Master's Thesis

Joint Modelling for Longitudinal and Time-to-Event Data. Application to Liver Transplantation Data

Author: **Laura Calaza Díaz**

Supervisors: **Carmen Cadarso Suárez**
Francisco Gude Sampedro

Master in Statistical Techniques
University of Santiago de Compostela
June 2014

"Caminante no hay camino,
el camino se hace al andar."
- Antonio Machado

Doña Carmen Cadarso Suárez, Profesora de Estadística e Investigación Operativa de la Universidad de Santiago de Compostela,

Don Francisco Gude Sampedro, Adjunto la Unidad de Epidemiología del Hospital Clínico Universitario de Santiago de Compostela,

Certifican:

Que el presente trabajo, titulado "Joint Modelling for Longitudinal and Time-to-Event Data. Application to Liver Transplantation data", realizado bajo su dirección, y que presenta Laura Calaza Díaz como Proyecto Fin de Máster correspondiente al tercer cuatri-mestre del Máster en Técnicas Estadísticas, reúne todos los requisitos exigidos por la normativa vigente.

Santiago de Compostela, 30 de junio de 2014

Contents

Abstract

A common objective in follow-up studies is to characterize the relationship between longitudinal measurements and time-to-event outcomes. For this aim, various methods were proposed in the statistical literature, such as an extended version of the Cox model with longitudinal covariates or a two-stage approach. However, these techniques have several limitations, including the possibility of biased estimations. To avoid these limitations, joint modelling approaches are becoming increasingly popular. In this work, we provide a brief overview of a joint model approach for longitudinal and time-to-event data, focusing on the survival process. Also, the predictive capacity of this model is studied and related computational aspects, including available software, are discussed. The main motivation behind this work relies on the application of the joint modelling to liver transplantation data, in order to investigate the abilities of postoperative glucose profiles to predict patients' survival.

Chapter 1

Introduction

In biomedical studies, periodically measured disease markers are used to monitor progression to the onset of disease or occurrence of death. Longitudinal studies are becoming increasingly popular, especially in biomedical research to gain a better understanding of the risk of disease and even death.

Many of these studies are aimed to characterize the relationship between longitudinal and time-to-event outcomes. In the literature, several methods are proposed to study the association between longitudinal responses and particularly time-to-event survival processes.

The extended Cox model (Andersen and Gill, 1982) and the two stage approach (Self and Pawitan, 1992), were proposed to handle this association, but these methodologies present some limitations. The extended Cox model assumes that the covariates are external and, for that reason, not related to the failure mechanism (Kalbfleisch and Prentice, 2002; Prentice, 1982), and that time-dependent covariates are measured without error. On the other hand, the two-stage approach is neither recommended. It estimates the joint model by using a two-step approach, first studying the longitudinal submodel with linear mixed effects, and then incorporates the random effects to the survival model. Because of this structure, no survival information is used to obtain longitudinal estimates, so informative dropout is not accounted for. These limitations may lead to biased or inefficient results.

Current research shows an increasing popularity of the joint likelihood approaches, due to their efficiency and their advantages comparing with the methodologies mentioned above. Joint models take into account the association between the longitudinal and the survival process by simultaneously determining the parameter estimates for both processes. For that reason, it alleviates the potential bias caused by both the extended Cox-model and the two-stage approach. Focusing our attention on those joint modelling proposals which used shared random effects to their parametrization, we find two different approaches depending on the research interest. Rizopoulos (2010) has proposed a joint model where the time-to-event process is of main interest and influenced by a longitudinal time-dependent covariate measured with error. Also Philipson et al. (2012) developed a shared random effects model where the focus is on both survival and longitudinal processes.

The goal of this work is to illustrate an appropriate methodology for follow-up studies which jointly analyse repeated measurements of biomarkers and event times of individuals. The motivation behind this proposal was an Orthotopic Liver Transplantation (OLT) database. OLT is the established treatment for end-stage liver disease and acute fulminant hepatic failure, and more than 80,000 OLTs have been performed in Europe. Advances in both medical management and surgical techniques have led to an increase in the number of long-term survivors (Dutkowski et al., 2010). Because liver transplant recipients live longer, it is necessary to understand and to anticipate causes of morbidity and mortality. Several investigators have consistently reported a significant association between increased glycemic variability and worse outcome in critically ill patients. In their analysis (Dossett et al., 2008; Egi et al., 2006; Krinsley, 2008; Meyfroidt et al., 2010), blood glucose variability is measured by using standard deviation, percentile values, successive changes in blood glucose, and by calculating the coefficient of variation. However, it is recognized that compared with the use of only single-moment biomarker values, serial biomarker evaluations may carry important additional information regarding the prognosis of the disease under study (Wolbers et al., 2010).

In this work we aim to investigate the abilities of postoperative glucose profiles to predict the death of patients who underwent an OLT, distinguishing between patients with and without a previous diagnosis of diabetes mellitus. Because of our interest in the patients survival, the relationships between glucose profiles and the risk of death were modelled by using the Rizopoulos (2010) proposal. In addition, the predictive capacity is analysed by means of time dependent ROC curves (Heagerty and Zheng, 2005).

The outline of this work is as follows. In Chapter 2, a review of longitudinal data analysis and survival analysis is made. This is necessary to introduce the respective sub-models of the joint regression model. In Chapter 3, details of the Joint Modelling approach proposed by Rizopoulos are described. Following-up with a brief description of the available package in the software R (R Core Team, 2014) that allows to implement this methodology, the JM package (Rizopoulos, 2010). Then, a detailed analysis of the results obtained throughout the OLT database is presented in Chapter 4. Finally, in Chapter 5 conclusions are presented and future lines are discussed.

Chapter 2

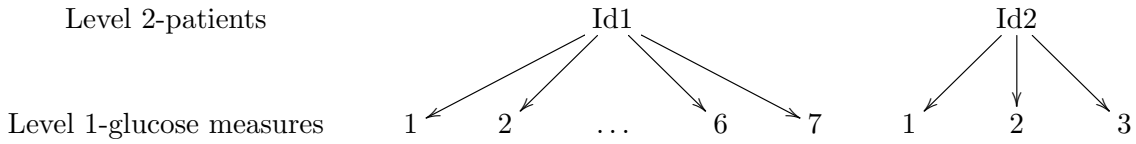
Background

In this chapter we will introduce a brief background to understand the joint modelling approach. Because it involves two major blocks, namely longitudinal data analysis and survival analysis, the following step is to summarize some key concepts which will lead us to understand the proposed joint methodology.

2.1 Longitudinal data analysis

Longitudinal data is often collected in clinical trials, especially in follow-up studies, in which individuals are measured repeatedly across time. For that reason, longitudinal studies are indispensable to study the change in an outcome over time. Due to no independence between these repeated measurements, it will be appropriate to apply mixed models, which constitute an adequate tool to model this dependence by considering a hierarchical structure on the database.

Hierarchies are used to perform the dependence relation between individuals and the groups they belong to. For example, in our study glucose profiles are taken in a sample of surgery patients, so we will identify a two-level structure: glucose measurements (level 1) grouped by individuals (level 2).



Observations are treated as clustered data, grouped into disjoint classes according to some classification criteria. A particular case are repeated measurements, where observations are made sequentially in the same individual, the cluster. Due to this fact, observations in the same cluster can not be usually considered independent.

Once the data structure is described, the key is to distinguish between the parameters of the model, classified into fixed effects and random effects. In this way, the response or dependent variable is assumed to be a function of fixed effects, non-observable cluster specific random effects, and an error term.

2.1.1 Fixed or random effects

Fixed effects are variables which only include the levels of interest, that is, the objective of the study lies in the cluster comparison, but not in generalizing results to the population. The objective is to study the average effect of predictors on the response. However, for random effects an infinite set of levels are assumed, so our study levels are seen as a sample from that population. The interest is to make inference for the complete population of levels. We are not interested in comparing means, but on how the random effect explains the variability in the dependent variable.

2.1.2 Linear mixed-effects models

In follow-up studies the observations of subjects are measured repeatedly over time. With this in mind, a simple linear regression can not be used due to the assumption of independent observations.

Linear mixed-effects models were created with the idea that each individual has its own subject-specific mean response profile over time. In the linear mixed-effects model extension (Harville, 1977; Laird and Ware, 1982; Verbeke and Molenberghs, 2000), the repeated measurements are fitted by using a linear regression model, where parameters vary over individuals. The general form is,

$$\begin{cases} y_i &= X_i\beta + Z_i b_i + \epsilon_i, \\ b_i &\sim N(0, D), \\ \epsilon_i &\sim N(0, \sigma^2 I_{n_i}), \end{cases} \tag{2.1}$$

where X_i and Z_i are the design matrices corresponding to the fixed and the random effects respectively and I_{n_i} is the order n_i identity matrix, where n_i denotes the number of observations in the ith subject (cluster), $i = 1, \ldots, n$. In addition, β is the fixed vector and b_i denotes the random effects coefficient. These random effects are assumed to be normally distributed, with mean zero and variance-covariance matrix D. Moreover, b_i are assumed to be independent of the error terms ϵ_i, i.e., $cov(b_i, \epsilon_i) = 0$. Equivalently,

$$\begin{cases} y &= X\beta + Zb + \epsilon, \\ b &\sim N(0, D), \\ \epsilon &\sim N(0, R). \end{cases} \tag{2.2}$$

Note that X is a $n \times p$ matrix, with p the number of fixed effects, and the fixed vector coefficients $\beta_j, j = 1, \ldots p$ denote the change in the average y_i when the corresponding covariate x_j is increased by one unit, while all other predictors are held constant. As Z is a $n \times k$ matrix, with k the number of random effects, and b_i represents how specific regression parameters of the ith subject deviates from those in the population. Moreover, $R = \sigma^2 I_{n \times n_i}$, with $I_{n \times n_i}$ denotes a $(n \times n_i)$-dimensional identity matrix.

Besides the above-mentioned advantage of using mixed models, in order to analyse observations with certain hierarchy, this methodology has several desirable features among which we highlight the following:

- Apart from describing how the mean response changes in the population of interest, it is possible to predict how individuals response trajectories change over time.

- There is no requirement of balanced data, that is, we do not require the same number of measurements on each subject, neither do they need to be taken at the same set of times.

Covariance matrix V

Marginally the covariance of y, $var(y) = V$, can be written as;

$$var(y) = var(X\beta + Zb + \epsilon)$$

Assuming that the random effects and the residuals are uncorrelated,

$$var(y) = var(X\beta) + var(Zb) + var(\epsilon).$$

Due to fact that β describes the fixed effects parameters, $var(X\beta) = 0$ and Z is a matrix of constants, therefore the covariance matrix is given by,

$$var(y) = V = Zvar(b)Z' + var(\epsilon) = ZDZ' + R$$

where D is the variance-covariance matrix and $var(\epsilon) = R = \sigma^2 I_{n\times n_i}$.

Random Intercepts and Random Intercepts and Slopes

By outlining, we can already distinguish two kinds of mixed models. The random intercepts model allows intercepts variation across groups. For a better understanding, the subject-specific fitted models are parallel to the average, that is the population fitted model.

In particular, a basic example of a random intercepts model was included, in order to illustrate the model fitting, which is formed by two clearly distinct parts,

$$y_i = \beta_0 + \beta_1 x_{ij} + b_{0i} + \epsilon_i$$

these are, a fixed part (which is the intercept and the coefficient of the explanatory variable times the explanatory variable) and a random part. The random part is composed of two random terms, just like the variance components model, on the one hand a variance of the level 1 random term $e_{ij} \sim N(0, \sigma^2)$ and, on the other hand a variance of the level 2 random term $b_i \sim N(0, \sigma_b^2)$. Accordingly in this case, in the mixed model formulation (2.1) the design matrices are replaced by,

$$X_i = \begin{bmatrix} 1_1 & x_{i1} \\ \vdots & \vdots \\ 1_{n_i} & x_{in_i} \end{bmatrix}, \quad Z_i = \begin{bmatrix} 1_1 \\ \vdots \\ 1_{n_i} \end{bmatrix}, \quad \beta = [\beta_0, \beta_1]^T.$$

and the random effects model covariance structure,

$$b_i \sim N(0, D_i), \text{ with } D_i = \sigma_b^2.$$

Following up with an intuitive extension, that also allows a random shift in the subject-specific slopes, known as random intercepts and random slopes model. In this case, our example will take the form,

$$y_i = \beta_0 + \beta_1 x_{ij} + b_{0i} + b_{1i} x_{ij} + \epsilon_i.$$

In this model we additionally have b_{1i} which represents the random slope effect of the coefficient x_{ij}, so actually two extra parameters should be estimated, the variance in intercepts between groups $\sigma^2_{b_0}$ and the variance in slopes between groups $\sigma^2_{b_1}$. In this case the model matrix Z_i has the form,

$$Z_i = \begin{bmatrix} 1_1 & x_{i1} \\ \vdots & \vdots \\ 1_{n_i} & x_{in_i} \end{bmatrix},$$

and the random effects model covariance structure,

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N(0, D_i), \text{ with } D_i = \begin{pmatrix} \sigma^2_{b_0} & \sigma_{b_0 b_1} \\ \sigma_{b_0 b_1} & \sigma^2_{b_1} \end{pmatrix}$$

where $\sigma_{b_0 b_1}$ denotes the covariance between the intercepts and slopes.

