Modeling the association between eGFR and survival in kidney research

Joint Models and Extensions

Master Thesis in Biostatistics (STA495)

by

Silvia Panunzi
silvia.panunzi@uzh.ch
matricula 15-741-820

supervised by

Prof. Armando Teixeira-Pinto
University of Sydney, School of Public Health
Prof. Torsten Hothorn
University of Zurich, Department of Biostatistics

University of Zurich[UZH]

Zurich, May 21, 2018

To the memory of my mother

"It always seems impossible until it is done"
— Nelson Mandela

Abstract

In clinical practice biomarkers are becoming of central importance for risk disease assessment, disease prevention, diagnosis and monitoring of therapies. Increasing interest is nowadays allocated to study potential associations between biomarkers and time to event outcomes. Joint models popularity is growing to follow clinical research needs. Joint models constitutes not only an appropriate tool for inference in the relationship between two different outcomes, such as longitudinal measurements and survival, but they have also been utilized to provide individualized predictions. They enable to study patients' dynamic survival probabilities, giving deeper insight in prevention studies.

Alternative methods have been proposed in literature with the same aim as joint models. Extended version of the Cox model with longitudinal covariates or two-stage estimation approaches are common examples. However, several limitations have been proved to exist for these techniques, making joint modelling approaches more and more popular. The key point for this set of models is that they have to handle at the same time correlated repeated measurements recorded for each subject during the period of follow-up, with possible missing observations, and incomplete time-to-event data, that often occur due to censored observations.

In this work, we provide extensive review of joint models approach for longitudinal and time-to-event data. We particularly focus on the relationship between patients' survival after kidney transplant and a biomarker of kidney function, the estimated glomerular filtration rate (eGFR), and assess the predictive capacity of this biomarker.

Contents

Chapter 1

Introduction

1.1 Overview

Life sciences analyses often require researchers and scientists to focus on the simultaneous analysis of multiple outcomes. In the contest of longitudinal and time-to-event data for example we usually observe repeated markers measures for the same set of individuals over time and try to relate them with observations on the time to a certain event. Joint models have recently being developed to overcome the inappropriateness of separate analysis that fail to take into account the association and dependence between the two components of the data. Joint modeling approach in fact, enables researchers to make the most efficient use of the complete set of data and to identify effects of variables when controlling for the interplay among the two processes object of investigation. From their introduction in the 90s (Self and Pawitan [1992], De Gruttola and Tu [1994], Faucett and Thomas [1996], Wulfsohn and Tsiatis [1997b]) joint models have been applied in a large scale of studies and widely extended to address challenging methodological and applied questions. Rizopoulos [2012] provides a comprehensive overview for a both theoretical and software exploiting. Interesting and very recent publications in this modeling framework can be found in the latest special issue of the Biometrical Journal (`http://onlinelibrary.wiley.com/doi/10.1002/bimj.v59.6/issuetoc?campaign=woletoc`). The fact that joint models constitute an active area of statistics research that has received a lot of attention in the recent years support us in pointing out how actual if our topic of study.

1.2 Motivating Study

Our research is primary motivated by the analysis of a transplantation database. In kidney disease new immunosuppressive drugs have decreased the incidence of rejection and have improved graft survival in the short term. However in the long term graft outcomes do not improve. Mortality of recipients with functioning grafts has been attributed mostly to cardiovascular diseases, graft losses to chronic allograft nephropathy [Marcén et al., 2010]. Monitoring the allograft function is very important after kidney transplant. Glomerular filtration rate (GFR), is considered the best index of kidney function, an indicator of long-term graft survival and an inde-

pendent risk factor for cardiovascular mortality [Santos and Martins, 2015]. In this study we want, specifically, to examine the relationship between estimated glomerular filtration rate (eGFR), derivative from serum creatinine blood measurement, and graft survival.

Data arises from the collaboration with the Center for Kidney Research (CKR), Westmead (Sydney), that conducts and implements high-priority research in the prevention, diagnosis, treatment, and care of people with or at risk of chronic kidney disease and related conditions. In chapter 6 we will introduce the Australian and New Zealand registry of transplants data and further explore our motivational dataset.

1.3 Goals

The goal of this manuscript is to illustrate an appropriate methodology to jointly analyse repeated measurements of biomarkers and event times of individuals and give them an experimental application studying the association between death or graft failure and a longitudinal biomarker, calculating updated event risks for transplanted patients and eGFR predictive performance. Joint modeling strategies and their extensions will be here exhaustively explained.

1.3.1 Research Questions

- Is there an association between time to death (or graft-failure) and the evolution of the eGFR biomarker?
- Is the analysis of the biomarker evolution helpful in predicting patients' conditional survival probabilities?

Chapter 2

Theory Background

This chapter is meant to give the reader an introductory understanding of classical models commonly used to separately analyze longitudinal and survival data. A non exhaustive literature studying special issues in the following topics will be also provided along with the discussion.

2.1 Longitudinal Data Analysis

Typically longitudinal data arise from collection of participants outcomes measured at different follow-up times.

Longitudinal studies allow us to investigate how for example treatment (or exposures) means differ at specific time points (cross-sectional effect) and how they change over time (longitudinal effect).They also enable us to distinguish changes over time within individuals (so called ageing effects) from differences among people in their baseline levels (cohort effects).

Simplest form in the family of prospective longitudinal studies is the analysis of changes from baseline to follow-up but even with this simple case we could have complications. In practice we usually have to deal with clinical trial data not balanced and not equally spaced. In fact it often happens that patients attend a different number of visits, in different times. One of the major challenges for the analysis of longitudinal outcomes is that these are often incomplete. Missing data can occur when patients are missing at intermittent times or when they eventually drop-out of the study. If subjects that are followed to the planned end of study differ from subjects with discontinue follow-up then a naive analysis may provide summaries that are not representative of the original target population. Mechanism leading to missing data should be carefully evaluated. A common classification of different missing data mechanisms distinguishes between three general cases:

- Missing completely at random (MCAR) when probability of missing is independent of any variable, this is the case for subjects who go out of the study after providing a pre-determined number of measurements or when laboratory measurements are lost simply due to equipment malfunction.

- Missing at random (MAR) when probability of missing depends on a set of covariates but not on the outcome, e.g. study protocols require patients whose

response value exceeds a threshold to be removed from the study.

- Missing not at random (MNAR) when probability of missing is actually related to the response variable we observe or not, on the data.

First situation is easily handled because under this state we can consider observed data as a random sample of the complete data, in the second case we cannot but we nevertheless can use likelihood inference methods and obtain valid results for subject-specific evolution and residuals. The third scenario is instead the most complex one, for appropriate inference on data it requires us to use procedures that explicitly model the joint distribution of the longitudinal and of the missingness process.

Longitudinal data are rather common in clinical research. As popular example HIV studies are usually mentioned, in these analyses CD4 cell counts are repeatedly measured over time because their value is causally associated with the evolution of AIDS virus.

Special methods are needed for the analysis of these data, in fact observations coming from the same subject can be thought as naturally correlated, therefore classical approaches as t-tests and linear regression models assuming independence between observations are no longer appropriate. Valid inferences can be achieved only if this correlation has been taken into account.

Two approaches are usually applied for longitudinal problems: Generalized Linear Mixed Models (GLMM) and Generalized Estimating Equations (GEE). Their purpose is to describe the dependence of the response on explanatory variables distinguishing between continuous, binary, normal and non normal outcomes and taking into account many further model specific assumptions. In the next subsection we are going to investigate more deeply the GLMM framework that will serve us for our research topic, for other approaches using GEE or Generalized Least Square (GLS) methods we remand the lecturer to a well-argued book such as [Diggle, 2002].

2.1.1 Generalized Linear Mixed-Effects Models

Generalized Linear Mixed Effects Models (GLMM) are the standard setting in longitudinal data analysis. These models can be seen as an extension of Generalized linear models (GLM) that incorporates random regression coefficients to characterize within-subject correlations in the data. The main idea is that each individual in the population has his own specific mean response profile over time that has to be modeled to obtain correct inferences. Let's use y_{ij} to denote the response of subject i $(i = 1, ..., n)$ at time t_{ij}, $j = 1, ..., n_i$. The evolution of each subject in time can be described by a linear model:

$$y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \tag{2.1}$$

where the terms $b_i = (b_{i0}, b_{i1})^T$ are the random effects that describe the variability of the individuals in population with some prespecified distribution, usually Normal. On the other hand parameters β_0 and β_1 are fixed effects describing the average population evolution process. If we have a continuous outcome Linear Mixed-Effects

Models (LMM) are implemented in the analysis. In a more general form we can rewrite the (marginal) model as:

$$\begin{cases} \mathrm{y}_i = X_i\beta + Z_i b_i + \epsilon_i, \\ \beta = (\beta_i, ...\beta_p)^T \\ \mathrm{b}_i \sim N(0, D) \\ \epsilon_i \sim N(0, \sigma^2 I_{n_i}) \\ b_i \text{ independent from } \epsilon_i \end{cases}$$

with X design matrix for fixed effects, Z design matrix for random effects and I_{n_i} is the order n_i identity matrix, where n_i denotes the number of observations in the i-th subject (cluster). Parameters interpretation comes straightforward: β_j with $j = 1, ..., p$ are interpreted as the change in the average y_i when x_j is increased by one unit; b_i express how a subset of the regression parameters for the i-th subject deviates from those in the population, so the sum $\beta_j + b_i$ describes the individual response. Models as the one above are already an extension of the simpler random intercept model, they allow intercepts variation across groups but also a random shift in the subject-specific slopes. When the chosen random-effects structure is not sufficient to capture the correlation in the data it is possible to change the model allowing for a more general, covariance matrix for the subject-specific error components, i.e. $\varepsilon_i \sim N(0, \Sigma_i)$, with Σ_i depending on i only through its dimensions n_i. In the literature, several different models have been proposed for different types of correlation functions. Some of the most frequently used are the first order Autoregressive, Exponential, and Gaussian correlation structures, but many more options are provided by standard statistical software. We assume that longitudinal responses of an individual are independent conditionally on its random effects:

$$p(y_i|b_i; \theta) = \prod_{j=1}^{n_i} p(y_{ij}|b_i; \theta), \tag{2.3}$$

using maximum likelihood principles we can derive the log-likelihood for the set of parameters

$$l(\theta) = \sum_{i=1}^{n} logp(y_i; \theta) = \sum_{i=1}^{n} log \int p(y_i|b_i; \beta, \sigma^2)p(b_i; \theta_b)db_i. \tag{2.4}$$

