onset of disease. Birkhauser, Boston.

[65] Statistiska Centralbyrån. Folk- och bostadsräkningen (FoB) 1975. Available at: https://www.scb.se/contentassets/c0dbe46b69f64b90b221bfaaff678d45/be0205_do_1975_bk_190107.pdf. Retrieved 28 Spetember 2020.

[66] Statistiska Centralbyrån. (1989). Yrkesklassificeringar i FoB 85 enligt Nordisk yrkesklassificering (NYK) och Socioekonomisk indelning (SEI). Stockholm.

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

[68] Taylor J., Park Y., Ankerst D., Proust-Lima C., Williams S. & Kestin L. et al. (2013). Real-Time Individual Predictions of Prostate Cancer Recurrence Using Joint Models. Biometrics, 69(1), 206-213. doi: 10.1111/j.1541-0420.2012.01823.x

[69] Therneau T. (2019). A Package for Survival Analysis in R. R package version 3.1-8. Available at: https://CRAN.R-project.org/package=survival. Retrieved 23 October 2020.

[70] Therneau T, Grambsch P (2000). Modeling Survival Data: Extending the Cox Model. Springer- Verlag, New York.

[71] Tsiatis A. & Davidian M. (2001). A semiparametric estimator for the pro- portional hazards model with longitudinal covariates measured with error. Biometrika 88, 447 – 458.

[72] Tsiatis A.A., DeGruttola V. & Wulfsohn M.S. (1995). Modeling the relationship of survival to longitudinal data measured with error: Applications to survival and CD4 counts in patients with AIDS. J Am Stat Assoc 90:27-37.

[73] Verbeke G & Molenberghs G (2000). Linear Mixed Models for Longitudinal Data. Springer- Verlag, New York.

[74] What Are Longitudinal Data? | National Longitudinal Surveys. (2020). Available at: https://www.nlsinfo.org/content/getting-started/what-are-longitudinal-data. Retrieved 3 September 2020.

[75] Werlabs AB. Apolipoprotein A1 (Apo A1) Available at: https://werlabs.se/halsokontroll/hjart-och-karlsjukdom/apo-a Retrieved 28 September 2020.

[76] Werlabs AB. Apolipoprotein B (Apo B) Available at: https://werlabs.se/halsokontroll/hjart-och-karlsjukdom/apo-b/ Retrieved 28 September 2020.

[77] Werlabs AB. Glukos (Fastesocker) - Blodsocker & Diabetes. Available at: https://werlabs.se/halsokontroll/diabetes-metabolstorning/glukos. Retrieved 28 September 2020.

[78] Werlabs AB. Högkänslight CRP - Inflammation. Available at: https://werlabs.se/halsokontroll/inflammation/hs-crp. Retrieved 28 September 2020.

A Appendix

A.1 Maximum Likelihood Estimation

The computation to obtain the maximum likelihood estimate for β in the linear mixed effects model is derived in this section. From Eq. (3.6), the likelihood of the linear mixed effects model which is a multivariate normal distribution.

$$L_i(\boldsymbol{\theta}) = p(y_i; \boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^{n_i}|V_i|}} \exp\left\{-\frac{(y_i - X_i\beta)^T V_i^{-1}(y_i - X_i\beta)}{2}\right\}.$$

.

The log-likelihood is then given as,

$$\log L_i(\boldsymbol{\theta}) = -\frac{n_i}{2}\log(2\pi) - \frac{1}{2}\log(|V_i|) - \frac{1}{2}(y_i - X_i\beta)^T V_i^{-1}(y_i - X_i\beta).$$

The partial derivative, with respect to β gives

$$\frac{d\log L_i(\boldsymbol{\theta})}{d\beta} = -\frac{1}{2}\left(-2X_i^T V_i^{-1} y_i + 2X_i^T V_i^{-1} X_i\beta\right) = X_i^T V_i^{-1} y_i - X_i^T V_i^{-1} X_i\beta$$

Set this expression to 0 and solve the ML estimator $\hat{\beta}$

$$\hat{\beta}_{ML} = \left(X_i^T V_i^{-1} X_i\right)^{-1} X_i^T V_i^{-1} y_i$$

And, finally for $i = 1, ..., n$ the ML estimator is

$$\hat{\beta}_{ML} = \left(\sum_{i=1}^{n} X_i^T V_i^{-1} X_i\right)^{-1} \sum_{i=1}^{n} X_i^T V_i^{-1} y_i$$