Estimation

Given that the ith subject outcomes have the same random effects they will be marginally correlated, so we assume that

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

i.e., longitudinal responses of a subject are independent conditionally on its random effect. It makes sense for using the marginal model

$$y_i = X_i\beta + \epsilon_i^*, \text{ where } \epsilon_i^* = Z_i b_i + \epsilon_i,$$

and with $cov(\epsilon_i^*) = V_i = Z_i D Z_i^T + \sigma^2 I_{n_i}$, to estimate the parameters of linear mixed-effects models. Using maximum likelihood (ML) principles, the observations of the ith subjects are not independent so the likelihood function needs to be a multivariate normal distribution for y_i. As random effects have expected values of zero and therefore do not affect the mean, this distribution has a mean vector $X_i\beta$ and a covariance matrix V_i, then

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

where $\theta^T = (\beta, V)$.

Taking into account that we assume independence across subjects, the likelihood function is simply the product of the density functions for each subject. The log-likelihood of a linear mixed model is given by:

$$l(\theta) = \sum_{i=1}^{n} \log p(y_i;\theta),$$

Given V_i, the estimates of fixed-effects parameters are obtained by maximizing the log-likelihood function (2.1.2), conditionally on the parameters in V_i, and have a closed-form solution:

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

Nevertheless, in the case of the random effects, we can not really speak about estimation but instead we can talk about prediction, because of them being random variables. One way to obtain the best linear unbiased predictor is by Henderson's mixed model equations, which, in turn, allow us to obtain the best linear unbiased estimator of $X\beta$,

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + D^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix},$$

and the obtained solutions are

$$\begin{aligned} \hat{\beta} &= (X'V^{-1}X)^{-1}X'V^{-1}y, \\ \hat{u} &= DZ'V^{-1}(y - X\hat{\beta}), \end{aligned}$$

depending on the variance parameters of $V = ZDZ' + R$.

Therefore, the next step is to estimate the parameters of the covariance matrix V. There are two common ways of estimation: the maximum likelihood (ML) and the restricted maximum likelihood (REML).

Employing the maximum likelihood method to obtain the ML estimators of the V parameters for a given value of β, will be biased for small samples. This bias arises because the ML estimate has not taken into account that β is estimated from the data as well. In contrast, the REML estimates the variance components based on the residuals obtained after the fixed effects were estimated, this is, $y - X\beta$, reason for why it is referred as marginal likelihood. Then, if the sample size is small, the REML would yield better estimates than the ML. It is worth noting that neither the ML nor the REML for the parameters in V can be written, in general, in closed form, so it is necessary to approximate them numerically. Two algorithms were implemented for linear mixed-effects models (Lindstrom and Bates, 1988) such as Expectation-Maximization and Newton-Raphson algorithms.

2.2 Survival Analysis

Survival analysis is a powerful tool for studies aimed at analysing event times. In particular, clinical follow-up studies may be interested in analysing the time until an event occurs, normally understood as death or contracting a disease. In these procedures, the factors or covariates effects on survival or risk are also studied.

In this way, the variable of interest, that is, the dependent variable, is the time until that event, namely *failure time*, *event time*, *survival time*. The presence of censoring in survival data is what makes the difference, and consequently requiring specific methodologies, such as survival analysis. To clarify, censored data is defined to be those data in which the event time of interest is not fully observed on all subjects under study (lost to follow up or drop out of the study, or if the study ends before they die or have an outcome of interest). To describe this censoring mechanism we refer to two possible classifications. First, regarding to the position of the observation of the time to the event, is either left- or right-censored (the survival time is less or greater than the observation time) and interval-censored data (in which the time to the event of interest is

known to occur between two certain time points). And secondly, differentiating between informative censoring, that occurs when subjects are lost to follow-up due to reasons related to the study, and noninformative censoring, when individuals drop out of the study for reasons unrelated to the study, but it can depend on covariates.

In our application data, the date of death is known for all patients finally included in our study, in case of an event prior to the end of study. There have been no drop outs and all censures are caused by the end of study. Given that, in order to conduct the study, we will consider noninformative right censoring.

2.2.1 Functions of interest

Let T be the survival time and C the censoring time. Define the follow-up time $Y = \min(T, C)$, and let $\delta = 1(T \leq C)$ denote the censoring indicator. Considering the probability density function of T, $f(t) = P(T = t)$, which represents the unconditional probabilities of death, the survival function is defined as the complement of the cumulative distribution function,

$$S(t) = P(T \geq t) = \int_t^\infty f(x)dx$$

giving the probability that the event has occurred by duration t.

An alternative characterization of the distribution of T is the hazard function. It describes the instantaneous risk for an event in the time interval $[t, t + dt]$ given that it has not occurred before, and is given by the following expression

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt},\ t > 0.$$

Due to fact that the hazard function characterizes the probability density function, $f(t) = \lambda(t)\Pi_{k=1}^{t-1}[1-\lambda(k)]$, the survival function can be expressed as a product of hazards,

$$S(t) = \frac{f(t)}{\lambda(t)} = \prod_{k=1}^{t-1}[1 - \lambda(k)]. \tag{2.3}$$

or as

$$S(t) = \exp\{-\Lambda(t)\} = \exp\left\{-\int_0^t \lambda(s)ds\right\} \tag{2.4}$$

where $\Lambda(\cdot)$ is the cumulative risk function, which describes the accumulated risk up until time t.

2.2.2 Survival estimation

Up to this point, results show that survival and hazard functions are equivalent alternatives to characterize the distribution of T. We now introduce the most well-known estimators of both functions.

- *Kaplan-Meier estimator*

  Let $T_i,\ i = 1, \dots n$ be the observations of T and, therefore, C_i the respective censoring time for subject i. In order to estimate the survival function, the estimator proposed by Kaplan and Meier (1958) takes into account for censoring by adjusting the number of subjects at risk,

$$\hat{S}_{KM}(t) = \prod_{i:t_{(i)} \leq t} \left[1 - \frac{d_i}{n_i}\right],$$

  where $t_{(i)}$ denote the distinct ordered times of death and, d_i and n_i denote the number of events and the number of individuals still at risk at time t_i respectively. The Kaplan-Meier estimator is a step function with discontinuities or jumps at the observed event times, coinciding with the empirical survival function if there is no censoring.

  A special mention needs to be made of the Kaplan-Meier statistical properties, its consistency has been proved by Peterson (1977), and Breslow and Crowley (1974) have shown that $\sqrt{n}(\hat{S}(t) - S(t))$ converges in law to a Gaussian process with expectation 0 and variance-covariance function that may be approximated using Greenwood's formula,

$$var(\hat{S}(t_{(i)})) = [\hat{S}(t_{(i)}]^2 \sum_{j=1}^{i} \frac{1 - \hat{\pi}_j}{n_j \hat{\pi}_j},$$

  where $\hat{\pi}_j = 1 - d_j/n_j$.

- *Nelson-Aalen Estimator*

  The Nelson-Aalen estimator could be thought as an alternative to estimate the cumulative hazard, it is given by the following expression:

$$\hat{\Lambda}(t_{(i)}) = \sum_{j=1}^{i} \frac{d_j}{n_j},$$

  in which the hazard must be intuitively interpreted as the ratio of the number of deaths to the number exposed. The estimator variance can also be approximated by using the Greenwood's formula. In such situation, Breslow (1972) suggested estimating the survival function as

$$\hat{S}(t) = \exp\{-\hat{\Lambda}(t)\}.$$

Fleming and Harrington (1984) showed the close relationship between both estimators, especially when the number of events is small relative to the number exposed.

2.2.3 Parametric Maximum Likelihood

Apart from using non-parametric estimators like the Kaplan-Meier and the Nelson-Aalen estimators, we can also assume a parametric form for the distribution of the survival time, $S(t)$, and then, the parameter estimation will be made by using maximum likelihood. Let the sub-index i refer to the subject indicator and, consequently, $\{Y_i, \delta_i\},\ i = 1, \ldots, n$ denote their survival information. Taking a random sample from a certain distribution, parameterized by θ, the likelihood funtion is given by,

$$l(\theta) = \prod_{i=1}^{n} f(Y_i; \theta)^{\delta_i} S_i(Y_i; \theta)^{(1-\delta_i)}.$$

Note that it takes to account for censoring information, by contributing with $f(T_i; \theta)$ when an event is observed at time T_i and with $S(T_i; \theta)$ when subjects survived up to that point, that is $T_i > Y_i = C_i$. This can be rewritten in terms of hazard function using the relations (2.3) and (2.4):

$$l(\theta) = \prod_{i=1}^{n} \lambda(Y_i; \theta)^{\delta_i} \exp\{-\Lambda(t)\}^{(1-\delta_i)}. \tag{2.5}$$

To address this issue, iterative optimatization procedures could be necessary to locate the maximum likelihood estimates $\hat{\theta}$, such as the Newton-Raphson algorithm (Lange, 2004).

2.2.4 Regression Methods

There are several ways to relate the outcome to predictors in survival analysis. We will focus on two, namely, the proportional hazards model, that is also known as relative risk model, and the accelerated failure time model.

- **Relative Risk Model**

  The best known procedure in survival analysis for modelling the relationship of covariates to a survival or other censored outcome is the Cox model (Cox, 1972), formulated as,

  $$\lambda_i(t|w_i) = \lambda_0(t) \exp(\gamma^T w_i), \tag{2.6}$$

  where λ_0 is an unspecified function of time called the baseline hazard, $w_i^T = (w_{i1}, \ldots, w_{ip})$ denotes the covariate vector for subject i and γ is a $p \times 1$ column vector of coefficients. In particular, $\exp(\gamma_j)$ denotes the ratio of hazards for one unit change in the j-th covariate at any time t. The model assumes that covariates have a multiplicative effect on the hazard for an event.

  Because the hazard ratio for two subjects with fixed covariates vectors w_i and w_j is constant over time, as can be proved,

  $$\frac{\lambda_i(t|w_i)}{\lambda_j(t|w_j)} = \frac{\lambda_0(t) \exp(w_i\gamma)}{\lambda_0(t) \exp(w_j\gamma)} = \exp\{\gamma^T(w_i - w_k)\} \tag{2.7}$$

  the model is also known as the proportional hazards, relative risk or relative hazard model.

Different types of estimation of the Cox model are possible following approaches based on parametric or semiparametric modelling. In the first case, we assume a model under the baseline hazard function, so the parameters' estimation is developed by maximizing the likelihood function (2.5). Nevertheless, the semiparametric modelling arises to avoid this baseline hazard' specification introduced by Cox (1972), showing that the estimation of γ can be based on the partial likelihood,

$$pl(\gamma) = \prod_{i=1}^{n} \left[\frac{\gamma^T w_i}{\sum_{i=1}^{n} I(Y_j \geq Y_i) \exp(\gamma^T w_j)} \right]^{\delta_i},$$

that is not to specify the distribution of T_i. For more detail of the efficiency of the Cox Model estimator, we refer to Kalbfleisch and Prentice (2002, chap. 5).

Proportional Hazard assumption and Time-Dependent Covariates

As Kleinbaum and Klein (2005) say

> *...the proportional assumption requires that the hazard ratio is constant over time, or equivalently, that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time.*

It is enough to point out at the expression (2.7), where the baseline hazard has cancelled out. However, in follow-up studies, where covariates depend on time, the Cox model will be inappropriate, due to these situations yield a hazard ratio that varies across time. In this way, if time-dependent covariates are considered, the Cox model form may still be used, but not satisfying the assumption (2.7). This adjustment is called the extended Cox model.

A time-dependent variable is defined as any variable whose value for a given subject may differ over time. We can distinguish two different categories of time-dependent covariates, namely *external* and *internal* covariates. In particular, a variable is called an external variable if its value changes because of "external" characteristics and affects several individuals simultaneously. In contrast, an internal variable change is due to "internal" characteristics or behaviour specific to the individual, typically arises as time-dependent measurements taken on the subjects under study. The most important characteristic of this type of variables is that they typically require the survival of the subject for their existence. In other words, a failure of the subject at time s corresponds to non-existence of the covariate at $t \geq s$.

Extended Cox Model

In this context, supposed to be both time-independent and time-dependent covariates, the extended Cox model is written as

$$\lambda_i(t|\mathcal{Y}_i(t), w_i) = \lambda_0(t) \exp\{\gamma^T w_i + \alpha y_i(t)\},$$

where $\mathcal{Y}_i(t) = \{y_i(s), 0 \leq s < t\}$, in which $y_i(t)$ denotes a vector of time-dependent covariates. These time-dependent covariates are encoded using the (start,stop] notation, that hold the information of the specific time intervals in which longitudinal measurements are recorded. As defined in (2.6), w_i denotes a vector of baseline covariates. The regression coefficients vector α is similar as for γ. Thus, assuming there is only a single time-dependent covariate, $\exp(\alpha)$ denotes the relative increase in the risk for an event at time t that results from one unit increase in $y_i(t)$ at this point.