Given as known the covariance for the random part $cov(Z_i b_i + \epsilon_i;) = V_i$ and

$$p(y_i; \theta) = (2\pi)^{-n_i/2}|V_i|^{-1/2} \exp\left\{-\frac{1}{2}(y_i - X_i\beta)^T V_i^{-1}(y_i - X_i\beta)\right\}, \tag{2.5}$$

the fixed-effects estimator is obtained by maximizing the function above conditionally on the parameters in V_i and correspond to the generalized least square estimator:

$$\hat{\beta}_{GLS} = \left(\sum_{i=1}^{n} X_i^T V_i X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i^T V_i y_i\right). \tag{2.6}$$

Concerning random-effects we can no longer talk about estimation, because we are dealing now with random variable and we have to predict them. Henderson's mixed model equations are used in practice to obtain best linear unbiased predictors (BLUPs) for b [Henderson et al., 2000] and to calculate the best linear unbiased estimator for $X\beta$ at the same time:

$$\hat{b} = DZ^T V^{-1}(y - X\beta). \tag{2.7}$$

If V_i is unknown, we can replace it by its maximum likelihood estimate $\hat{V}_i$ asymptotically unbiased, otherwise restricted maximum likelihood (REML) can be applied to address a more general situation in small samples. This method estimates the variance components based on the residuals obtained after the estimation of the fixed effects $(y - X\beta)$. Neither of those methods however have a close form, so in order to obtain numerical optimization approaches are needed, such as the Expectation-Maximization [Dempster et al., 1977] or the Newton-Raphson algorithms [Lange, 2004].

2.1.2 Using splines to model non linear associations

In using maximum likelihood for simultaneous estimation of the parameters the form of the design matrix X is explicitly involved. One consequence of this is that if we use the wrong form for X, we may not even get consistent estimators for the parameters of interest.

A linear model, with its respective X matrix, assumes by definition a linear relationship between outcome and covariates, but it is not unusual that association between the outcome and covariate varies across covariates. To handle non linear relationships it is necessary to incorporate the concept of *smoothing* in the framework of mixed models. In the following we literature concerning the use of splines as accurately described in Gurrin et al. [2005]. The relationship between a continuous response Y and a single covariate x can be modeled by

$$\mathbb{E}[Y_i] = f(x_i) + \epsilon_i, \tag{2.8}$$

with $(i = 1, \ldots, n)$ and $f(\cdot)$ as an arbitrary smooth function giving the conditional mean of Y, ϵ_i are independent error random variables with mean zero and variance σ^2_ϵ. In order to estimate f in this non-parametric regression model we can use a spline estimator of the form

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} b_k (x - K_k)_+ \tag{2.9}$$

where

$$(x - K_k)_+ = \begin{cases} 0, if\ x \leq K_k \\ (x - K_k), if\ x > K_k \end{cases}$$

and $K_1, ..., K_k$ are knots, this equation describes a sequence of line segments linked together at the knots, to form a continuous function; it can be extended to take the

form of a piecewise polynomial of degree p:

$$f_q(x;\beta,b) = \beta_0 + \beta_1 x + ... + \beta_1 x^q + \sum_{k=1}^{K} b_k (x - K_k)_+^q \tag{2.11}$$

where $\beta = (\beta_0, ..., \beta_q)^T$ and $b = (b_1, \ldots, b_k)^T$ denote the vectors of coefficients and $[1, x, \ldots, x^q, (x - K_k)_+^q]$ denote *basis* functions. Splines of order q, or degree $q-1$, are defined as linear combination of them.

Types of splines differ each other in the choice of knots and roughness penalization [Gurrin et al., 2005]. Natural cubic splines [Green and Silverman, 1993] are the most used ones; for further investigation please refer to Wood [2006]. Finding the right level of smoothing is critical in the modelling process, smoothing too much can lead to lose underlying temporal dynamics, while smoothing too little can lead to wrong conclusions [Berk, 2013].

For applying a spline smoothing procedure we have to estimate $\hat{\beta}$ and $\hat{b}$ coefficients. Let's take a set of n response $\mathbf{y} = (y_1, \ldots, y_n)$ and covariates $\mathbf{x} = (x_1, \ldots, x_n)$ and normally distributed errors and random effects b and define the $n \times 2$ fixed effects design matrix as $X = [\mathbb{1} \quad \mathbf{x}]$ and the $n \times K$ random effects matrix as $Z = [(\mathbf{x} - k_1\mathbb{1})_+, \ldots, (\mathbf{x} - k_K\mathbb{1})_+]$, a connection between mixed models and spline smoothing methods can be established by considering $\hat{\beta}$ and $\hat{b}$ as the estimators that minimize the so called penalized least squares (PLS) function

$$PLS(\beta, b) = ||y - X\beta - Zb||^2 + \frac{\sigma_\epsilon^2}{\sigma_b^2}||b||^2. \tag{2.12}$$

Penalization process consist in constraining the magnitude of the random effect coefficients in b not to grow too large; a penalty $\|b\|^2$ results, for example, from the Gaussian distribution assumption. The advantage for spline smoothers in linear mixed models is that the ratio of variance components $\sigma_\epsilon^2/\sigma_b^2$ can be selected using REML estimation [Gurrin et al., 2005].

We can rewrite the estimators as

$$(\hat{\beta}, \hat{b})^T = (C'C + \frac{\sigma_\epsilon^2}{\sigma_b^2} G)^{-1} Cy,$$

with $C = [X \quad Z]$ and $G = diag(0_p, 1, \ldots, 1)$, 0_p representing the p-dimensional zero vector where p is the dimension of the vector β of fixed regression coefficients. This equation can be recovered by substituting the covariance matrix for the random errors $R = \sigma_\epsilon^2 I$ and the one for the random effects $D = \sigma_b^2 I$ into the mixed model equations.

In terms of interpretability an important disadvantage arises when such elaborate non linear parameterization of the subject-specific mean structure of longitudinal submodel is assumed. In particular, when polynomials or splines are used to capture non linear subject-specific evolution, the random effects do not have a the usual straightforward interpretation.

2.2 Analysis of Time-to-Event Data

Survival time, or event time, is defined as the time accounted from an initial start point to the occurrence of the event of interest. Time to death or to treatment failures are often primary outcomes in clinical studies, for example in randomized trials that compare a new drug with placebo for its ability to maintain remission in patients. Time-to-event data differs from classical Generalized Linear Model (GLM) responses that only concern one outcome variable in that they consist of both a continuous variable (time to event or censoring time) and a binary variable (indicating whether the observed time is the event time or not), which is not covered by GLM. The most important characteristic that distinguishes the analysis of time-to-event outcomes from other areas in statistics is indeed censoring. By this we mean that the event time of interest is not fully observed for all subjects under study. Several definitions have been formalized to describe this phenomenon and to define its different characterizations with their specific assumptions. Censoring implies that standard tools, such as the sample average, the t-test, and linear regression cannot be used and inferences may be sensitive to misspecification of the distribution of the event times.

We can distinguish censoring types by means of:

- Location of the true event time with respect to the censoring time: Right, Left and Interval censoring. Left and right censoring are special cases of interval censoring, with the beginning of the interval at zero or the end at infinity, respectively. Right censoring occurs when a subject leaves the study before an event occurs, or the study ends before the event has been experienced.

- Probabilistic relation between the true event time the censoring time: Informative and Non-informative (or Random) censoring, similar to MNAR and MAR in missing values analysis.

Depending on the type of censoring mechanism, different inferential procedures should be followed. The majority of the literature has focused on methods that can handle right censored data because they are the most common encountered. Throughout the remaining part of the thesis we will be focusing on event times that may be be subject to right censoring.

In the following subsections we introduce the notation for key components and models in the context of survival analysis.

2.2.1 Key functions

Let T be a non-negative random variable representing the event time, this is said to be right-censored by C, censoring time, if T is not observed but the relation $T > C$ is known. We say that we observe $T^* = min(T, C)$ in the sense that observation time is the first event occurring between survival and censoring time. Event indicator $\delta = I(T \leq C)$ is zero when we have censoring and 1 when the event occurs.

Key quantities used in survival analysis are the Survival and the Hazard function:

$$S(t) = P(T > t) = 1 - F(t) = S(t) = \int_t^\infty f(x)dx, \tag{2.13}$$

$$h(t) = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt | T \geq t\}}{dt}. \tag{2.14}$$

$S(t)$ is defined as the probability that event will occur after a certain time point t while h(t) represents the instantaneous hazard rate or in easier words, it tells us how likely an individual will experience the event in the next time point, given that he has not experienced it previously. The following relations must be considered:

$$S(t) = \exp\left\{ - \int_0^t h(x)dx \right\}, \tag{2.15}$$

$$h(t) = \frac{f(t)}{S(t)} = -\frac{\partial log S(t)}{\partial t}. \tag{2.16}$$

Another main quantity that is usually taken into consideration is the cumulative hazard function

$$H(t) = \int_0^t h(x)dx = -logS(t).$$

The most well-known estimators of both functions are the Kaplan-Meier and the Nelson-Aalen estimator.

- Kaplan Meier estimator

$$\hat{S}_{KM}(t) = \begin{cases} 1, & t < t_1, \\ \prod_{t_j \leq t} \left(1 - \frac{d_j}{r_j}\right), & t > t_1, \end{cases} \tag{2.17}$$

  where we consider $0 < t_1 < ... < t_D$ as the distinct uncensored event times, r_j is the total number of individuals "at risk" prior to time t_j and d_j is the number of observed events at t_j. 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.

- Nelson Aalen estimator

$$\hat{H}(t) = \sum_{t_j \leq t} \frac{d_j}{r_j}. \tag{2.18}$$

In the context of joint modeling we will make use of counting process theory [Aalen, 1978]. Some notation is introduced here for preliminary understanding. Let $N_i(t) = I(T^* \leq t, \delta_i = 1)$ be the counting process representing the observed events by time t for subject i and $Y_i(t) = I(T^* \geq t)$ the at-risk process, equal to 1 if subject i is indeed considered at risk prior to time t. Nelson-Aalen estimate can be then rewritten as

$$\hat{H}(t) = \int_0^t \frac{\partial N(x)}{Y(x)}, \tag{2.19}$$

with $N(t) = \sum_{i=1}^{n} N_i(t)$ and $Y(t) = \sum_{i=1}^{n} Y_i(t)$. The Kaplan Meier estimate can now be alternatively computed as

$$\hat{S}_{KM}(t) = \prod_{x \leq t} 1 - d\hat{H}(x). \tag{2.20}$$

Alternatively to the above non-parametric estimators a parametric form for the distribution of the survival time could be assumed. In that case we use maximum likelihood theory for parameters estimation. Suppose that we have n units with lifetimes and unit i-th observed for a time t_i. If the unit died at t_i its contribution to the likelihood function is the density, written as the product of the survivor and hazard functions $L_i = f(t_i) = S(t_i)h(t_i)$, while if the unit is still alive at t_i we will have $L_i = S(t_i)$. Combining the information from the censored and uncensored observations, we obtain the likelihood function:

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta) = \prod_{i=1}^{n} p(T_i;\theta)_i^{\delta} S_i(T_i;\theta)^{1-\delta_i}, \tag{2.21}$$

or in terms of log-likelihood

$$l(\theta) = \sum_{i=1}^{n} \delta_i \log p(T_i;\theta) + (1-\delta_i) \log S_i(T_i;\theta) = \sum_{i=1}^{n} \delta_i \log h_i(T_i;\theta) - \int_0^{T_i} h_i(x;\theta)dx \tag{2.22}$$

Several iterative optimization procedures, such as the Newton-Raphson algorithm [Lange, 2004], can then be used to locate the maximum likelihood estimates $\hat{\theta}$.

2.2.2 Proportional Hazards Cox models

Up to this point we have considered a population, where the lifetimes of all units are governed by the same survival function. In reality, we may have survival models characterized by the presence of a vector of covariates or explanatory variables that may affect survival time and require us to consider the problem of modeling these effects. There are different approaches for survival regression. Accelerated Failure Time models, for example, assume that the effect of a covariate is to accelerate or decelerate the life course of a disease by some constant. These are essentially standard regression models applied to the log of survival time, and except for the fact that observations are censored don't imply new estimation problems. In this subsection we will focus on the more popular Cox Proportional Hazards (Cox PH) models. Cox PH model was originally formulated by Cox [1992] as:

$$h_i(t|Z_i) = h_0(t) \exp\left(\beta^T Z_i\right). \tag{2.23}$$

In the expression above, $Z_i = (Z_{i1}, ..., Z_{ip})$ is the vector of covariates, $\beta = (\beta_1, ..., \beta_p)$ is the vector of regression coefficients and $h_0(t) = h(t|0)$ is the baseline hazard or baseline risk function, corresponding to the hazard function of a subject with $\beta^T Z_i = 0$. If we rewrite the model in the log scale,

$$\log h_i(t|Z_i) = \log h_0(t) + (\beta_1 Z_{i1} + ... + \beta_p Z_{ip}), \tag{2.24}$$

we can easily derive interpretation for the regression coefficients; we state that β_j is the change in the log hazard at any time t when variable Z_{ij} is increased by one unit (if continuous) and all others are held constant. More generally we will say that $\exp(\beta_j)$ denotes the Hazard Ratio (HR) of subject i with covariate $(Z_i + 1)$ compared to subject i with covariate Z_i:

$$HR = \frac{h_i(t|Z_{i1} + \cdots + Z_{ij} + 1 + \cdots + Z_{ip})}{h_i(t|Z_{i1} + \cdots + Z_{ij} + \cdots + Z_{ip})} = \exp(\beta_j) \tag{2.25}$$

The main assumptions for Cox PH includes:

- The censoring time C and the event time T are conditional independent given the covariates.
- The survival curves for two strata must have hazard functions that are proportional over time.
- The distribution of censoring time is unrelated to the unknown parameters related to survival.

In the relative risk model above we are not specifying any distribution for T^*. The baseline hazard function would have a different form depending on the specific distribution for the survival time. If for example T^* follows a Weibull distribution, we would have that $h_0(t) = \phi\sigma_t t^{\sigma_t - 1}$. To simplify estimation without having to specify any formal expression for $h_0(\cdot)$ and considering only parameters of interest, Cox [1992] proposed the following Partial Likelihood definition for the parameters β:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\left(\beta^T Z_i\right)}{\sum_{j \in R_i} \exp\left(\beta^T Z_j\right)} \right]^{\delta_i}, \tag{2.26}$$

where R_i is the risk set at the time just prior to t; or equivalently expressed as Partial Log-Likelihood

$$pl(\beta) = \sum_{i=1}^{n} \delta_i \left[\beta^T Z_i - log \sum_{T_j \geq T_i} \exp(\beta^T Z_j) \right] \tag{2.27}$$

which is the Profile Likelihood obtained by maximizing the joint likelihood with respect to H_0 for a fixed β and assuming there are no ties in the observed event times. Procedure follows exactly in the same way as with the full likelihood. To obtain the maximum likelihood estimates the scores equations $\partial pl(\beta)/\partial\beta^T = 0$ have to be solved in order to find a solution for $\hat{\beta}$ that is asymptotically normally distributed with mean β_{true}, the true parameter vector, and variance $[I(\beta_{true})]^{-1}$, the inverse of the expected information matrix. Computing the expected value requires to know the censoring distribution, but since we don't have it here standard errors are typically estimated using the observed information $I(\hat{\beta})^{-1}$.

2.3 Joint Modeling motivation

Why then to create another set of models if we already have these well performed statistical tools available?

Let's take transplantation studies as example. Our interest is the relation intercoming between longitudinal measurements taken at the visit times and the probability to experience graft failure or death. A standard analysis in this contest tends to ignore the longitudinal information and to use only the last available measurement as a baseline covariate in the survival model. In this simple way we would be discarding a considerable amount of valuable information. An alternative straightforward approach could be to put all longitudinal measurements at any time as covariates in the time-to-event model, but this would require many additional degrees of freedom and of course could lead to multicollinearity regression issues. Time-dependent Cox models are the usual setting to incorporate information coming from a covariate changing over time. The hazard function is here expressed as

$$h_i(t|Z_i(t)) = h_0(t) \exp\big(\beta^T Z_i(t)\big), \tag{2.28}$$

where $Z_i(t)$ is the time-dependent covariate and $h_0(t)$ is again the unspecified baseline hazard function. This model requires by construction knowledge of the covariate process for all subjects in the risk set at the time of each failure. However in clinical trials it is common to have a marker measured at only discrete time points so that no measurements for the covariate would exist for those in the risk set when an event occurs in a middle time between scheduled follow-up visits. A possible solution for this is to carry forward the last longitudinal measurement preceding failure time and treat it as if it was the current value, but this would simply ignore measurement errors that characterize the covariate values [Tsiatis et al., 1995], affecting standard errors of the estimates of interest and causing the relative risk parameter to be biased towards zero [Prentice, 1982].

Time-dependent cox models also implicitly assume to deal with exogenous covariates, defined by Kalbfleisch and Prentice [2002] as those which path at any future time point t is not affected by the occurrence of an event at time $s < t$. They should satisfy the relation

$$\Pr(Y_i(t)|Y_i(s), T_i^* \geq s) = \Pr(Y_i(t)|Y_i(s), T_i^* = s). \tag{2.29}$$

This hypothesis holds for time-dependent variables such as environmental factors, while it is clearly unfulfilled by biomarkers, which value at any time is influenced by the occurrence of the event in the past. A longitudinal endogenous variable is therefore said to be informatively censored at the event time, with patients non-random dropouts; in this case the hazard function cannot be directly relation to the survival by the usual formula

$$S_i(t|Y_i(t)) = \exp\left\{ -\int_0^t h_i(s|Y_i(s))ds \right\}, \tag{2.30}$$

so the log-likelihood construction used before is no more appropriate.

Stated above are the main reasons why joint models have been introduced in statistical theory. First attempt method to overcome the disadvantages of the naive

methods was done by Tsiatis et al. [1995] who proposed a two-stage estimation analysis. This basically consists in retrieving the empirical Bayes estimates from the longitudinal model and plugging in the predictions into the survival one and eventually using Monte Carlo simulations to sample from the posterior distribution of the random effects, to account for the fact that we use estimates instead of true values.

Chapter 3

Joint Models Framework

Since the seminal paper Wulfsohn and Tsiatis [1997a], longitudinal covariates have played an increasingly important role in the modeling of survival data. Based on the theoretical background we just introduced we are now able to present the standard joint model, which incorporates all the basics from survival and longitudinal analysis. In this chapter we will start by introducing the general model formulation, in a second section we will discuss and compare two different estimation's methods: the full maximum likelihood approach and its ancestor, the two-stage approach; we will also briefly mention optimization and numerical integration algorithms and computational issues arising when dealing with complex models. Inference on model's parameters, their interpretation and possible diagnostics tools to judge model's appropriateness will follow. Additionally, we will try to include a complete list of model's extensions that cover situations very often encountered in the real world contest of analysis.