The primary reason for distinguishing among defined, internal or external variables is that the extended Cox model has strong assumptions. On the one hand, the time-dependent variables are external, so they are not related to the failure mechanism (Kalbfleisch and Prentice, 2002). This is an unrealistic assumption for the longitudinal process, especially, for follow-up studies. Moreover, the value of each covariate would be known at every failure time for all subjects, a problem in unbalanced data. On the other hand, an unacceptable condition in internal variables is that the extended Cox model will not be able to take into account the measurement error of the longitudinal covariates, and thus can introduce bias.

- **Accelerated Failure Time Model**

Another alternative modelling framework for event time data is the accelerated failure time (AFT) models. These models specify that predictors act multiplicatively on the failure time (additively on the log of the failure time). The predictors alter the rate at which a subject proceeds along the time axis, i.e., they accelerate or decelerate the time of failure (Kalbfleisch and Prentice, 2002). The model is defined as,

$$log(T_i) = \gamma^T w_i + \sigma_t \epsilon_{ti}$$

where parameter γ_t is a scale parameter and ϵ_{ti} is assumed to follow a specific distribution. Then, the parameter γ_j denotes the change in the expected log failure time for a unit change in the corresponding covariate w_{ij}. Equivalently, a unit change in w_{ij} increases the failure time by a factor of $\exp(\gamma_j)$. Then, in terms of the risk rate function, we can postulate the accelerated failure time as,

$$\lambda_i(t|\mathcal{Y}_i(t), w_i) = \lambda_0(V_i(t)) \exp\{\gamma^T w_i + \alpha y_i(t)\},$$

with $V_i(t) = \int_0^t \exp\{\gamma^t w_i + \alpha y_i(s)\} ds$. For more information we refer to Cox and Oakes (1984).

In both Proportional Hazards and Accelerated Failure Time models an unspecified baseline risk function $\lambda_0(\cdot)$ is used, that can be assumed of a specific parametric form or modelled flexibly. We want to point out that the Weibull distribution (and consequently its special case, the exponential distribution) is the only distribution that can be used to describe both PH and AFT models.

Chapter 3

Joint Modelling for Longitudinal and Time-to-Event Data

Many clinical and epidemiologic studies, generate both longitudinal (repeated measurements) and survival (time-to-event) data. Up to now, well-established methods have been introduced to study the longitudinal process and the survival process separately. However, these may be inappropriate when the longitudinal variable is correlated with the survival process, either with the subject's status as well as the possibility of study dropout.

As mention at the end of the section 2.2, a possibility to study the association between longitudinal measurements and survival process is the extended Cox model. But it is not appropriate, especially for internal time-dependent covariates because it can result in biased estimations.

A possible alternative developed by Self and Pawitan (1992) is the use of the two-stage approach. Despite reducing any bias by using a survival model that incorporates a longitudinal covariate that has been measured with error, this is not an unbiased approach. No survival information is used to determine longitudinal estimates, so informative drop-out is not accounted for, causing biased estimates.

The joint likelihood approach arises to alleviate the potential bias caused by the time-dependent Cox model and two-stage approach. This is done by taking into account the association between the survival and the longitudinal processes by simultaneously determining the parameter estimates for both processes. In the literature we can find several types of joint approaches depending on the parametrization of the joint likelihood of the longitudinal and survival processes. An overview of the development of joint models is made by Tsiatis and Davidian (2004). The authors focus on models for the longitudinal process and the hazard for the time-to-event that depend jointly on shared, underlying random effects. As mentioned in this article, it has been demonstrated that these models lead to correction of potential biases for enhanced efficiency.

In the literature, two different proposals of joint approaches which used shared random effects to their parametrization were found. The difference between them is the research interest. Rizopoulos (2010) has proposed a joint model where the time-to-event process is of main interest and influenced by a longitudinal time-dependent covariate

measured with error. Philipson et al. (2012) developed a shared random effects model where the focus is on both survival and longitudinal processes.

In this work, our main goal is to study the patients' survival. This is the reason why, in this chapter, we consider the joint approach proposed by Rizopoulos (2010), due to the fact that this model is focused on the survival process.

First, the submodels are specified, particularly with regard to the Rizopoulos's proposal. Then, in the next section, the maximum likelihood estimation of the joint model's parameter is discussed. A brief summary of inference and the diagnosis for the joint model approach is presented in the subsequent sections. Following up with a short exposition of how to provide dynamic predictions, that is, predictions updated utilizing the new information recorded for each subject. Finally, the main ideas of using the JM Package (Rizopoulos, 2010) are exposed in the next section, synthesising the computational problems already presented by the author.

3.1 Submodels specification

The joint model consists of two linked submodels, known as the longitudinal submodel, and the survival submodel. To introduce this methodology we will use the same notation as in Chapter 2, although with minor changes, this is schematically presented below to keep it in mind,

- Let T_i be the event time, C_i the censoring time and $\delta_i = 1(T_i \leq C_i)$ the event indicator for the ith subject.
- Let $y_i(t)$ be the observed value of the time-dependent covariate at time point t, equivalently, $y_{ij} = \{y_i(t_{ij}), j = 1, \ldots, n_i\}$. Thus, $m_i(t)$ denote the *true* and unobserved value of the respective longitudinal outcome at time t, uncontaminated with the measurement error value of the longitudinal outcome so it is different from $y_i(t)$.

3.1.1 The Survival Submodel

Our aim is to associate the *true* and unobserved value of the longitudinal outcome at time t, $m_i(t)$, with the risk for an event T_i, as stated in section 2.2,the relative risk model can be written as,

$$\lambda_i(t|\mathcal{M}_i(t), w_i) = \lambda_0(t) \exp\{\gamma^T w_i + \alpha m_i(t)\},\ t > 0, \tag{3.1}$$

where $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ denotes the history of the true (unobserved) longitudinal process up to time t. Let $\lambda_0(\cdot)$ denote the baseline risk function and w_i the vector of baseline covariates. The interpretation of the regression coefficients is exactly the same,

- $\exp(\gamma_j)$ denotes the ratio of hazards for one unit change in the j-th covariate at any time t.
- $\exp(\alpha)$ denotes the relative increase in the risk for an event at time t that results from one unit increase in $m_i(t)$ at this point.

In the expression (3.1) we can note that it depends only on a single value of the time-dependent marker $m_i(t)$. To take into account the whole covariate history $\mathcal{M}_i(t)$ to determine the survival function, the relation (2.4) can be used to obtain,

$$\begin{aligned} S_i(t|\mathcal{M}_i(t), w_i) &= P(T_i > t|\mathcal{M}_i(t), w_i) \\ &= \exp\left(-\int_0^t \lambda_0(s)\exp\{\gamma^T w_i + \alpha m_i(s)\}ds\right) \end{aligned} \tag{3.2}$$

Reminding again that both are written as a function of a baseline hazard $\lambda_0(t)$. Regardless of the fact that the literature recommends to leave $\lambda_0(\cdot)$ completely unspecified, in order to avoid the impact of misspecifying the distribution of survival times, in the joint modelling framework it can lead to an underestimation of the standard error of the parameter estimates (Hsieh et al., 2006). There are several options to use a risk function corresponding to a known parametric distribution, such as

- The *Weibull model*, let Y follow a Weibull distribution with parameters t and p, $Y \sim W(\lambda, p)$, the hazard is obtained as,

  $$\lambda(t) = \lambda p(\lambda t)^{p-1}$$

  where, if $p > 1$ indicates that the failure rate increases with time; decreasing if $p < 1$, and constant over time if $p = 1$, called also *exponential model*.

But it is more desirable to flexibly model the baseline risk function. Among the proposals encountered, we would like to highlight those that follow,

- The *piecewise-constant model*, where the baseline risk function takes the form:

  $$\lambda_0(t) = \sum_{q=1}^{Q} \xi_q I(\nu_{q-1} < t \leq \nu_q),$$

  where $0 = \nu_0 < \nu_1 < \ldots < \nu_Q$ denotes a partition of the time scale, with ν_Q being lager than the largest observed time, and ν_q denotes the value of the hazard in the interval $(v_{q-1}, v_q]$.

- The *regression splines model*, where the log baseline risk function $\log \lambda_0(t)$ is given by,

  $$\log \lambda_0(t) = \kappa_0 + \sum_{d=1}^{m} \kappa_d B_d(t, q),$$

  where $\kappa^T = (\kappa_0, \kappa_1, \ldots, \kappa_m)$ are the spline coefficients, q denotes the degree of the B-splines basis functions $B(\cdot)$, and $m = \ddot{m} + q - 1$, with $\ddot{m}$ denoting the number of interior knots.

In both models, the specification of the baseline hazard becomes more flexible as the number of knots increases. In particular, in the limiting case of the piecewise-constant model where each interval contains only a single true event time, this model is equivalent to leaving $\lambda_0(\cdot)$ unspecified and estimating it using nonparametric maximum likelihood. In both approaches, we should keep a balance between bias and variance and avoid overfitting. Although there is not an ideal strategy, Harrel (2001) gives a standard

rule of thumb based on keeping the total number of parameters (included in the linear predictor and in the model for h_0(t)), between 1/10 and 1/20 of the number of events in the sample. Therefore, the knots' position will be based on percentiles of the event times.

3.1.2 The Longitudinal Submodel

In the above definition of the survival model 3.1 we used the true unobserved value of the longitudinal covariate $m_i(t)$. Taking into account that the longitudinal information $y_i(t)$ is collected with possible measurement errors, the first step towards measuring the effect of the longitudinal covariate to the risk for an event is to estimate $m_i(t)$, in order to reconstruct the complete *true* history $\mathcal{M}_i(t)$ to each subject. Then, the linear mixed model can be rewritten as,

$$\begin{cases} y_i(t) & = & m_i(t) + u_i(t) + \epsilon_i(t), \\ m_i & = & x_i^T(t)\beta + z_i^T(t)b_i, \\ b_i & \sim & N(0, D), \\ \epsilon_i & \sim & N(0, \sigma^2 I_{n_i}). \end{cases}$$

This mixed model formulation allows to state that the longitudinal outcome $y_i(t)$ is equal to the true level $m_i(t)$ plus an error term. The main difference from the model (2.1) is that, in addition to the random error term $\epsilon_i(t)$, we could incorporate an additional stochastic term $u_i(t)$. This last term is used to capture the remaining serial correlation in the observed measurements, which random effects are unable to capture. Considering that $u_i(t)$ is considered as a mean-zero stochastic process, independent of b_i and $\epsilon_i(t)$.

3.2 Estimation

In the previous chapter the estimation of the parameters has been based on the maximum likelihood approach for both processes. Rizopoulos (2012b) has also used the likelihood method for joint models, as perhaps the most commonly used approach in the joint literature.

In this section, we first describe the joint likelihood process in order to estimate the joint model's parameters. It is followed by a brief presentation of how to estimate the random effects in joint modelling.

3.2.1 Joint Likelihood Formulation

The likelihood method for joint models is based on the maximization of the log-likelihood of the joint distribution of the time-to-event and longitudinal data $\{Y_i, \delta_i, y_i\}$.

Let the vector of time-independent random effects, b_i, account for the association between the longitudinal and the event process, and the correlation between the repeated measurements in the longitudinal outcome. Strictly, we have that the longitudinal process and the survival process are conditionally independent given b_i,

$$p(Y_i, \delta_i, y_i|b_i; \theta) = p(Y_i, \delta_i|b_i; \theta)p(y_i|b_i; \theta)$$

with

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

where $\theta = (\theta_t^T, \theta_y^T, \theta_b^T)^T$ denotes the full parameter vector for the event time outcome, the longitudinal outcomes and for the random-effects covariance matrix respectively.

Under the modelling assumptions presented in the previous section and these above conditional independence assumptions, the joint log-likelihood contribution for the i-th subject has the form,

$$\begin{aligned}\log p(Y_i, \delta_i, y_i;\theta) &= \log \int p(Y_i, \delta_i, y_i, b_i;\theta) db_i \\ &= \log \int p(Y_i, \delta_i|b_i;\theta_t, \beta) \left[\prod_j p\{y_i(t_{ij})|b_i;\theta_y\}\right] p(b_i;\theta_b) db_i\end{aligned} \tag{3.3}$$

where the likelihood of the survival part takes the form

$$p(Y_i, \delta_i|b_i;\theta_t, \beta) = \{\lambda_i(Y_i|\mathcal{M}_i(Y_i);\theta)\}^{\delta_i} S_i(Y_i|\mathcal{M}_i(Y_i);\theta)$$

with $\lambda_i(\cdot)$ and $S_i(\cdot)$ obtained by (3.1) and (3.2). On the other hand, the joint density for longitudinal responses together with the random effects is performed through the following expression,

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

where q_b denotes the dimensionality of the random-effects vector, and $\|\cdot\|$ denotes the Euclidean vector norm.