3.1 The General Model

The joint modeling approach postulates a relative risk model for the event-time outcome directly associated with the longitudinal process. Using similar notation as in 2 we will denote as T_i^* the true event time for the i-th subject and as $y_i(t)$ the endogenous time-dependent covariate values at time point t for the i-th subject. In real clinical studies, longitudinal marker values are rarely available at event time but rather at specific occasions t_{ij} denoted as follow-up times, actually observed measurements are denoted as $y_{ij} = y_i(t_{ij})$ with $j = 1, \ldots, n_i$ number of visits for each patients.

Joint models belong to a broaden class of models called shared random effects with a key underline assumption:

$$f(Y, T, b) = f(T|b)f(Y|b)f(b); \tag{3.1}$$

i.e. the event time responses are independent from the longitudinal response, conditionally on the random effects.

More precisely a joint likelihood for the two outcome is now formulated as product of two conditional independent distributions:

$$p(T_i, \delta_i, y_i|b_i; \theta) = p(T_i, \delta_i|b_i; \theta)p(y_i|b_i; \theta), \tag{3.2}$$

with

$$p(y_i|b_i;\theta) = \prod_j p(y_i(t_{ij})|b_i;\theta) \tag{3.3}$$

and $\theta = (\theta_t, \theta_y, \theta_b)^T$ the parameter vector for the event time variable, the longitudinal variable and for random effects covariance matrix, respectively. Also, intermittent missing data (censoring and visiting processes) are assumed non-informative and independent from true event times and future longitudinal measurements. Violation of the latter assumption would in fact imply dependence of subject's prognosis with latent characteristics.

The standard joint model set up links together a linear-mixed model for the longitudinal outcome with a cox proportional hazard model for the time-to-event outcome. The survival submodel is:

$$\begin{aligned} h_i(t|M_i(t), w_i) &= \lim_{dt\to 0} \Pr t \le T_i^* < t + dt | T_i^* \ge t, M_i(t), w_i/dt, \\ &= h_0(t) \exp\left\{\gamma^T w_i + \alpha(m_i(t))\right\}. \end{aligned} \tag{3.4}$$

where $M_i(t) = \{m_i(t), 0 < s < t\}$ denotes the history of the true (unobserved) longitudinal process up to time t with $m_i(t)$ being the true current value, w_i is a vector of baseline covariates and $h_0(t)$ the well known baseline hazard. From the above we can further determine the survival function by the relation

$$S_i(t) = \Pr(T_i^* > t | M_i(t), w_i) \tag{3.5}$$

$$= \exp\left\{ -\int_0^t h_0(s) \exp\{\gamma^T w_i + \alpha(m_i(s))\} ds \right\}, \tag{3.6}$$

which implies that patient's survival depend on the entire history of the covariate. In classical survival analysis the usual practice is to leave the baseline hazard function $h_0(t)$ completely unspecified, this avoids the restriction resulting from specifying a certain form for the baseline hazard and at the same time still can offer valid statistical inference through the use of partial likelihood. In the joint models context this choice will generally lead to underestimation of the standard errors of the model parameters ([Yuen and Mackinnon, 2016],[Hsieh et al., 2006]). The risk function can be set as known parametric distributions such as the Weibull, Log-Normal, Gamma or alternatively to more flexible non parametric distributions as step-functions and linear splines. Two common options often encountered are in fact:

- the Piecewise-constant model, where $h_0(t) = \sum_{q=1}^{Q} \xi_q I(v_{q-1} < t \le v_q)$, with $0 = v_0 < v_1 < \cdots < v_Q$ denoting a partition of the time scale (v_Q larger than the largest observed time) and ξ_q is the value of the hazard in the interval $(v_{q-1}, v_q]$.

- the Regression splines model, where $log(h_0(t)) = k_0 + \sum_{d=1}^{m} k_d B_d(t, q)$, with $k = (k_0, \ldots, k_m)$ denoting the spline coefficients and q the degree of the B-spline basis functions. In the general model we have to keep in mind that a standard rule of thumb is to retain a total number of parameter between 1/10 and 1/20 of the number of events in the sample, in order to avoid over-fitting, therefore choosing a baseline hazard spline formulation with too many degrees of freedom could be inappropriate [Rizopoulos, 2012].

Once determined the submodel for the survival data (Eq.3.4) we formulate the longitudinal submodel as:

$$\begin{aligned} y_i(t) &= x_i^T(t)\beta + z_i^T(t)b_i + \epsilon_i(t) \\ &= m_i(t) + \epsilon_i(t), \end{aligned} \tag{3.7}$$

with $b_i \sim N(0, D)$ and $\epsilon_i(t) \sim N(0, \sigma I_{n_i})$. In the above we notice that the design vectors for the fixed and random effects $x_i(t)$ and $z_i(t)$ are time-dependent, furthermore the same assumptions for LMM models we saw in 2 are still valid. The model highlight the idea to decompose the longitudinal outcome for each patient at event time between true level $m_i(t)$ and error term $\epsilon_i(t)$. This is in fact the main improvement of joint models over extended cox models, where the estimation error was not accounted for the longitudinal process.

3.2 Estimation: Two stages approach

The main idea of the beginning version of joint models was to build the likelihood of the above mentioned models separately. A two stages approach has the advantage to be quick and relatively easy to implement with standard software, but on the other hand has been widely demonstrated how it leads to less efficient estimates. This procedure is also denoted as ordinary regression calibration (ORC) and works as follows.

At Stage I we obtain $\hat{\theta}_y$ maximizing the log-likelihood:

$$l_y(\theta_y) = \sum_i^n p(Y_i|b_i;\theta)p(b_i|\theta),$$

this requires numerical integration as Gaussian quadrature rules for example, we then obtain the corresponding empirical Bayes estimates

$$\hat{b}_i = argmax_\theta \big\{ log(p(Y_i|b_i;\theta)) + log(p(b_i|\theta)) \big\}$$

and compute the predicted values $\hat{y}_i = x_i\hat{\beta} + z_i\hat{b}$.

At Stage II we fit the relative risk model plugging in the fitted values $\hat{y}_i(t)$ as time-dependent covariates and maximize the partial likelihood to get an estimate for γ. A remarkable disadvantage here is that it we do not correct for event-dependent drop-out, and uncertainty in the estimated MLEs and BLUPs (best linear unbiased predictors) is not carried forward to the survival model, resulting in estimates that are too precise. Also, the form of the BLUPs depends critically on the validity of normally distributed random effects and error terms, which becomes less satisfactory as time increases and subjects suffer informative drop-out [Tsiatis and Davidian, 2004].

3.3 Estimation: Full Likelihood approach

Both Bayesian and Frequentist approaches could be applied for the estimation of the joint likelihood, however our focus is on the latter one. This choice doesn't intend to

underestimate the importance of Bayesian techniques when dealing with complicated situations as for example high-dimensional random effects, where it can be worth to consider Monte Carlo sampling methods for numerical integration instead of Gaussian quadrature or Laplace Approximation. The methods just mentioned will be here investigated.

Under the assumptions presented in previous sections and under conditional independence of the two outcomes given the underline random effects, the joint-likelihood contribution for the i-th subject can be formulated as follows

$$\begin{aligned} logp(T_i, \delta_i, y_i|b_i; \theta) &= log \int p(T_i, \delta_i, y_i, b_i; \theta) db_i \\ &= log \int p(T_i, \delta_i|b_i; \theta_t, \beta) db_i \Big[\prod_j p(y_i(t_{ij})|b_i, \theta_y) \Big] p(b_i; \theta_b) db_i. \end{aligned} \tag{3.8}$$

The three components in the equation above are:

- the likelihood for the survival part

$$\begin{aligned} p(T_i, \delta_i|b_i; \theta_t, \beta) &= h_i(T_i|M_i(t); \theta)^{\delta_i} S_i(T_i|M_i(t); \theta) \\ &= \Big(h_0(T_i) \exp\{\gamma^T w_i + \alpha m_i(T_i)\} \Big)^{\delta_i} \\ &\times \exp\Big(- \int_0^{T_i} h_0(s) \exp\big\{\gamma^T w_i + \alpha m_i(T_i)\big\} ds \Big), \end{aligned}$$

- the joint density for the longitudinal responses and random effects

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

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

To maximize the log-likelihood function $l(\theta) = \sum_i logp(T_i, \delta_i, y_i|b_i; \theta)$ for all the observed data with respect to θ, Wulfsohn and Tsiatis [1997a] have proposed a two steps iterative procedure.

3.3.1 EM algorithm

The purpose of the EM algorithm is to estimate parameters of interest by maximizing the likelihood of the observed data. This is done by iterating between an E-step, where we compute the expected log-likelihood of the complete data conditional on the observed data and the current estimate of the parameters, and an M-step, where new parameter estimates are computed by maximizing this expected log-likelihood.

- **E-step**: estimates parameters $\theta = (\theta_t, \theta_y, \theta_b)^T$ of the complete data Y^o and Y^m by using only observed data, we compute the expected value of the observed data log−likelihood

$$Q(\theta|\theta^{(it)}) = \mathbb{E}\{logp(y;\theta)|y^o;\theta^{(it)}\} = \int logp(y^m, y^o;\theta)p(y^m|y^o;\theta^{(it)})dy^m. \tag{3.10}$$

  Referring to the joint likelihood formulation, in particular, we want to maximize $l(\theta)$ by maximizing

$$\begin{aligned} Q(\theta|\theta^{(it)}) &= \sum_i \int logp(T_i, \delta_i, y_i, b_i;\theta)p(b_i|T_i, \delta_i, y_i;\theta^{(it)})db_i \\ &= \sum_i \int \left\{logp(T_i, \delta_i|b_i;\theta_t, \beta) + logp(y_i|b_i;\theta_y) + logp(b_i;\theta_b)\right\} \\ &\times p(b_i|T_i, \delta_i, y_i;\theta^{(it)})db_i, \end{aligned} \tag{3.11}$$

  treating random effects as missing data.