Then, the (overall) log-likelihood for all the observed data is formulated as,

$$l(\theta) = \sum_i \log p(Y_i, \delta_i, y_i;\theta). \tag{3.4}$$

The maximization of this function (3.4) with respect to θ requires a combination of numerical integration and optimization algorithms, because both the integral with respect to the random effects in (3.3) and in the survival function given by (3.2) do not have an analytical solution. Despite some authors have employed standard numerical integration techniques, such as Monte Carlo or Gaussian quadrature, the Expectation-Maximization (EM) algorithm described by Wulfsohn and Tsiatis (1997) has been traditionally preferred. The intuitive idea behind the EM algorithm is to maximize the log-likelihood in two steps: the Expectation step, where missing data are filled, so we replace the log-likelihood of the observed data with a surrogate function, and the Maximization step, where this surrogate function is then maximized.

Furthermore Rizopoulos et al. (2009) has introduced a direct maximization of the observed data log-likelihood which is a quasi Newton algorithm. Therefore hybrid optimization approaches start with EM and then continue with direct maximization.

- *Optimization Control*

  To control the optimization process, the EM algorithm starts with a fixed number of iterations, and if convergence is not achieved, it switches to a quasi-Newton algorithm until convergence is obtained. The following two criteria are used to declare convergence,

$$\max\{|\theta^{(it)} - \theta^{(it-1)}|/(|\theta^{(it-1)}| + \epsilon_1)\} < \epsilon_2,$$
$$l(\theta^{(it)}) - l(\theta^{(it-1)}) < \epsilon_3\{|l(\theta^{(it-1)})| + \epsilon_3\},$$

  where $\theta^{(it)}$ denotes the parameters values at the ith iteration. In addition, the values for ϵ_1, ϵ_2 that are frequently used are about 10^{-3} or 10^{-4}, and for ϵ_3 it is about 10^{-8}.

- *Numerical Integration*

  As mentioned before, a numerical approach is necessary to approximate the integrals of the survival function (3.2), as well as the integral with respect to the random effects (3.3), the latter becoming more computationally demanding as its dimensionality increases.

  In addition to the possibility of using the Gauss-Hermite (GH) quadrature to approximate these integrals' solutions, Rizopoulos (2012a) proposed an alternative approach, called the adaptive Gauss-Hermite (aGH) rule, that decreases the computational burden to some degree. For more details regarding the specification of the approximation rules we refer to Rizopoulos (2012b, Sec. 4.3.5).

3.2.2 Estimation of the Random Effects

The estimation of the random effects presented in Rizopoulos (2012b) is based on Bayes theory. Assuming that $p(b_i;\theta)$ is the prior distribution, and that $p(b_i|Y_i,\delta_i,y_i;\theta)p(y_i|b_i;\theta)$ is the conditional likelihood part, the corresponding posterior distribution is,

$$\begin{aligned} p(b_i|Y_i,\delta_i,y_i;\theta) &= \frac{p(Y_i,\delta_i|b_i,\theta)p(y_i|b_i;\theta)p(b_i;\theta)}{p(Y_i,\delta_i,y_i;\theta)} \\ &\propto p(Y_i,\delta_i|b_i,\theta)p(y_i|b_i;\theta)p(b_i;\theta) \end{aligned}$$

it does not have a closed form solution so it has to be numerically computed. Two types of estimators typically used are,

$$\begin{cases} \overline{b}_i &= \int b_i p(b_i|Y_i,\delta_i,y_i;\theta), \text{ and} \\ \hat{b}_i &= \arg\max_b\{\log p(b_i|Y_i,\delta_i,y_i;\theta)\}, \end{cases}$$

that is, mean versus mode.

3.3 Model testing

It is shown in section 3.2.1 that the parameters in the joint models can be estimated by maximum likelihood. Apart from the likelihood ratio procedure for model testing, Rizopoulos explained that there are other alternatives to test the null hypothesis

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_a : \theta \neq \theta_0,$$

presented below,

- *Likelihood Ratio Test*, defined as

$$LRT = -2\{l(\hat{\theta}_0 - l(\hat{theta}))\},$$

  where $\hat{\theta}_0$ and $\hat{\theta}$ denote the maximum likelihood estimates under the null and alternative hypothesis, respectively.

- *Score Test*, defined as

$$U = \mathcal{S}^T(\hat{\theta}_0)\{\mathcal{I}(\hat{\theta}_0)\}^{-1}\mathcal{S}(\hat{\theta}_0), \quad \text{with } \mathcal{I}(\hat{\theta}) = -\sum_{i=1}^{n} \frac{\partial S_i(\theta)}{\partial \theta}\bigg|_{\theta=\hat{\theta}},$$

  where $\mathcal{S}(\cdot)$ and $\mathcal{I}(\cdot)$ denote the score function and the observed information matrix of the model under the alternative hypothesis.

- *Wald Test*, defined as

$$W = (\hat{\theta} - \theta_0)^T \mathcal{I}(\hat{\theta})(\hat{\theta} - \theta_0).$$

Under the null hypothesis, they are asymptotically χ^2_p-distributed, with p denoting the number of parameters being tested. In particular, the Wald test for a single parameter θ_j is equivalent to $(\hat{\theta}_j - \theta_{0j})/\hat{s.e.}(\hat{\theta}_j)$, which under the null hypothesis follows an asymptotic standard normal distribution.

Despite of being asymptotically equivalent, the behaviour of the tests is different in finite samples. The election of any of these procedures depends on the limitations of each one. Specifically, regarding the computational cost of fitting, the Wald test only requires to fit the model under the null hypothesis, and the score test under the alternative. However, the likelihood ratio test requires to fit the model under both hypotheses, being more computationally expensive. But other issues must be considered, such as the Wald test that does not take into account the variability introduced by estimating the variance components, apart from ignoring the fact that we need to estimate the survival process. Also, the implementation of the score test needs extra steps to calculate the required components.

A general drawback of these tests is that they are only appropriate for the comparison of two nested models. In order to carry out the comparison of non-nested models, information criteria could be used, such as the Akaike's Information Criterion (AIC; Akaike (1974)), and the Bayesian Information Criterion (BIC; Schwarz (1978)), defined as,

$$AIC = -2(\hat{\theta}) + 2n_{par},$$
$$BIC = -2(\hat{\theta}) + n_{par}\log(n),$$

where n_{par} denotes the number of parameters in the model.

Apart from these topic procedures to models' comparison, we could be also interested in testing whether an extra random effect should be included in the joint model. However, this specific field is forgotten in the joint modelling framework, so it could be an interesting future line of research.

3.4 Joint Model Diagnostics

The previous sections have provided a guide to learn how to formulate the joint model to optimally study the relationship between longitudinal and time-to-event data. But, as ever, after fitting a regression model it is important to determine whether all the necessary model assumptions are valid before performing inference.

A standard tool to perform model diagnostics are residual graphical methods, as residual plots, and formal statistical tests. Despite of it being intensively studied for longitudinal and survival analysis, this topic has not received special attention in the joint modelling literature. It is noteworthy that Rizopoulos et al. (2010) have developed a multiple imputation residuals tool to asses these joint model's assumptions.

The diagnostic plots to check the fit of mixed models and relative risk models can be used to construct diagnostic plots to inspect the fit of joint models. As explained in the mentioned paper, a problem is that in the joint modelling framework, it is assumed that the occurrence of events is related with the underlying evolution of the subject-specific longitudinal profiles, which corresponds to a non-random dropout mechanism (i.e. missing not at random mechanism, MNAR). The implication of the non-random nature of the dropout mechanism is that the observed data, upon which the residuals are calculated, do not constitute a random sample of the target population, so the residuals are not expected to exhibit standard properties, such as zero mean and independence.

To overcome these problems Rizopoulos et al. (2010) proposed a new method for calculating residuals and producing diagnostic plots in joint models, based on creating random versions of the completed data set by multiple imputation of the missing longitudinal responses under the fitted joint model. They proposed two different procedures depending on which type of visit times we are considering in the study: fixed or random visit times.

The graphical residual analysis allows, as is always the case, to check possible trends. We refer to Rizopoulos et al. (2010) for technical details.

3.5 Dynamic Predictions

In joint modelling approaches the objective is to study the association between the survival process and longitudinal outcomes. Such models can be used to provide predictions for the survival and longitudinal outcomes. Rizopoulos (2011) proposed to provide predictions of a joint model with a dynamic nature. This dynamic nature comes from updating the prediction utilizing new information recorded for the patient as time progresses. That is, considering the effect of repeated measures taken in time t to the survival up to time t. Thus, the conditional probability is of primary interest, described as,

$$\pi_i(u|t) = P(T_i^* \geq u / T_i^* > t, \mathcal{Y}_i(t), \omega_i, D_n), \quad t > 0,$$

where u is the followed-up time ($u > t$), D_n denotes the sample on which joint model was fitted. The author uses a Bayesian formulation of the problem and Monte Carlo estimates of $\pi_i(u/t)$, for more details we refer to the article Rizopoulos (2011), as well as to Rizopoulos (2012b, chap.7).

3.6 JM Package

Although software capable of fitting joint models has recently been developed, we find different approaches to model specification across software packages. The available procedures in the statistical software packages R and Stata take a similar approach, this is, a random effects joint model. Among them we mention the R packages **JM** (Rizopoulos, 2010) and **JMBayes** (Rizopoulos, 2014) of Dimitris Rizopoulos, and the **joineR** of Philipson et al. (2012), apart from the Stata Module **STJM** of Crowther (2012). This work is not meant to be all-inclusive, but we want to illustrate the range of available techniques to apply joint modelling.

Focusing attention in the methodology presented, we give an overview of the implementation of the theory in the R package JM. First, illustrating the package design and its main functions. Following up with a summary of its limitations due to convergence problems.

3.6.1 Design

The R package **JM** constitutes a useful tool for the joint modelling of longitudinal and time-to-event data, in addition it contains all the methodology explained above. In order to adjust the sub-models, two additional packages are necessary. The linear mixed-effects modelling is based on the outpout of the function `lme()` from package `nlme`(ref), and the survival fit is implemented by either function `coxph()` or function `survreg()` of package `survival`(ref). Then, the joint model is fitted by `jointModel()`, which includes as main arguments the two separately fitted models to extract all the required information. It also incorporates an argument `method` to specify the type of the survival submodel, this is, the survival distribution and the regression model, and the algorithm numerical integration method, among the avaliable options are: `"piecewise-PH-GH"`, `"spline-PH-GH"`, `"Cox-PH-GH"`, `"weibull-PH-GH"`, and `"weibull-AFT-GH"`.

Once we have the returned object of class `jointModel`, the common options are available, such as the general results (functions `print()` and `summary()`), the estimated coefficients for the two sumodels (`coef()` and `fixef()`, the multiply-imputed residuals that account for nonrandom dropout (`residuals()`), the function `anova()` that computes the marginal Wald test and the likelihood ratio test based on fitted joint models. Moreover, the diagnostic plots (`plot()`), the predictions (`predict()`) and the log-likelihood value of the fitted model and the Akaike's and Bayesian information criteria (`logLik()` and AIC()). For more details and to get information about other additional functions, we refer to Rizopoulos (2012b, Appendix C).

3.6.2 Convergence Problems of the implemented JM Algorithm

The function `jointModel()` makes an automatic choice for default control arguments, as the number of quadrature points, number of iterations, convergence tolerances, etc. But it does not work always in practice, so in some occasions it is neccesary to take the values control. To assist in a possible divergence of the algorithm, the function incorporates a control argument `verbose`, which allow to print the optimization path towards the maximum. Possible solutions could be, among others:

i) Change the statting values,

ii) increase the number of EM iterations,

iii) choose other locations for the knots in the piecewise-constant or spline-based baseline hazard functions.

Chapter 4

Application to real data

In Chapter 3, we have introduced the joint regression model and the available package to furthermore apply to the data of liver transplantation. The main objective of this work is to investigate the abilities of postoperative glucose profiles to predict the death of patients who underwent Orthotopic Liver Transplantation (OLT), distinguishing between patients with and without a previous diagnosis of diabetes mellitus.

In order to analyse the data, we first describe the variables contained in the database, including the percentages of their respective missing values. To conduct this problem, we include a section of the procedure of missing imputation. Following up with the joint model approach proposed, and subsequently the results obtained and computational aspects are presented in the next sections.

4.1 Liver Transplantation Data

From the institutional clinical database, adult patients who underwent OLT in the Hospital Clínico Universitario de Santiago, between July 1994 and July 2011, were identified. Patients who were lost to follow-up and those who died in the first 72 hours were excluded. Registry data that did not conform within a range of expected results were rejected and reevaluated. A total of 632 patients were available for study. The participants were observed until either the primary endpoint (death) was reached or 31 July 2012 (median [range], 5.6 [0.1, 17.5] years). This study was approved by the Institutional Review Board (Comité Ético de Investigación Clínica de Galicia, Santiago, Spain).

The primary outcome studied was death from any cause before July 2011. Patients were followed up by the study team throughout their hospital stay. After discharge, vital status information was acquired by reviewing the Galician Health Registry, by contacting patients or their families individually and, if the patient had been rehospitalized, by reviewing the hospital records of major clinical events.

In this trial, the variables measured were:

- *Sex*, males or females.
- *Age*, age in years of the subject at the moment of the OLT.
- *bmi*, body mass index in kg/m^2.

- *diab*, diabetes mellitus, indicates if subjects was previously diagnosed of diabetes by physician.
- *meld*, model for end-stage liver disease, is a (continuous) score of severity of illness.
- *TIF*, cold ischemia time in minutes.
- *TH*, erythrocites transfusion (units).
- *TP*, platelets transfusion (units).
- *TVP*, thrombosis of vein porta.
- *NPTt*, parenteral nutrition (days).

To study the etiology of liver transplantation, the following binary variables were considered:

- *oh*, alcohol consumption.
- *hcv*, hepatitis C virus.
- *carc*, carcinoma.
- *vnc*, virus no C.
- *col*, cholangiocarcinoma.
- *cole*, cholestasis.
- *ote*, other causes.

Additionally, the glucose measurements were taken, once the previous day to the transplant, and the during the seven days after the liver transplantation. All of them recorded in the morning between 8 and 9 am. Particularly, emphasising that glucose profiles have not the same length, meaning that we are dealing with unbalanced dataset. Patients were classified as known diabetic patients if they had been informed of this diagnosis by a physician before admission or were on oral antihyperglycemic agents, insulin, or diet therapy.

4.1.1 Imputation

Missing data were observed in the following covariates (% of missing values): *meld* (57.12%); *TH* (4.27%); *TP* (3.80%); *TIF* (2.69%); *NPTt* (1.42%); *hcv*, *ote* and *TVP* (0.16% each one), apart from missing values in glucose levels, multiple imputation was used to estimate the missing values. To deal with these missing data, Multivariate Imputation by Chained Equations approach was made using the R package *mice* (van Buuren and Groothuis-Oudshoorn, 2011), with the number of imputations by default ($m = 5$):

```
> library(mice)
> transplante<-read.spss("transplante.sav")
> imputation=mice(transplante)
```

Then, the choice among these created data sets was made by selecting those in which the observed and imputed values of the variable *meld* were closest. In particular, the agreement was made regarding to the density plots shown in Figure 4.1, finally choosing the last imputed data set:

```
> require(lattice)
> densityplot(imputation, ~meld)
> transplanteimp=complete(imputation,5)
```

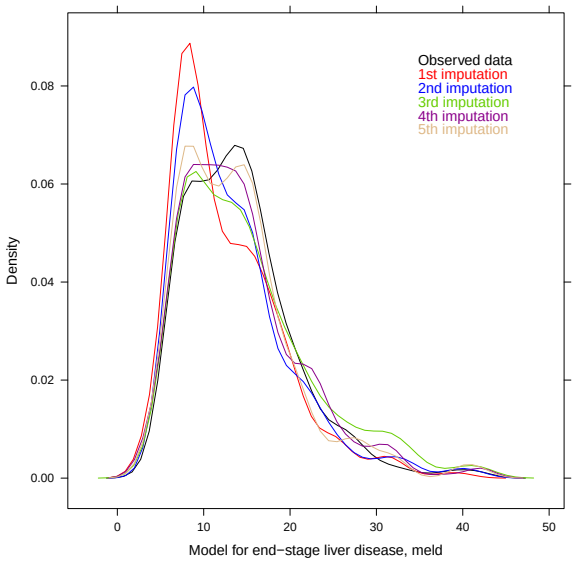


Figure 4.1: Plots of the densities for the observed and imputed data of the variable *meld*.

We had to define as "Not Available, NA" the glucose measurements imputed in patients who died before the 7-day period after surgery, staying only with the measurements before their death.

4.1.2 Descriptive Analysis

Table 4.1 displays patients characteristics. We observed that during follow-up 218 patients died. In order to explore the survival process we assessed each factor through univariate Cox regression, evaluating the magnitude of the association between covariates and survival through hazard ratios (HR), together with their corresponding 95% confidence intervals (CI). Besides this, we also show the p-values of likelihood ratio test to check which variables could be consider as predictors of the liver transplantation survival.

Table 4.1: Baseline Characteristics of patients who underwent OLT.

Characteristics	Descriptive Statistics	HR (95%CI)	p-value
Women, n (%)	158 (25%)	1.14 (0.85 − 1.55)	0.380
age, years	54 (45 − 60)	1.02 (1.01 − 1.03)	0.002
bmi, kg/m^2	27 (25 − 29)	0.97 (0.94 − 1.00)	0.081
diab, n (%)	125 (19.8%)	0.97 (0.69 − 1.36)	0.850
meld	13 (9 − 17)	1.03 (1.02 − 1.05)	< 0.001
oh, n (%)	386 (61.1%)	0.84 (0.64 − 1.10)	0.200
hcv, n (%)	130 (20.6%)	1.34 (0.98 − 1.81)	0.064
carc, n (%)	174 (27.5%)	1.37 (1.03 − 1.83)	0.033
vnc, n (%)	33 (5.2%)	1.06 (0.61 − 1.86)	0.840
col, n (%)	6 (0.9%)	1.58 (0.51 − 4.96)	0.430
cole, n (%)	31 (4.9%)	1.05 (0.57 − 1.92)	0.880
ote, n (%)	42 (6.6%)	0.97 (0.55 − 1.70)	0.910
TIF	7.18 (6 − 9)	1.05 (0.99 − 1.12)	0.098
TH	6 (3 − 10)	1.02 (1.01 − 1.04)	< 0.001
TP, n (%)	242 (38.3%)	1.07 (0.82 − 1.41)	0.609
TVP, n (%)	56 (8.9%)	1.02 (0.64 − 1.62)	0.936
NPTt	4 (3 − 6)	1.07 (1.05 − 1.09)	< 0.001

Data are expressed as median (IQR).

IQR= interquartile range; HR (95%CI) = hazard rate (95% confidence interval).

Univariate analysis showed that mortality risk was significantly higher in older patients ($p = 0.002$), patients with higher meld scores ($p < 0.001$), with carcinoma ($p < 0.033$), in those who needed a higher amount of blood transfusion units ($p < 0.001$), and with more time (days) needing parenteral nutrition ($p < 0.001$).

In Figures 4.2 and 4.3, the postoperative glucose profiles for individuals with and without previous diabetes, and for those who died in the first year post-transplantation can be seen , respectively.

4.2 Joint Modelling Approach

We applied the proposed joint model to the OLT data, with the aim of investigating the effect of repeated glucose measurements on time to all-cause death. In order to carry out this work, we introduced a joint model approach Rizopoulos (2012b) with different modelisations of survival sub-models and different linear mixed effects models for the longitudinal sub-model. All these different models were fitted by using the **JM** package, the R project joint modelling implementation.

From now on, all model approaches were made for patients with and without a previous diagnosis of diabetes. To this aim, the data was split in two different databases: "dm1" and "dm0", for diabetic and non-diabetic patients respectively. This is given that in non-diabetic subjects has been reported to exhibit high levels of glucose in response to situations of severe stress such as undergo major surgery or hospitalize in the intensive care unit, ICU (known as, reactive hyperglycemia). This behaviour is completely

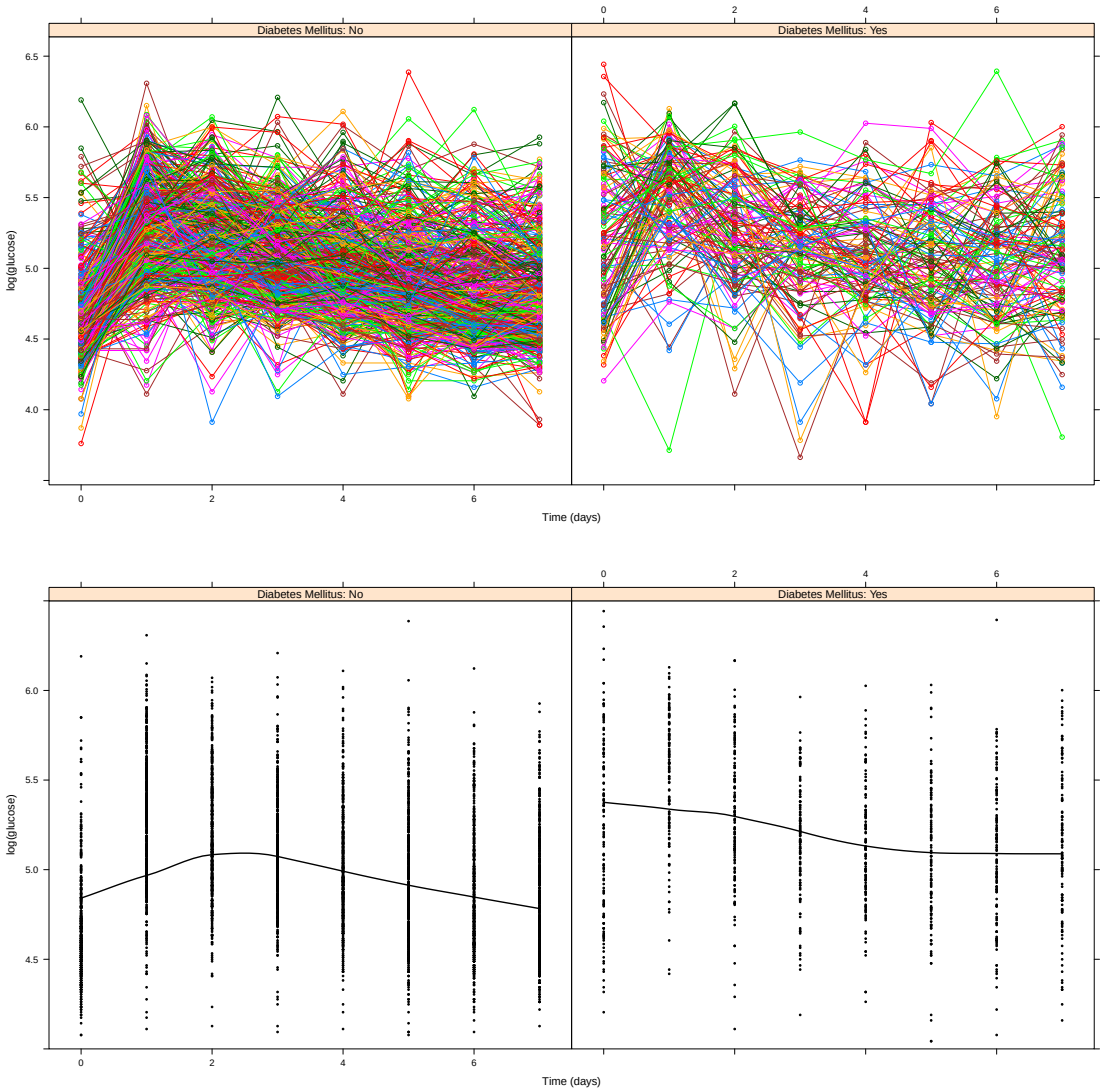


Figure 4.2: Subject specific trajectories of Glucose levels for patients with and without diabetes and the overall trajectories of Glucose with a p-spline method.

different in patients with diabetes. These events can be displayed in Figures 4.2, 4.3 and 4.4, where it is clearly visible that non-diabetic subjects have a different profile of glucose in the first week depending on having exitus or not.

We will show simultaneously both situations with their corresponding models. In addition, the R code below each section was added to illustrate the process.

4.2.1 Survival Submodel

For instance, in the descriptive analysis section we investigate through a univariate analysis which factors under investigation describe the survival process, but necessarily ignoring the impact of any others. Our purpose now is to determine which covariates potentially affect patient prognosis, distinguishing between diabetic and non-diabetic

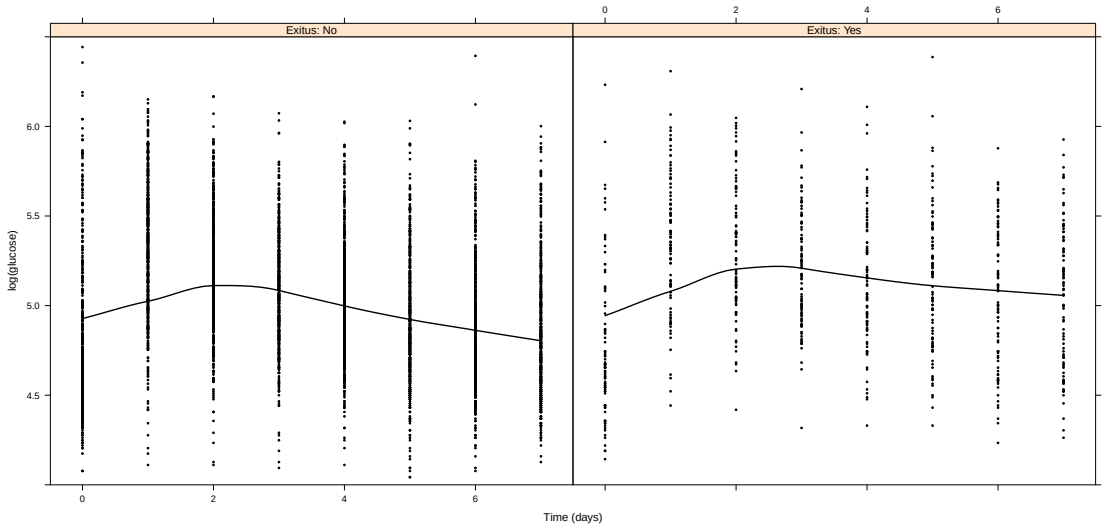


Figure 4.3: Overall trajectories of Glucose levels for patients with exitus versus nonexitus.

patients. Based on clinical experts' opinion, we introduced specific variables in the multivariate Cox regression models. Then, significant covariates were selected by using a backward stepwise procedure. The final models considered were the following:

$$\begin{aligned}\lambda_{i,diab}(t) &= \lambda_0(t)\exp(\gamma_1 hcv_i + \gamma_2 meld_i + \alpha\log(Glucose)_i(t)),\\ \lambda_{i,nodiab}(t) &= \lambda_0(t)\exp(\gamma_1 age_i + \gamma_2 carc_i + \gamma_3 meld_i\\ &\quad + \gamma_4 bmi_i + \gamma_5 TH_i + \alpha\log(Glucose)_i(t)),\end{aligned}$$

where $\lambda_0(t)$ is the baseline risk function, t is the time-to-event and $log(Glucose)$ is the true (unobserved) value of the longitudinal outcome. For the further introduction of the main R code, we presented the corresponding to the survival sub-models:

```
> fitSurvdiab<-coxph(Surv(timeExitus, exitus)
+ ~ hcv+meld,data=dm1,x=TRUE)
> fitSurvnodiab<-coxph(Surv(timeExitus, exitus)
+ ~ edad+carc+meld+imc+TH,data=dm0,x=TRUE,model=TRUE)
```

As mentioned in the **JM** package presentation in section 3.6, different types of survival sub-models can be fitted, such as Weibull model with a relative risk function and a spline-approximated baseline risk function, as well as the type of numerical integration method to approximate integrals. In Section 4.3, we present the results obtained from these approaches, among which the final model is choosed by comparing their Akaike Information Criterion (AIC; Akaike (1974)).