- **M-step**: we obtain the updated parameters $\theta^{(it+1)}$ by maximizing the expected value first computed $\theta^{it+1} = argmax_\theta Q(\theta|\theta^{it})$.

  This imply splitting the complete data log-likelihood into three parts

$$logp(T_i, \delta_i, y_i, b_i;\theta) = logp(T_i, \delta_i|b_i;\theta_t, \beta) + logp(y_i|b_i;\theta_y) + logp(b_i;\theta_y) \tag{3.12}$$

  with maximization that therefore involves for each parameter only the pieces where it appears.

Closed forms are available for the variance of residuals of the longitudinal model and the variance-covariance matrix of the random effect:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\sum_i \int (y_i - X_i\beta - Z_i b_i)^T (y_i - X_i\beta - Z_i b_i)p(b_i|T_i, \delta_i, y_i;\theta)db_i}{\sum_{i=1} n_i} \\ &= \frac{\sum_i (y_i - X_i\beta)^T (y_i - X_i\beta - 2Z_i\bar{b_i}) + tr(Z_i^T Z_i Var(b_i|T_i, \delta_i, y_i;\theta)) + \bar{b_i}^T Z_i^T Z_i \bar{b_i}}{\sum_{i=1} n_i}, \end{aligned}$$

where tr denote the trace of a matrix and $\bar{b_i} = \mathbb{E}$ the expectation function.

The estimate for the covariance matrix of the random-effects is

$$\hat{D} = \frac{\sum_i (b_i|T_i, \delta_i, y_i;\theta)\bar{b_i} + \bar{b_i}\bar{b_i}^T}{n}.$$

However since fixed effect coefficients are involved in both the longitudinal and survival models no closed form solution can be found for them, therefore a Newton-Raphson updating scheme is implemented for both:

$$\hat{\beta}^{(it+1)} = \beta^{(it)} - \left\{\partial S(\hat{\beta}^{(it)})/\partial\beta\right\}^{-1} S(\hat{\beta}^{(it)})$$

with

$$\begin{aligned} S(\beta) &= \sum_i X_i^T \{y_i - X_i\beta - Z_i\ b_i\}/\sigma^2 + \alpha\delta_i x_i(T_i) \\ &\quad - \exp(\gamma^T w_i)\Big[\iint_0^{T_i} h_0(s)\alpha x_i(s)\exp\left(\alpha x_i^T(s)\beta + z_i^T(s)b_i\right) \\ &\quad \times p(b_i|T_i,\delta_i,y_i;\theta)dsdb_i\Big]. \end{aligned}$$

This also applies to the parameters of the survival model:

$$\hat{\theta}_t^{(it+1)} = \theta_t^{(it)} - \{\partial S(\hat{\theta}_t^{(it)})/\partial\theta_t\}^{-1} S(\hat{\theta}_t^{(it)})$$

with

$$\begin{aligned} S(\gamma) &= \sum_i w_i^T\Big[\delta_i - \exp(\gamma^T w_i)\iint_0^{T_i} h_0(s)\exp\left(\alpha x_i^T(s)\beta + z_i^T(s)b_i\right) \\ &\quad \times p(b_i|T_i,\delta_i,y_i;\theta)ds\ \ db_i\Big]; \end{aligned}$$

$$\begin{aligned} S(\alpha) &= \sum_i \delta_i x_i^T(T_i)\beta + z_i^T(T_i)b_i \\ &\quad - \exp(\gamma^T w_i)\Big[\iint_0^{T_i} h_0(s)\exp\left(\alpha x_i^T(s)\beta + z_i^T(s)b_i\right) \\ &\quad \times p(b_i|T_i,\delta_i,y_i;\theta)ds\ \ db_i\Big]; \end{aligned}$$

$$\begin{aligned} S(\theta_{h_0}) &= \sum_i \delta_i \partial h_0(T_i;\theta_{h_0}/\partial\theta_{h_0}^T \\ &\quad - \exp(\gamma^T w_i)\Big[\iint_0^{T_i} \partial h_0(T_i;\theta_{h_0}/\partial\theta_{h_0}^T\exp\left(\alpha x_i^T(s)\beta + z_i^T(s)b_i\right) \\ &\quad \times p(b_i|T_i,\delta_i,y_i;\theta)ds\ \ db_i\Big]. \end{aligned}$$

The EM algorithm essentially treats random effects as missing values. Once derived parameters θ for the joint model, patient-specific trajectories b_i (random variables) can be predicted using the Bayesian paradigm. Assuming $p(b_i,\theta)$ to be the prior distribution, and $p(T_i,\delta_i|b_i;\theta)p(y_i|b_i;\theta)$ the conditional likelihood part we formulate their posterior distribution as

$$p(b_i|T_i,\delta_i,y_i;\theta) \propto p(T_i,\delta_i|b_i;\theta)p(y_i|b_i;\theta)p(b_i,\theta). \tag{3.13}$$

An empirical Bayes approach is then used to estimate the posterior mean and mode that describes the posterior distribution stated above:

- $\bar{b}_i = \int b_i p(b_i|T_i,\delta_i,y_i;\theta)db_i$,
- $\hat{b}_i = argmax_b\{logp(b|T_i,\delta_i,y_i;\theta)$.

3.3.2 Numerical Integration

We briefly introduce numerical integration techniques used in joint models to approximate intractable integrals. Standard and adaptive Gauss Hermite quadrature rules are usually applied. These consists in approximating the integral in the definition of the score vector by a weighted sum of integral evaluations at prespecified abscissas [Rizopoulos, 2012]. First of all we notice that the score function for the model's parameters can be written as

$$S(\theta) = \sum_i \int \partial\{logp(T_i, \delta_i|b_i;\theta) + logp(y_i|b_i;\theta) + logp(b_i;\theta)\}/\partial\theta^T p(b_i|T_i, \delta_i, y_i;\theta)db_i \tag{3.14}$$

$$= \sum_i \int A(\theta, b_i)p(b_i|T_i, \delta_i, y_i;\theta)db_i, \tag{3.15}$$

with $A(\cdot)$ denoting the complete data score vector. For any form of the A() function of the random effects, the integral in the definition of the score vector can be approximated by a weighted sum of integral evaluations at prespecified abscissas

$$\mathbf{E}\{A(\theta, b_i)|T_i, \delta_i, y_i;\theta\} \approx 2^{q_b/2} \sum_{t_1,\ldots,t_q} \pi_t A(\theta, b_t\sqrt{2})p(b_t\sqrt{2}|T_i, \delta_i, y_i;\theta)exp(||b_t||^2),$$

where $\sum_{t_1,\ldots,t_q} = \sum_{t_1=1}^{K} \cdots \sum_{t_q=1}^{K}$, K denoting the number of quadrature points and $b_t^T = (b_{t1}, ..., b_{tq})$ the abscissas with corresponding weights π_t. Accuracy of the approximation improves as K is increased while computational cost exponentially increases with q_b (dimension of random effects) but also the locations of the quadrature points with respect to the location of the main mass of the integral could be critical. If $A(\theta, b_i)p(b_i|T_i, \delta_i, y_i;\theta)$ is concentrated far from zero or its width is quite different from the weight function $exp(||b||^2)$, an adaptive procedure is applied.

The new approximation will be

$$\begin{aligned}\mathbf{E}\{A(\theta, b_i)|T_i, \delta_i, y_i;\theta\} \approx\ & 2^{q_b/2}|\hat{B}_i|^{-1} \sum_{t_1,\ldots,t_q} \pi_t A(\theta, \hat{b}_i\sqrt{2}\hat{B}_i^{-1}b_t) \\ & \times p(b_t\sqrt{2}|T_i, \delta_i, y_i;\theta)exp(||b_t||^2),\end{aligned}$$

where $\hat{b}_i = argmax_b\{logp(T_i, \delta_i, y_i, b;\theta)\}$ is the mode of the random effects and $\hat{B}_i$ is the Choleski factor of the estimated hessian matrix $\hat{H}_i$.

3.4 Inference

Model testing on the null hypothesis on $H_0 : \theta = \theta_0$ versus $H_A : \theta \neq \theta_0$ can involve three different statistics

- the Likelihood Ratio Test: LRT $= -2\{l(\hat{\theta}_0) - l(\hat{\theta})\}$,
  where $\hat{\theta}_0$ and $\hat{\theta}$ are the maximum likelihood estimates under the null and alternative hypothesis.

- the Score Test: $\mathrm{U} = S^T(\hat{\theta}_0)I(\hat{\theta}_0)^{-1}S(\hat{\theta}_0)$,
  with $I(\cdot)$ denoting the observed information matrix of the model under the alternative hypothesis.
- the Wald Test: $\mathrm{W} = (\hat{\theta} - \theta_0)^T I(\hat{\theta})(\hat{\theta} - \theta_0)$.

All of them, under the null hypothesis, follow asymptotically a chi-squared distribution, with p (number of parameters to be tested) degrees of freedom. The likelihood ratio test is more computationally expensive because it requires the model to be fitted under both hypotheses, on the other hand the Wald test does not take into account the variability introduced by estimating the variance components.

These tests are only appropriate for the comparison of two nested models but to carry out the comparison of non-nested models, information criteria can be used, such as the Akaike's Information Criterion (AIC, [Akaike, 1974]), and the Bayesian Information Criterion (BIC, [Schwarz, 1978]):
$\mathrm{AIC} = -2(l(\hat{\theta})) + 2p$,
$\mathrm{BIC} = -2(l(\hat{\theta})) + plog(n)$.