4.2.2 Longitudinal Submodel

First of all, it must be clear that we want to test survival of the patients who underwent liver transplantation by measuring the glucose profiles, among other baseline covariates. Then, it is obvious that these observations can be considered to be grouped by patients.

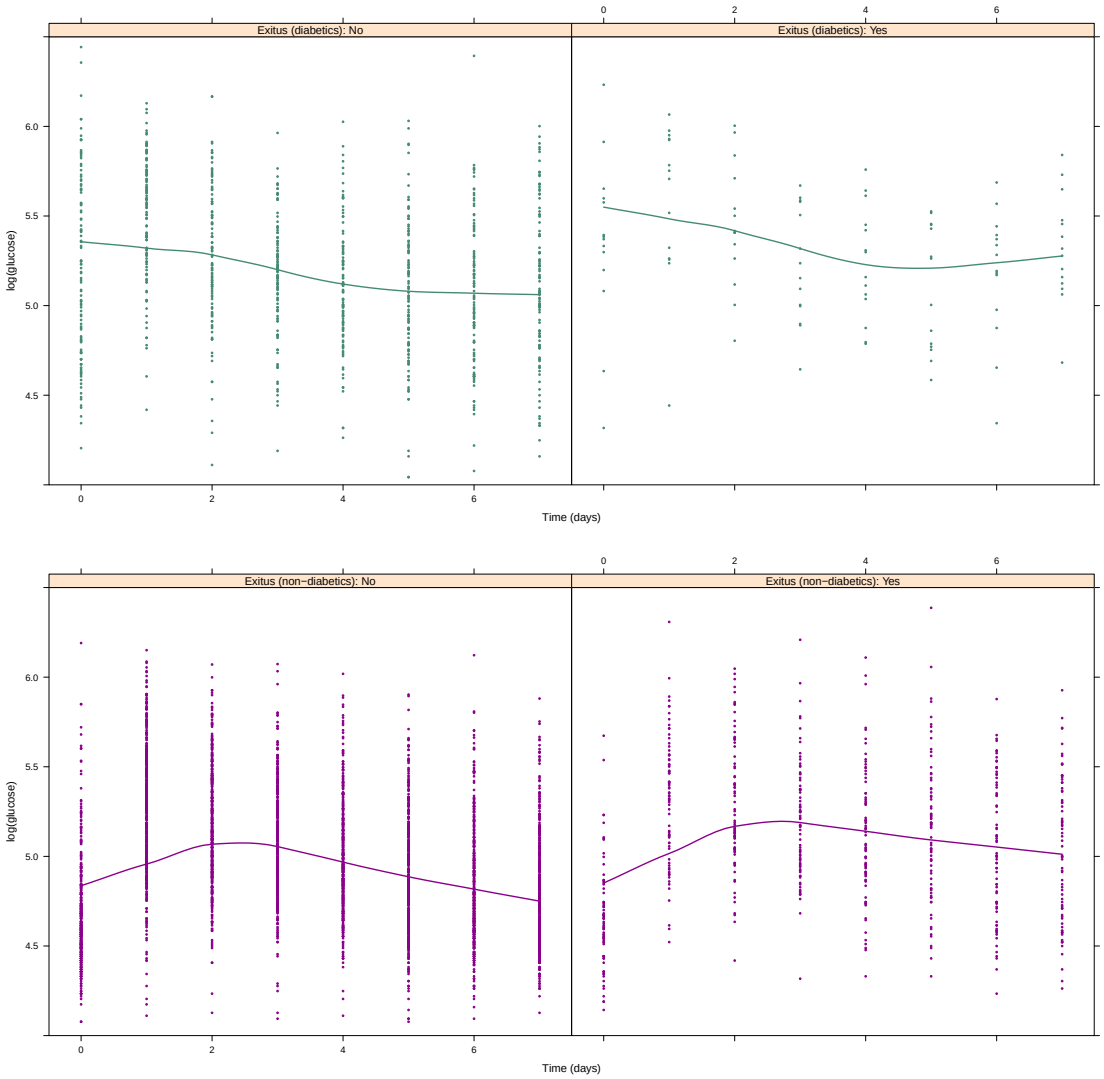


Figure 4.4: Overall trajectories of Glucose levels for patients with and without diabetes, differing by exitus versus nonexitus.

Once we have determined the cluster, thus, the hierarchical structure of the data, we must decide which type of analysis is going to be developed, a fixed effects model or a random effects model. Given that we wish to make inference about whole population of patients who undergo a liver surgery, it is clear that the clustering by patients must be incorporated as a random effect.

We implemented different longitudinal sub-models to study the longitudinal outcome by including the subject-specific random effects. Prior to this, it is necessary to implement these longitudinal models to reshape the data in order to obtain a longitudinal format in which we have the glucose measurements, the id number and the measuring times:

```
> transplanteimp.long<-reshape(transplanteimp,
+ varying = list(names(transplanteimp)[c(24,25,26,27,28,29,30,31)]),
```

```
+ direction = "long", v.names = c("Glu"), idvar = "id", times = c(0:7))
```

The proposed models are named as,

a) *Random Intercept Model*
In this model the intercepts are allowed to vary based on patients, and therefore, the scores on the dependent variable for each individual observation are predicted by the intercept that varies across individuals.

The sub-models can be written as:

$$\begin{aligned}\log(Glucose)_{i,diab} &= \beta_0 + \beta_1 time_i + \beta_2 hcv_i + \beta_3 meld_i + b_{0i} + \epsilon(t_{ij}),\\ \log(Glucose)_{i,nodiab} &= \beta_0 + \beta_1 time_i + \beta_2 age_i + \beta_3 carc_i + \beta_4 meld_i + \\ &\quad + \beta_5 bmi_i + \beta_6 TH_i + b_{0i} + \epsilon(t_{ij}),\end{aligned}$$

where *time* is the time that repeated measurements are taken and b_{0i} is the random intercept effect for each patient.

```
> fitLME.int.diab<-lme(log(Glu)~time+hcv+meld,random=~1|id,
+ data=subset(transplanteimp.long,transplanteimp.long$dm=="Yes"),
+  na.action=na.omit)
> fitLME.int.nodiab<-lme(log(Glu)~time+edad+carc+meld+imc+TH,
+ random=~1|id,data=subset(transplanteimp.long,
+ transplanteimp.long$dm=="No"), na.action=na.omit)
```

However, one of the problems comes from assuming that slopes are fixed. For this reason, the following sub-models were developed.

b) *Random Intercept and Slope Model*
Besides consider random intercepts, this model also allows random slopes to vary across subjects. The corresponding sub-models are given by,

$$\begin{aligned}\log(Glucose)_{i,diab} &= \beta_0 + \beta_1 time_i + \beta_2 hcv_i + \beta_3 meld_i + b_{0i} + \\ &\quad + b_{1i} t_{ij} + \epsilon_i(t_{ij}),\\ \log(Glucose)_{i,nodiab} &= \beta_0 + \beta_1 time_i + \beta_2 age_i + \beta_3 carc_i + \beta_4 meld_i + \\ &\quad + \beta_5 bmi_i + \beta_6 TH_i + b_{0i} + b_{1i} t_{ij} + \epsilon_i(t_{ij}),\end{aligned}$$

which additionally incorporate $b_{1i}t_{ij}$ that represents the random slope effect of the different Glucose trajectories of each patient.

```
> ctrl <- lmeControl(opt='optim')
> fitLME.slope.diab<-lme(log(Glu)~time+hcv+meld,random=~time|id,
+ data=subset(transplanteimp.long,transplanteimp.long$dm=="Yes"),
+ na.action=na.omit,control=ctrl)
> fitLME.slope.nodiab<-lme(log(Glu)~time+edad+carc+meld+imc+TH,
+ random=~time|id,data=subset(transplanteimp.long,
+ transplanteimp.long$dm=="No"),na.action=na.omit,control=ctrl)
```

c) *Spline Model*
Becasue non-linear patterns of Glucose trajectories, as we can observe in Figure

4.2, a splined-based approach is also considered in order to obtain more flexible regression models. In particular, that proposed by Rizopoulos and Ghosh (2011), which consider natural cubic splines:

$$\begin{aligned} log(Glucose)_{i,diab} &= (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})B_n(t, d1) + (\beta_2 + b_{12})B_n(t, d_2) \\ &\quad + (\beta_3 + b_{i3})B_n(t, d_3) + \beta_4 hcv_i + \beta_5 meld_i + \epsilon_i(t), \\ log(Glucose)_{i,nodiab} &= (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})B_n(t, d1) + (\beta_2 + b_{12})B_n(t, d_2) \\ &\quad + (\beta_3 + b_{i3})B_n(t, d_3) + \beta_4 age_i + \beta_5 carc_i + \\ &\quad + \beta_6 meld_i + \beta_7 bmi_i + \beta_8 TH_i + \epsilon_i(t), \end{aligned}$$

where $\{B_n(t, d_k); k = 1, 2, 3\}$ denotes a B-spline basis matrix for a natural cubic spline (de Boor, 1978).

```
> fitLME.spline.diab<-lme(log(Glu)~ns(time,3)+hcv+meld,
+ random=list(id=pdDiag(form=~ns(time,3))),
+ data = subset(transplanteimp.long, transplanteimp.long$dm
+ =="Yes"),na.action=na.omit)
> fitLME.spline.nodiab<-lme(log(Glu)~ns(time,3)+edad+carc+
+ meld+imc+TH,random=list(id=pdDiag(form=~ns(time,3))),
+ data = subset(transplanteimp.long,transplanteimp.long$dm
+ =="Yes"),na.action=na.omit)
```

4.3 Results

In this Section we applied the joint modelling approach described in Section 3, to assess the effect of glucose profiles in survival after Liver transplantation.

As we have seen above, different longitudinal sub-models are analysed with an only intercept, intercept and slope analysis and a non-linear subject specific evolutions for the Glucose levels. Using Akaike Information Criterion (AIC), we chose the longitudinal sub-model for each group with the less AIC value as shown in Table 4.2.

Table 4.2: AIC values of different Longitudinal submodels.

	AIC	
Model	Diabetes	No Diabetes
Intercept	1040.912	3307.567
Intercept + Slope	1044.749	3259.037
Spline	1022.747	2230.935

According to these results, it is an evidence of our choice of spline-based longitudinal sub-models. Then, we fitted the joint model by comparing different sub-models for the survival process of transplant data such as mentioned above: I) a time-dependent relative risk model with Weibull baseline risk function and II) a time-dependent relative risk model with the log baseline risk function that is approximated using B-splines. The R code is illustrated in a general context, but it is enough to replace the respective sub-models for diabetic and non-diabetic patients listed above:

```
>fit.JM.weibull<-jointModel(fitLME.spline,fitSurv,timeVar="time")
```

```
>fit.JM.spline<-jointModel(fitLME.spline,fitSurv,
+timeVar="time",method="spline-PH-aGH")
```

Following the same procedure, we obtained the final model which the less AIC value as shown in Table 4.3.

Table 4.3: AIC values of different survival submodels.

	AIC	
Model	Diabetes	No Diabetes
fit.JM.weibull	1762.431	5356.511
fit.JM.spline	-	8977.696

The results indicate that final models for diabetic and non-diabetic patients take a relative risk model with Weibull baseline risk function and with a spline longitudinal sub-model. In Table 4.4 and 4.5 we synthesized all the information of both joint models.

Table 4.4: Fitted values of the final model for the joint model approaches for diabetic patients.

Joint Models (*JM*) - Diabetic patients			
		Coef	Std. error
Longitudinal Process	Intercept (β_0)	5.1813	0.0582
	β_1	-0.4112	0.0487
	β_2	-0.1384	0.0760
	β_3	-0.2273	0.0376
	β_4	-0.0788	0.0491
	β_5	0.0099	0.0034
Survival Process	Glucose	0.0002	0.0020
	hcv	0.8528	0.3355
	meld	0.0191	0.0256
LogLikelihood		-865.2155	

We observed that non-diabetic patients with higher Glucose levels have a worse survival through the hazard ratios , HR= 1.002 (95% CI: $1.0004 - 1.0036$). However, for diabetic patients the association between the Glucose levels and survival process is not statistically significant, HR=1.0002 (95% CI: $0.9963 - 1.0041$).

4.3.1 Joint Model Diagnostics

Once both joint models are fitted, the next step is to verify if all the necessary model assumptions are valid. Standard types of residuals plots can be used to validate the assumptions behind mixed models and relative risk models when these are separately fitted.

In order to validate the proportional hazards assumption of the survival submodels, a graph of the Schoenfeld residuals was displayed to check the overall goodness-of-fit of our relative risk submodels. The results obtained are shown in Figures 4.5 and 4.6. Because zero slopes in the generalized linear regression of the scaled Schoenfeld

Table 4.5: Fitted values of the final model for the joint model approaches for non-diabetic patients.

Joint Models (*JM*) - Non-Diabetic patients			
		Coef	Std. error
Longitudinal Process	Intercept (β_0)	4.4018	0.0525
	β_1	-0.1875	0.0184
	β_2	0.7047	0.0331
	β_3	-0.3460	0.0136
	β_4	0.0036	0.0006
	β_5	-0.0048	0.0170
	β_6	0.0054	0.0011
	β_7	0.0027	0.0018
	β_8	0.0040	0.0008
Survival Process	Glucose	0.0020	0.0008
	age	0.0308	0.0080
	carc	0.6401	0.1892
	meld	0.0643	0.0115
	bmi	-0.0523	0.0226
	TH	0.0273	0.0068
LogLikelihood		−2656.256	

residuals, and considering imputation process, the covariates are under a proportional hazard assumption.

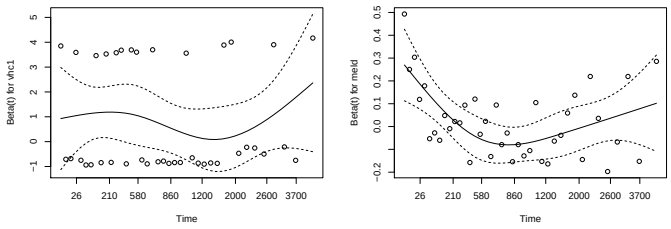


Figure 4.5: Schoenfeld residuals for the survival submodel for patients with diabetes.

On the other hand, the residuals for the longitudinal part can also be checked by diagnostic plots for linear mixed-effects models. However, the dropout mechanism are not accounted for residuals for the longitudinal process. To overcome this problem, the multiple imputation residuals for fixed visit times proposed by Rizopoulos et al. (2010) are used. For both diabetic and non-diabetic patients multiply-imputed standardized marginal residuals are plotted in Figure 4.7 (with their black dashed line loess smooth), together with the observed standardized marginal residuals (and their black loess smooth). A comparison between the two curves reveals that trends in the observed residuals plots are attributed to nonrandom dropout and not to a model lack-of-fit.

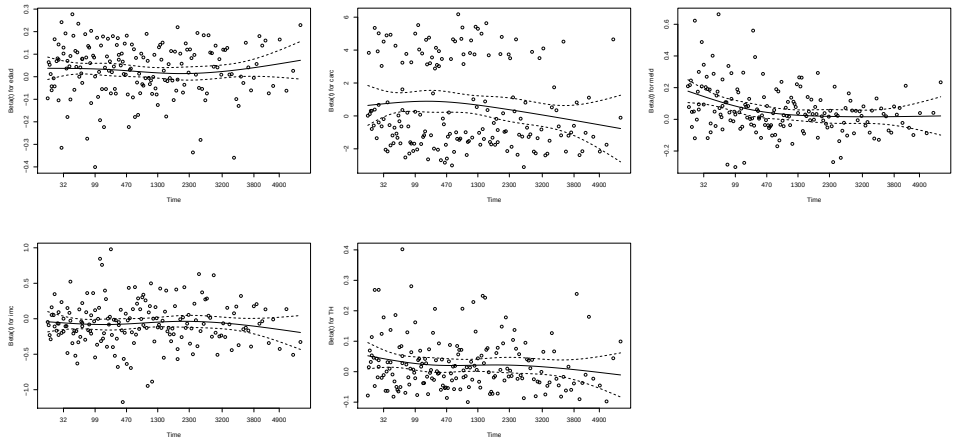


Figure 4.6: Schoenfeld residuals for the survival submodel for patients without diabetes.

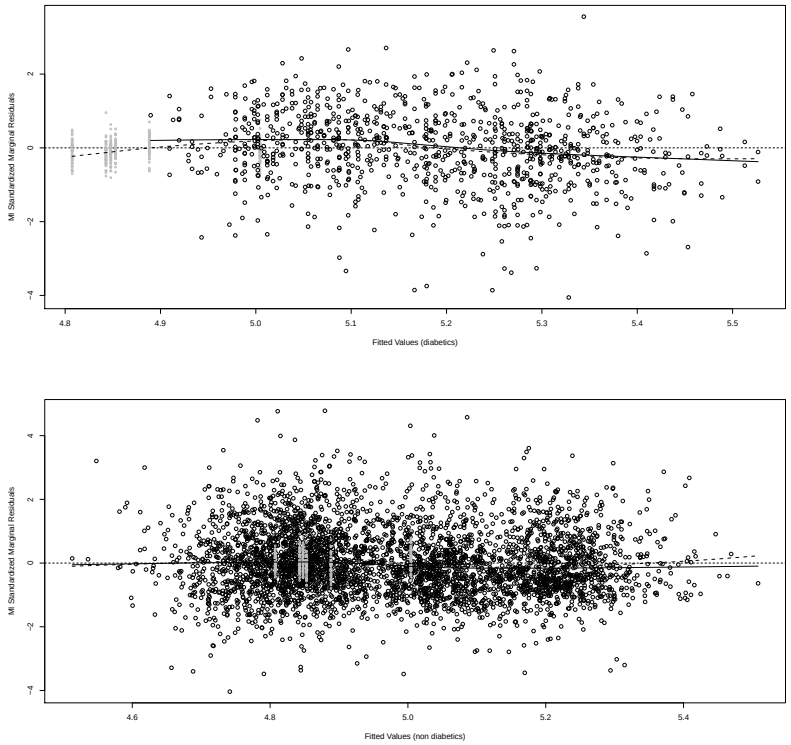


Figure 4.7: Observed standardized marginal residuals (black points), augmented with all the multiply imputed residuals (grey points). The superimposed solid lines represent a loess fit based only on the observed residuals (black line), and a weighted loess fit based on all residuals (black dashed line).

4.3.2 Predictions

Distinguishing the predictions between patients with and without diabetes, we compared the joint modelling approaches, described in Section 3, with the following survival models, to notice the advantage and disadvantages of each:

Survival model (1)

In these models we introduced the baseline covariates: *hcv* and *meld* for diabetic patients and *age*, *carc*, *meld*, *bmi* and *TH* for non-diabetic patients.

$$\begin{aligned}\lambda_{i,diab}(t) &= \lambda_0(t)\exp(\gamma_1 hcv_i + \gamma_2 meld_i),\\ \lambda_{i,nodiab}(t) &= \lambda_0(t)\exp(\gamma_1 age_i + \gamma_2 carc_i + \gamma_3 meld_i\\ &\quad + \gamma_4 bmi_i + \gamma_5 TH_i).\end{aligned}$$

Survival model (2)

In the second model we added a baseline Glucose level (*glu0*) as another covariate.

$$\begin{aligned}\lambda_{i,diab}(t) &= \lambda_0(t)\exp(\gamma_1 hcv_i + \gamma_2 meld_i + \gamma_3 glu0),\\ \lambda_{i,nodiab}(t) &= \lambda_0(t)\exp(\gamma_1 age_i + \gamma_2 carc_i + \gamma_3 meld_i\\ &\quad + \gamma_4 bmi_i + \gamma_5 TH_i + \gamma_6 glu0).\end{aligned}$$

In order to carry out such comparison, we used the linear predictors at time t to compute the ROC curves and the Area Under Curve for each time point (Heagerty and Zheng, 2005). This calculation is implemented in R package *risksetROC* (Heagerty and Saha-Chaudhuri, 2012). As we can observe in Figure 4.8, the behaviour is completely different depending on a positive or negative diabetes diagnosis.

Dynamic predictions

In joint models the conditional probabilities of the survival process are dynamically updated. Because of their dynamic structure, these predictions can assist the clinicians to make decisions. Also they provide clear observations of the association between the longitudinal and the survival process.

In Figure 4.9 two patients with different Glucose behaviours can be observed: a diabetic patient (subject 51) and a non-diabetic patient (subject 40), with their longitudinal observations to observe the effect of the longitudinal outcome to the risk of death of these patients. We observe a lower survival probability for the patient who has higher Glucose levels.

In this study, we observe the association in the dynamic predictions' graphics, however we can not show the dynamic nature of these predictions because the Glucose measurements are only taken in the 7 post-operative days.

4.4 Computational Aspects

Joint modelling approach for longitudinal and time-to-event data requires a combination of a double numerical integration and optimization. The function `jointmodel()` implements a hybrid optimization procedure to locate the maximum likelihood estimates, starting with EM algorithm and if not converge switches to a quasi-Newton algorithm until it converges. These requirements of both double optimization and numerical integration may lead us to experience convergence problems. Although the function `jointmodel()` permits choices for default control arguments such as number of

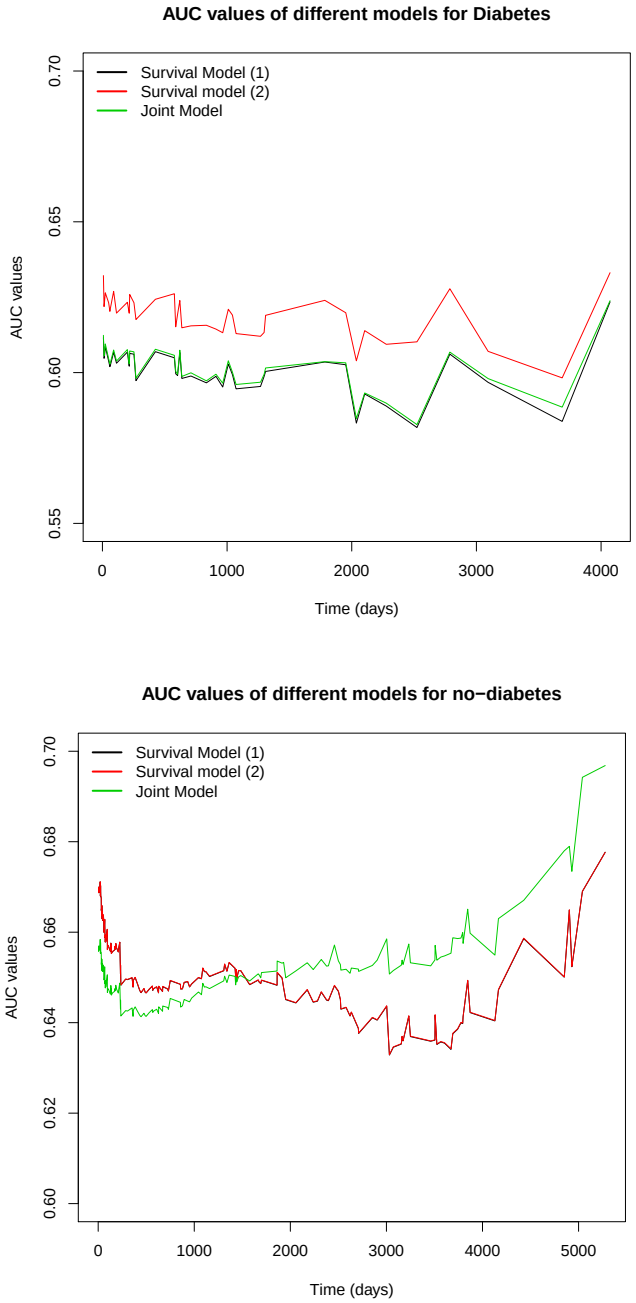


Figure 4.8: Time dependent AUCs for each model separately for patients with and without diabetes.

quadratic points, number of iterations, convergence tolerances, it is not guaranteed to work in all datasets. These aspects are described and discussed in (Rizopoulos, 2012b, pp. 61-87) . The author also mentions that in the majority of the cases, the converge problems can be avoided by changing the stating values, increasing EM iterations or choosing other locations for the knots if the piecewise constant of spline-based baseline hazard functions are used. In this study we experienced converge problems while fitting a spline-based baseline hazard function for the survival sub-model.