3.5 Model Diagnostics

When fitting a regression model it is always important to determine whether all the model assumptions are valid. Residual graphical methods are a standard tool to perform model diagnostics that has being intensively studied for longitudinal and survival analysis. Standard model diagnostics for mixed effects and relative risk models can be used as for the complete data set:

- Residuals for the longitudinal part Standardized marginal residuals
$$r_i^{ym} = \hat{V}_i^{-1/2}(y_i - X_i\hat{\beta}), \tag{3.16}$$
where $\hat{V}_i^{-1/2} = Z_i bDZ_i^T + \hat{\sigma}^2 I_{n_i}$ represents the marginal covariance matrix of y_i, can be used to investigate miss-specification of the mean structure $X_i\beta$ and to check assumptions about V_i within-subjects covariance matrix. Standardized subject specific residuals
$$r_i^{ys}(t_{ij}) = \{y_i(t_{ij}) - x_i^T(t_{ij})\beta - z_i^T(t_{ij})\hat{b}_i\}/\hat{\sigma}, \tag{3.17}$$
instead can be used to validate homosckedasticity and normality assumptions.
- Residuals for the survival part Martingale residuals defined as
$$r_i^{tm} = \delta_i - \int_0^{T_i} h_i(s|\hat{M}(s);\hat{\theta})ds, \tag{3.18}$$
are frequently used to verify whether functional forms of the covariates are appropriate. On the other hand Cox-Snell residuals in the form
$$r_i^{tcs} = \int_0^{T_i} h_i(s|\hat{M}(s);\hat{\theta})ds = \int_0^{T_i} \hat{h}_0(s)\exp\{\gamma^T w_i + \hat{\alpha}\hat{m}_i(s)\}ds, \tag{3.19}$$

represent the values of the estimated cumulative risk function ($H(t) = -log(S(t))$) at the observed event times T_i and can be used to assess the fit of the relative risk model. By definition r_i^{tcs} follow a unit exponential distribution however judging the model from their distribution can be misleading without keeping in mind that when T_i are censored they are censored too. It would be helpful instead to compare the unit exponential distribution of the survival function $S_{exp}(t) = -exp(t)$ with the Kaplan-Meier curve to see if they match.

In the joint modeling framework it is assumed that the occurrence of events is related with the underlying evolution of the subjects specific longitudinal profiles, which corresponds to a non-random dropout mechanism (MNAR). Residual plots can be misleading because patients that dropped out may have different longitudinal evolution than patients who do not, in other words observed data are not a random sample of the target population. Rizopoulos [2012] proposed an interesting method for calculating residuals and producing diagnostic plots in joint models, creating random versions of the completed data set by multiple imputation of the missing longitudinal responses under the fitted model, allowing to analyze possible trends. Assuming longitudinal measurements scheduled at prefixed visit times, for each subject under study we have observations up to the last visit time before T_i. Multiple imputation is carried with a Bayesian approach, by repeated sampling from the posterior distribution of missing data given observed data:

$$\begin{aligned} p(y_i^m|y_i^o, T_i, \delta_i; \theta) &= \int p(y_i^m|b_i, y_i^o, T_i, \delta_i; \theta) p(b_i|y_i^o, T_i, \delta_i; \theta) db_i \\ &= \int p(y_i^m|b_i; \theta)(b_i|y_i^o, y_i^m; \theta) db_i. \end{aligned} \tag{3.20}$$

When n is sufficient large the posterior of the parameters can be approximated by a normal distribution and derived by the following scheme:

1. Draw $\theta^{(l)} \sim N(\hat{\theta}, \hat{v}ar(\theta))$.
2. Draw $b_i^{(l)} \sim b_i|y_i^o, T_i, \delta_i, \theta^{(l)}$.
3. Draw $y_i^{m(l)}(t_{ij}) \sim N(\hat{m}_i^{(l)}(t_{ij}), \sigma^{2(l)})$, for visit times $t_{ij \geq T_i}$ not observed for the current individual.

Each step is repeated for a total number of iterations, simulated $y_i^{m(l)}(t_{ij})$ together with y_i^o can now be used to calculate imputed residuals.

3.6 Model Extensions

Starting from the basic structure of a joint model we could build a more flexible parameterization in different ways.

3.6.1 Association structures

Classical extensions proposed by Rizopoulos [2012] involve the inclusion of features of the longitudinal covariate history that were not captured by only considering its current value at each event time point t, for example:

- the current slope of the marker as $h_o(t)\exp\{\gamma^T w_i + \alpha m_i(t)'\}$ to capture situations where at a specific time point patients show similar true marker levels, but differ in the rate of change of the marker
- the marker with a time lag Δt as $h_o(t)\exp\{\gamma^T w_i + \alpha m_i(t-\Delta t)\}$
- the cumulative effect of the marker as $h_o(t)\exp\{\gamma^T w_i + \alpha \int_0^t m_i(s)\}$ where the area under the longitudinal trajectory up to each event time is regarded as a summary measure of the whole marker history.
- a possible interaction with baseline covariates $h_o(t)\exp\{x_{1i}^T\beta + \alpha(x_{2i}\cdot m_i(t))\}$
- only random effects as $h_o(t)\exp\{\gamma^T w_i + \alpha^T b_i\}$ that, except for the baseline hazard, results in a time-constant risk model and therefore can be used to facilitate estimation [Barrett et al., 2015] with the drawback that the association parameter is not interpretable in case of a flexible spline parameterization of individual trajectories.

The association structure can be chosen at priori, based on problem knowledge, or after fitting the model using AIC and BIC criteria. In a Bayesian setting approaches as Bayesian model averaging and Bayesian shrinkage have been developed to include a combination of different associations or to include different association structures and achieve a parsimonious model by placing shrinkage priors on the association parameters.

3.6.2 Submodels structures

The standard joint model set up can also be extended with regard to the two specific submodels. In many situations, for example, it can be necessary to relax the normality assumption of the random effects [Tang et al., 2017] or to use a different longitudinal response distribution (e.g for a categorical marker) substituting the LMM with a more general GLMM model (see chapter 2). For a GLMM longitudinal submodel we specify by $y_i = y_{ij}, j = 1, ..., n_i$ the vector of observed longitudinal responses for the i-th subject and denote the probability density function in the exponential family form, expressing the conditional mean by a general monotonic function $g(\cdot)$ that works as link to the linear predictor.

The joint model is therefore formulated as:

$$\begin{cases} \mathrm{p}(\mathrm{y}_i|b_i;\beta,\phi) = \exp\{\sum_{j=1}^{n_i}\left[y_{ij}\psi_{ij}(b_i) - c\{\psi_{ij}(b_i)\}\right]/a(\phi) - d(y_{ij},\phi)\}, \\ \mathrm{m}_i(t) = \mathbb{E}(y_i(t)|b_i) = g^{-1}(x_i(t)^\beta + z_i(t)_i^b), \\ \mathrm{h}_i(t) = h_0(t)\exp\left\{\gamma^T w_{1i} + f(m_i(t-c), b_i, w_{2i};\alpha)\right\}, \end{cases}$$

where $\psi_{ij}(b_i)$ and ϕ denote the natural and dispersion parameters, $c(\cdot)$, $d(\cdot)$ and $a(\cdot)$ are functions for the members of the exponential family, such as Binomial, Poisson, Gamma, and normal distributions, $b_i \sim N(0, D)$ and α now measures the strength

of the association between the risk for an event at time t and the expected value of the longitudinal outcome at the same time point.

Furthermore it is not uncommon that scientists want to investigate more than one biomarker in association with the same event outcome, a multivariate longitudinal model should be formulated in these cases; to reduce computational burden in this contest Proust-Lima et al. [2009] proposed a multivariate joint model where the longitudinal outcomes are considered as realizations of a single latent process that represents the common unobserved factor that drives the observed longitudinal trajectories. Different association structures are also allowed in multivariate models, we refer to Hickey et al. [2016] for an extensive review of possible multivariate association structures.

Finally, quantile regression joint models have been also developed to assess the association between quantiles of the longitudinal profile with the hazard [Farcomeni and Viviani, 2015].

With regard to the relative risk model, potential extensions could be considered to account for interval-censored outcomes, recurrent events and competing risks. In the competing risk setting, for example, if we assume K different causes of failure, a survival model for each of the causes is postulated ($h_{ik}(t)$, $k = 1, \ldots, K$) and the likelihood of the event process is constructed as the product of the single likelihoods:

$$p(T_i, \delta_i | b_i; \theta) = \prod_{k=1}^{K} \left[h_{0k}(T_i) \exp\{\gamma_k^T w_i + \alpha_k m_i(T_i)\} \right]^{I(\delta_i = k)} \times \exp\left(\sum_{k=1}^{K} \int_0^{T_i} h_{0k}(s) \exp\{\gamma_k^T w_i + \alpha_k m_i(s)\} ds \right).$$

This is done in practice by previous transforming the dataset, each patients have to be represented by a number of rows equal the number of causes, creating a stratification factor for the competing risks variable and a binary status variable equal to 1 if the corresponding event occurred.

When the proportional hazard assumption fails accelerated failure models (AFT) can be applied as alternative to Cox models. In this framework the effect of covariates is specified as additive on the the log failure time:

$$log T_i^* = \gamma^T w_i + \sigma_t \epsilon_{ti},$$

where σ_t is a scale parameter and $_{ti}$ can be assumed to follow a normal, Student's-t or extreme value distribution. The subject's risk rate function can be re-expressed as

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

with $V_i(t) = \int_0^t \exp\{\gamma^T w_i + \alpha(m_i(t))\} ds$. In contrast to the classical model here $h_0(\cdot)$ is evaluated at $V_i(t)$, so that the entire covariate history $M_i(t)$ is assumed to influence the subject-specific hazard.

Besides the flexibility of the longitudinal model we have seen above a generalization of the association is required when we want to consider a time-varying relationship between the biomarker and the time-to-event. The most flexible framework for joint models allowing for flexible longitudinal trajectories and potentially

nonlinear time-varying association modeled by penalized splines has been only recently developed [Köhler et al.]. The very general setup the hazard of an event at t as

$$h_i(t) = \exp(\eta_i(t)) = \exp\left\{\eta_{\lambda i}(t) + \eta_{\gamma i} + \eta_{\alpha i}(t) \cdot \eta_{\mu i}(t)\right\},$$

with the full predictor η including a predictor η_λ for all time-varying survival covariates or those with a time-varying coefficient (also the log baseline hazard), a predictor for baseline survival covariates η_γ, a predictor η_α for potentially time-dependent association between the hazard and the longitudinal marker η_μ, the latter is allowed to follow a non parametric distribution eventually. Identifiability problems occur if constraints are not assumed in the additive structure of the model. All nonlinear terms are imposed to sum to zero over all observations for predictors in both submodels, for B-splines [De Boor et al., 1978] the basis matrix X_{km} is transformed into a $n \times (p_{km} - 1)$ matrix for which it holds $X_{km}\mathbb{1}_{\mathbb{k}-\mathbb{1}} = 0$ [Wood, 2006] and adjusting the penalty.

Estimation of joint models with such complex structure becomes very challenge with a frequentist estimation approaches due to the necessary integration over potentially high-dimensional random effects distributions, Bayesian estimation is therefore often employed (R packages **bamlss** and **JMbayes**).

Chapter 4

Dynamic Predictions

Accuracy of risk assessment for prevention and treatment of different diseases is required in clinical practice. Doctors make decisions regarding treatments, tests or alternative therapies based on risk scores. These risk scores are mostly influenced by common measured variables as age, sex, BMI, smoking habit, genetic components, biomarkers.

For statisticians the goal becomes to give updated estimates of survival probabilities for a new patient as long as additional information is recorded for that subject. In other words the question is: "what is the likelihood of developing an adverse outcome among individuals who survived up to a specific time and given the available extra information up to that time?".

An early approach for solving this question has been landmarking [van Houwelingen and Putter, 2011]. Landmark analysis consists in extrapolating survival probabilities from a Cox model fitted to the patients from the original dataset who are still at risk at a specific time point called "landmark time". A relatively newer method to produce dynamic predictions for survival probabilities is based instead, on the class of joint models for longitudinal and time-to-event data. In aim of this chapter is to explore predictions models in their dynamic version, comparing and judging the available techniques and to provide guidance on model choices and performance evaluation by mean of prediction criteria for discrimination and calibration.

4.1 What are individual dynamic predictions?

Let $D_n = \{T_i, \delta_i, y_i; i = 1, ..., n\}$ denote a sample from the target population, where T_i indicates again the observed event time and y_i the longitudinal outcome measured at time t_{ij} for the i-th subject. Our interest in this case is to derive predictions for a new subject i from the same population on whom a set of longitudinal measurements $Y_i(t) = \{y_i(t_{ij}); 0 \leq t_{ij} \leq t, j = 1, ..., n_i\}$ have been observed. In the context of *endogenous* covariates, it's important to highlight that a measurement recorded at time t implies survival of the patient up to this time point, this is the reason why we focus on conditional subject-specific probability of surviving time $u > t$ given survival up to t:

$$\pi_i^{JM}(u|t) = \Pr(T_i^* \geq u | T_i^* > t, Y_i(t), w_i, D_n; \theta_{true}), \tag{4.1}$$

where w_i are baseline covariates and θ_{true} true parameters. Making use of the conditional independence assumption, between the longitudinal and survival outcome, and omitting w_i for easy of notation we can rewrite the right hand side of the equation as

$$\begin{aligned}\Pr(T_i^* \geq u | T_i^* > t, Y_i(t), D_n; \theta) &= \int \Pr(T_i^* \geq u | T_i^* > t, Y_i(t), b_i; \theta) \\ &\times p(b_i | T_I^* > t, Y_i(t); \theta) db_i \\ &= \int \Pr(T_I^* \geq u | T_I^* > t, b_i; \theta) \\ &\times p(b_i | T_I^* > t, Y_i(t); \theta) db_i \\ &= \int \frac{S_i\{u | M_i(u, b_i); \theta\}}{S_i\{t | M_i(t, b_i); \theta\}} p(b_i | T_i^* > t, Y_i(t); \theta) db_i,\end{aligned} \tag{4.2}$$

with $S_i(\cdot)$ being the patient's survival function and $M_i(\cdot)$ indicating the longitudinal biomarker history as formulated in the longitudinal repeated measures model 3.4. An estimate for the conditional probability can be derived using the empirical Bayes estimate for b_i or alternatively implementing a Monte Carlo simulation scheme with the following steps

Step 1. First we take K samples of $\theta^{(k)}, k = 1, \ldots, K$ from the MCMC sample of $p(\theta | D_n)$ that asymptotically corresponds to a normal posterior distribution $N(\hat{\theta}, I_n)$ where the variance-covariance matrix $I_n = -H_{hessian}^{-1}$ equals the observed information matrix

$$I_n = \Big\{ -\sum_{i=1}^{n} \frac{\partial^2 log p(y_i; T_i; \delta_i; \theta))}{\partial \theta^T \partial \theta} \Big|_{\theta = \hat{\theta}} \Big\}^{-1}. \tag{4.3}$$

Step 2. Second we draw K realizations $b_i^{(k)}$ for the random effects of the new subject i from their posterior distribution

$$p(b_i | T_i > t; Y_i(t); \theta^{(k)}) \propto \Big\{ \prod_{j=1}^{n_i(t)} p(y_{ij} | b_i; \theta^{(k)}) \Big\} S_i\{t | M_i(t; b_i); \theta^{(k)}\} p(b_i; \theta^{(k)}), \tag{4.4}$$

where $n_i(t)$ is the number of subject i measurements available by time t.

Step 3. . Third we derive an estimate of $\pi_i^{JM}(u|t)$

$$\hat{\pi}_i^{JM}(u|t) = \frac{1}{K} \sum_{k=1}^{K} \frac{S_i\{u | M_i(u, b_i); \theta^{(k)}\}}{S_i\{t | M_i(t, b_i); \theta^{(k)}\}}. \tag{4.5}$$

Given the structure of a joint model it can also be possible to compute predictions for the projected longitudinal profile of the marker, to initiate for example a preventive treatment and prevent worsening of the disease. For a subject i still alive

by follow-up time t the conditional expected value of the longitudinal outcome at time $u > t$ would be

$$w_i(u|t) = \mathbb{E}\{y_i(u)|T_i^* > t, Y_i(t), D_n; \theta_{true}\}. \tag{4.6}$$

In the same way as we have done with survival probabilities we can rewrite the expected value under the conditional independence assumption as

$$\begin{aligned}\mathbb{E}\{y_i(u)|T_i^* > t, Y_i(t), D_n\} &= \int \mathbb{E}\{y_i(u)|T_i^* > t, Y_i(t), b_i; \theta\} p(b_i|T_i^* > t, Y_i(t); \theta) db_i \\ &= \mathbb{E}\{y_i(u)|b_i\} p(b_i|T_i^* > t, Y_i(t); \theta) db_i \\ &= x_i^T(u)\beta + z_i^T(u) \int b_i p(b_i|T_i^* > t, Y_i(t); \theta) db_i. \end{aligned} \tag{4.7}$$

Assuming a large enough sample size such that $\theta|D_n \approx N(\hat{\theta}, I_n)$ as before, we derive the following Monte Carlo estimate for $w_i(u|t)$ using a three-step scheme as above:

$$\hat{w}_i^{JM}(u|t) = \frac{1}{K} \sum_{k=1}^{K} w_i(u|t)^{(k)}, \tag{4.8}$$

where $w_i(u|t)^{(k)} = x_i^T(u)\beta^{(k)} + z_i^T(u)b^{(k)}$ are computed for each sample k in the simulation.

4.2 Comparison between Landmarking and Joint Modeling

In this section we will follow Rizopoulos et al. [2013] and Rizopoulos et al. [2017] and Maziarz et al. [2017] comparing landmarking and joint modeling approaches for computing dynamic predictions. We have already mentioned in the introduction to this chapter that an alternative methods exists to predict dynamic survival probabilities. Landmark analysis is in contrast of joint modeling a simpler Landmark analysis can be considered a simpler procedure in this context because it's easily implemented by applying a standard time-dependent Cox, that considers only subjects at risk at the landmark time t ($R(t) =: T_i > t$).

The model is fitted to these individuals by setting the landmark time as time zero (baseline):

$$h_i(u) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr(u \leq T_i^* < u + \Delta t | T_i^* \geq u, Y_i(t)) = h_0(u) \exp\{\gamma^T w_i + \alpha \tilde{y}_i(t)\}, \tag{4.9}$$

with $\tilde{y}_i(t)$ being the last available longitudinal response. After the fitting, an estimate of $\pi_(u|t)$ is then simply obtained using the Breslow estimator for the cumulative baseline hazard:

$$\hat{\pi}_i^{LM}(u|t) = \exp\Big[- \hat{H}_0(u) \exp\{\hat{\gamma}^T w_i + \hat{\alpha} \tilde{y}_i(t)\}\Big], \tag{4.10}$$

with

$$\hat{H}_0(u) = \sum_{i \in R(t)} \frac{I(T_i \leq u)\delta_i}{\sum_{l \in R(u)} \exp\{\hat{\gamma}^T w_l + \hat{\alpha}\tilde{y}_l(t)\}}. \tag{4.11}$$

This formulation can be further extended to allow the baseline hazard to be a function of the visit time and not only of the last measurement, relaxing the proportional hazards assumption [Zheng and Heagerty, 2005], or to account for measurement error in the time-varying covariate by considering the predicted longitudinal outcome as seen in the two-stages approach for joint models (Section 3.2). A formulation for this landmarking mixed model is obtained by substituting $\tilde{y}_i(t)$ above with $\hat{m}_i(t) = y_i(t) - \epsilon_i(t)$

$$\hat{\pi}_i^{LMixed}(u|t) = \exp\Big[- \hat{H}_0(u) \exp\{\hat{\gamma}^T w_i + \hat{\alpha}\hat{m}_i(t)\}\Big], \tag{4.12}$$

A clear disadvantage of $\hat{\pi}_i^{LM}(u|t)$ compared to $\hat{\pi}_i^{JM}(u|t)$ is that it considers only a part of the total information resulting in lack of efficiency. However joint models makes more model assumptions. A misspecification or a different association structure in the joint model have been demonstrated in literature to strongly affect predictions.