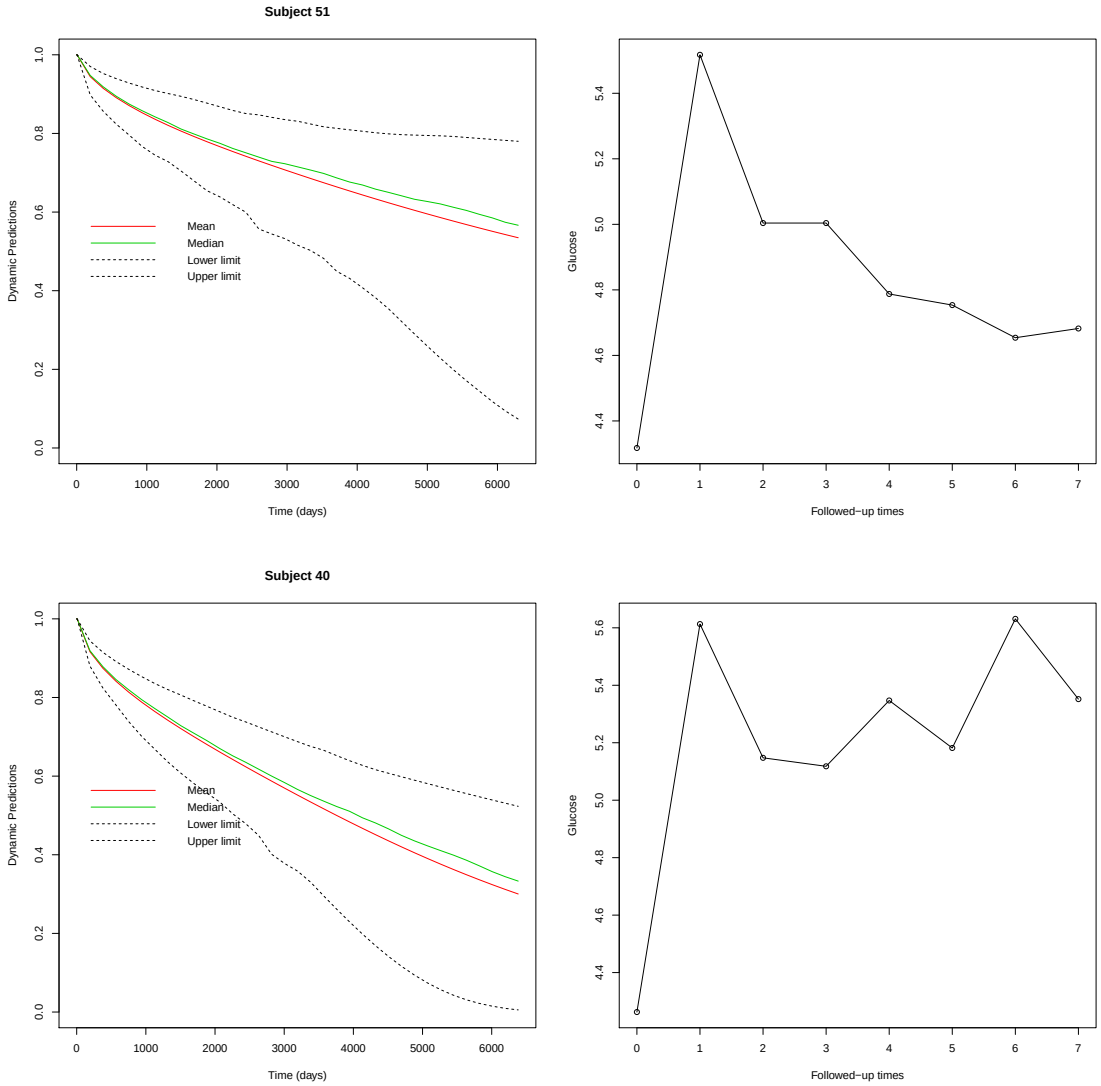


Figure 4.9: Dynamic predictions for the final model for a diabetic patient (subject 51) and a non-diabetic patient (subject 40).

```
>fit.JM.spline1<-jointModel(fitLME.spline,fitSurv,timeVar="time",
 +method="spline-PH-aGH",verbose=T)
```

Setting the "verbose" option to TRUE, the function gives us the optimization path:

```
iter: 1
log-likelihood: -2302.596
betas: 5.2325 -0.412 -0.1326 -0.2248 -0.0766 0.0059
sigma: 0.3613
gammas: 1.2437 8e-04
alpha: 0.6122
gammas.bs: -12.7378 -11.4164 -11.4444 -11.027 -10.9222 -10.7309 -10.6004
          -10.4899 -10.4587
D: 0.0393 0 0 0.0318
```

```
...

iter: 119
log-likelihood: 0
betas: 5.2326 -0.4117 -0.1327 -0.2248 -0.0763 0.0059
sigma: 0.4144
gammas: 1.2437 8e-04
alpha: 0.6122
gammas.bs: -12.7378 -11.4164 -11.4444 -11.027 -10.9222 -10.7309 -10.6004
        -10.4899 -10.4587
D: NaN NaN NaN NaN

iter: 120
log-likelihood: 0
betas: 5.2326 -0.4117 -0.1327 -0.2248 -0.0763 0.0059
sigma: 0.4144
gammas: 1.2437 8e-04
alpha: 0.6122
gammas.bs: -12.7378 -11.4164 -11.4444 -11.027 -10.9222 -10.7309 -10.6004
        -10.4899 -10.4587
D: NaN NaN NaN NaN


quasi-Newton iterations start.

Error en optim(thetas, LogLik.splineGH, Score.splineGH, method = "BFGS",  :
  valor no finito provisto por optim
```

We can observe in the output the first iteration has a reasonable log-likelihood value but from the iteration 3 the log-likelihood values are 0 and then higher estimations for the longitudinal sub-model coefficients have obtained. As mentions Rizopoulos (2012b), this problem usually comes from a failure of the numerical integration rule or from parameter scale problem.

Besides its several advantages in estimations, the joint modelling doesn't have a fast computation in which some further studies required. For this particular study, the timing of the joint model with different survival sub-models and a cubic spline model for the longitudinal process are as follows, 2.43 minutes for B-spline survival model and 10.55 minutes for the Weibull model for non-diabetic patients, and 0.95 minutes for the Weibull model for diabetic patients. The timings above-mentioned are based on Intel (R) Core(TM) 2.40 GHz 8GB RAM, Windows 7.

Chapter 5

Conclusions

In this work the joint modelling approach proposed by Rizopoulos (2010) is presented. This methodology constitutes a useful tool to study the relationship between longitudinal and survival data. This was proved by fitting joint regression models to study the survival of patients with and without a previous diagnosed of diabetes mellitus who underwent Orthotopic Liver Transplantation (OLT).

From these models, for non-diabetic patients we have obtained a significant effect of Glucose levels on survival. This effect improves the predictive performance of the model according to AUC values. Accordingly with the results obtained, it has been observed different behaviours of AUC values for patients with and without diabetes. It can be seen in Figure 4.8, in which for non-diabetic patients the discrimination capacity is low for a model with a single Glucose measure. However, predictions improve by adding longitudinal observations of the Glucose levels. Indeed it does not occur for diabetic patients, justified by non significant effect of Glucose levels on survival.

From the clinical point of view, results show that in individuals previously undiagnosed diabetes, the profiles of the glucose in the immediate post-operative (one week) patients who underwent OLT may be useful in predicting mortality. These findings reinforce the hypothesis of "reactive hyperglycemia", in which blood glucose in days following transplantation presents a different behaviour in those with higher mortality. Thus suggesting that these profiles can emerged in prognostic markers of mortality in subjects undergoing stress for hospital admission to a major surgery or to the intensive care unit (ICU).

Moreover, dynamic predictions are shown to illustrate that it is important to use all available information to produce predictions of survival probabilities. Despite not being the primary interest of this work, this could be useful to the physicians to gain a better understanding of the disease dynamics. As an example, we can observe in Figure 4.9, the dynamic predictions of the survival process of patient 40, who has no diabetes, and presents a high post-transplant Glucose levels in the first week. Additionally, computational aspects are mentioned, in particular, those difficulties encountered in implementing the joint proposal.

To conclude, a further progress in this area is essential for both statistical and medical aspects. As a step toward this objective, we would like to call for providing a test to

determine whether an extra random effect should be included in the joint model. And, to take full advantage of the results, a future research line will be to carry over into the field of reclassification. In order to establish risk scores, and therefore to rank risk.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, 10:1100–1120.

Breslow, N. (1972). Discussion of paper 'regression models and life-tables' by D. Cox. *Journal of the Royal Statistical Society, Series B*, 34:216–217.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, 2(3):437–453.

Cox, D. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34(2):187–220.

Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

Cox, D. and Snell, E. (1968). A general definition of residuals. *Journal of Royal Statistical Society, Series B*, 30:248–275.

Crowther, M. J. (2012). STJM: Stata module to fit shared parameter joint models of longitudinal and survival data. http://ideas.repec.org/c/boc/bocode/s457342.html. Last checked: 16.06.2014.

de Boor, C. (1978). *A Practical Guide to Splines*. Springer, Berlin.

Dossett, L., Cao, H., Mowery, N., Dortch, M., Morris, J. J., and May, A. (2008). Blood glucose variability is associated with mortality in the surgical intensive care unit. *American Journal of Surgery*, 74(8):679–685.

Dutkowski, P., Rougemont, O. D., and Clavien, P. (2010). Current and future trends in liver transplantation in europe. *Gastroenterology,*, 138(3):802–809.

Egi, M., Bellomo, R., Stachowski, E., French, C., and Hart, G. (2006). Variability of blood glucose concentration and short-term mortality in critically ill patients. *Anesthesiology*, 105(2):244–252.

Fleming, T. and Harrington, D. (1984). Nonparametric estimation of the survival distribution. *Communications in Statistics: Theory and Methods*, 13:2469–2486.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of The American Statistical Association*, 72:320–340.

Heagerty, P. J. and Saha-Chaudhuri, P. (2012). Riskset ROC curve estimation from censored survival data. http://cran.r-project.org/web/packages/risksetROC/index.html. [Last checked: 16.06.2014].

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105.

Hsieh, F., Tseng, Y.-K., and Wang, J.-L. (2006). Joint modelling of survival and longitudinal data: likelihood approach revisited. *Biometrics*, 62:1037–103.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York, 2nd edition edition.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation for incomplete observations. *Journal of the American Statistical Association*, 93:457–481.

Kleinbaum, D. G. and Klein, M. (2005). *Survival Analysis. A Self-Learning Text*. Springer, New York, 2nd edition.

Krinsley, J. (2008). Glycemic variability: a strong independent predictor of mortality in critically ill patients. *Critical Care Medicine*, 36(11):3008–3013.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, 38:963–974.

Lange, K. (2004). *Optimization*. Springer-Verlag, New York.

Lindstrom, M. and Bates, D. (1988). Newton-Raphson and EM algortihms for linear mixed-effects models for repeated measures data. *Journal of the American Statistical Association*, 83:1014–1022.

Meyfroidt, G., Keenan, D., Wang, X., Wouters, P., Veldhuis, J., and den Berghe, G. V. (2010). Dynamic characteristics of blood glucose time series during the course of critical illness: effects of intensive insulin therapy and relative association with mortality. *Critical Care Medicine*, 38(4):1021–1029.

Peterson, A. V. (1977). Expressing the Kaplan–Meier estimator as a function of empirical subsurvival functions. *Journal of Americal Statistical Association*, 72(360):854–858.

Philipson, P., Sousa, I., Diggle, P., Williamson, P., Kolamunnage-Dona, R., and Henderson, R. (2012). joineR: Joint modelling of repeated measurements and time-to-event data. http://cran.r-project.org/web/packages/joineR/vignettes/joineR.pdf. [Last checked: 16.06.2014].

Prentice, R. (1982). Covariate measurement errors and parameter estimates in a failure time regression model. *Biometrika*, 69:331–342.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. http://www.jstatsoft.org/v35/i09/. [Last checked: 16.06.2014].

Rizopoulos, D. (2011). Dynamic predictions and properspective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67:819–829.

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.

Rizopoulos, D. (2012b). *Joint Models for Longitudinal and Time-to-Event Data. With Applications in R.*

Rizopoulos, D. (2014). JMbayes: Joint modeling of longitudinal and time-to-event data under a bayesian approach. http://cran.r-project.org/web/packages/JMbayes/JMbayes.pdf. [Last checked: 16.06.2014].

Rizopoulos, D. and Ghosh, P. (2011). A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*, 30:1366–1380.

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

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2010). Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics*, 66:20–29.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Self, S. and Pawitan, Y. (1992). *Modelling a marker of disease progression and onset of disease.* AIDS Epidemiology: Methodological Issues, Birkhaser, Boston.

Tsiatis, A. A. and Davidian, M. (2004). Joint modelling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834.

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

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer-Verlag, New York.

Wolbers, M., Babiker, A., Sabin, C., and et al. (2010). Pretreatment CD4 cell slope and progression to AIDS or death in HIV-infected patients initiating antiretroviral therapy. The CASCADE collaboration: a collaboration of 23 cohort studies. *PloS Medicine*, 7(2):e1000239.

Wulfsohn, M. and Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 53:375–387.