Apart from the *endogeneity* concept that we already used in motivating joint models, a definition formulated by Jewell and Nielsen [1993], stated that of valid prediction function must satisfy the consistency condition:

$$h_i(t + s|Y_i(t)) = \mathbb{E}\{h_i(0|Y_i(t + s))|Y_i(t)\}. \tag{4.13}$$

This means that prediction at time $t + s$ should be calculated integrating out over the probability distribution of the longitudinal outcome in the interval $(t, t + s)$, this only possible if we derived it from the joint distribution of the outcomes, not modeled in standard and mixed landmarking.

4.3 Prediction accuracy measures

Normally, prediction models performance is assessed focusing on their *discrimination* (how well can the model discriminate between patients who had the event from those who did not) and *calibration* (how well the model predicts the observed data) power. Specific estimates for predictive measures in the joint models framework have been studied by Blanche et al. [2015]:

- **Discrimination measure**
  For any cut-off value c in $[0, 1]$ we classify subject i as a case if $\pi_i(t+\Delta t|t) \leq c$ and to the opposite we classify i as a control if $\pi_i(t + \Delta t|t) > c$. As generally done, we can compute the area under the receiver operating characteristic curve (AUC) by selecting a random pair of subjects i, j with measurements up to time t. Varying c, we have that if patient i experiences the event within the time interval $t + \Delta t$ whereas j does not the model should assign to patient j higher probability of surviving longer than $t + \Delta t$:

  $$AUC(t, \Delta t) = \Pr[\pi_i(t+\Delta t|t) < \pi_j(t+\Delta t|t)|\{T_i^* \in (t, t+\Delta t\,]\} \cap \{T_j^* > t + \Delta t\}]. \tag{4.14}$$

The estimates is derived by decomposition:

$$\widehat{AUC}(t, \Delta t) = \widehat{AUC_1}(t, \Delta t) + \widehat{AUC_2}(t, \Delta t) + \widehat{AUC_3}(t, \Delta t) + \widehat{AUC_4}(t, \Delta t), \tag{4.15}$$

where $\widehat{AUC_1}(t, \Delta t)$ is the proportion of concordant pairs of subjects out of those with sortable observed event times; whether $\widehat{AUC_2}(t, \Delta t)$, $\widehat{AUC_3}(t, \Delta t)$, $\widehat{AUC_4}(t, \Delta t)$ are the pairs of subjects who, due to censoring, cannot be compared, they are weighted with the probability that they would be comparable.

- **Calibration measure**
  Considering longitudinal information available up to time t, predictive accuracy at a specific time $u > t$ is given by the expected prediction error:

$$PE(u|t) = \mathbb{E}[L\{N_i(u) - \pi_i(u|t)\}], \tag{4.16}$$

where the expectation is taken with respect to the distribution of the event times, $N_i(t) = I(T_i^* > t)$ is the event status at time t and $L(\cdot)$ denotes a loss function.

Accounting for censoring an estimate has been derived as follows:

$$\begin{aligned} \widehat{PE}(u|t) \;=\; & \frac{1}{n(t)} \sum_{i:T_i \geq t} I(T_i \geq u) L\{1 - \hat{\pi}_i(u|t)\} + \delta_i I(T_i < u) L\{0 - \hat{\pi}_i(u|t)\} \\ & + (1 - \delta_i) I(T_i < u) \big[\pi_i(u|T_i) L\{1 - \pi_i(u|t)\} \\ & + \{1 - \pi_i(u|T_i)\} L\{0 - \pi_i(u|t)\}\big]; \end{aligned} \tag{4.17}$$

with $n(t)$ that is the risk set at time t.

The sum is composed by three main terms, the first denotes patients who were alive after time u, the second patients who dead before u and the third patients who were censored in the interval $[t, u]$.

Chapter 5

Simulation study

5.1 Simulation design for joint models

Simulation studies are a common strategy to evaluate the performance of a statistical model.

A joint model simulation requires itself to set up two different simulations, a mixed-effects model for the continuous longitudinal outcome and a relative risk model for individual survival probabilities. In both settings outcomes and covariates are simulated in order to reflect plausible scenarios. Visiting times for the longitudinal process can be differently simulated from a given distribution, for example gamma or uniform, or as a fixed grid of time points. Furthermore different For the survival process things are more complicated.

A reference papers which accurately describes how to possibly generate survival event times are Crowther and Lambert [2013] and Bender et al. [2005]. Classical choices for event times distributions are the Exponential, the Weibull and the Gompertz distribution, however these often simplify too much real clinical contexts. Bender et al. [2005] gives a first description of the basis theory for survival times simulations. Starting from the formula of the cumulative hazard function

$$H(T|X) = H_0(t)\exp(X\beta) = \int_0^t h_0(u)du\exp(X\beta), \tag{5.1}$$

with T being the survival simulated time, it has been shown that

$$F(T|X) = 1 - S(T|X) = 1 - \exp[-H(T|X)] = u, \tag{5.2}$$

where $u \sim U(0,1)$, which results in the conditional survival function $S(T|X) = u$.

Consequently, we find T solving the equation

$$T = H_0^{-1}[-log(U)\exp(-X\beta)]. \tag{5.3}$$

A necessary condition is the invertibility of the cumulative hazard function. This is the case for classical survival times distributions, but can be a problem in other situations requiring root-finding techniques to solve for T [Crowther and Lambert, 2013], e.g. "Brent's univariate root-finding method" that uses a function to iteratively find a solution for the equation $S(t) - U = 0$.

In practice, we can easily simulate a Cox model with a Weibull baseline hazard function, $h_0(t) = abt^{b-1}$ where a and b are the shape and scale parameters respectively. The Cox model then, can be assumed to depend on the current value of the longitudinal biomarker, or eventually on the slope or cumulative effect. To simulate the event times, we first simulated a subject-specific survival probability, s_i, from a $Uniform(0, 1)$ distribution and solved for T_i using for example R functions *integrate()* and *uniroot()*, from the following equation:

$$s_i - \exp\{-\int_0^{T_i^*} abu^{b-1} \exp(\alpha_1 y_1^*(u))du\} = 0. \tag{5.4}$$

In the following we report the specific R code to compute the inverse of the survival function, where:

- **invS** is the inverse of the survival function and **h** is the hazard function
- **t** is the upper time limit of the integral
- **u** is the upper time limit of the integral
- **i** is the subject identifier
- **n** is the number of subjects in the sample
- **XX** and **ZZ** are the fixed and random effects matrices for the longitudinal process
- **phi** is the scale of a Weibull baseline hazard
- **ff** is the association parameter between the true value of the longitudinal outcome $f1$ and the event process

```
invS <- function (t, u, i) {
  h <- function (s) {
    f1 <- as.vector(XX %*% betas + rowSums(ZZ * b[rep(i, nrow(ZZ)), ]))
    exp(log(phi) + (phi - 1) * log(s) + eta.t[i] + f1 * alpha)
  }
  integrate(h, lower = 0, upper = t)$value + log(u)
}
u <- runif(n)
trueTimes <- numeric(n)
for (i in 1:n) {
  Up <- 50
  tries <- 5
  Root <- try(uniroot(invS, interval = c(1e-05, Up), u = u[i], i = i)$root, TRUE)
  while(inherits(Root, "try-error") && tries > 0) {
    tries <- tries - 1
    Up <- Up + 200
    Root <- try(uniroot(invS, interval = c(1e-05, Up), u = u[i], i = i)$root, TRUE)
  }
```

```
  trueTimes[i] <- if (!inherits(Root, "try-error")) Root else NA
}
na.ind <- !is.na(trueTimes)
trueTimes <- trueTimes[na.ind]
```

Censoring times were independently simulated from another uniform distribution and the event variable is then created following the rule in survival analysis $Y = min(T, C)$:

```
set.seed(1)
Ctimes <-runif(n, first_measurement_time, last_measurement_time)
Time <- pmin(trueTimes, Ctimes)
event <- as.numeric(trueTimes <= Ctimes) # event indicator
```

With the results from simulated datasets we can verify common properties and assumptions underlying the fitted model, for example bias, coverage and estimates robustness and its ability to handle highly complex longitudinal trajectories, mimicking in every detail the real association process.

5.2 Simulation in practice

In the following we are going to implement a joint model simulation in practice. Primary objective of this simulation is to compare the two-stages with the EM estimation approach for joint models. As we have already mentioned in chapter 2, the two-stages approach does not consider bias due to possible non random drop-outs and to the fact that using estimates for the values of the longitudinal outcomes it also does not considered that these are affected by measurement errors [Qiu et al., 2016].

We performed a 100 simulations for 100 patients that were assumed to be followed up for a maximum period of 15 years, resembling the setting in Rizopoulos et al. [2017]:

- longitudinal measurements were recorded at baseline and afterwards at 14 random follow-up times.
- we simulated normally distributed random effects.
- in the longitudinal process we used B-splines of time with two internal knots placed at 2.5 and 6 years and boundary knots at 0.5 and 13 years. The fixed effects matrix included a dummy variable to indicate treatment group and its interaction with visit time (as modeled by splines). Natural cubic splines for time were also used to simulate the random effects matrix.
- the survival process was formulated as $h_i(t) = h_0(t) \exp[\gamma_1 Treatment + \alpha_1 y_i(t)]$, with a Weibull baseline hazard.
- censoring times were simulated from a uniform distribution.

- we dropped longitudinal measurements taken after the observed event time for each subject.

True parameters for γ, ϕ scale for the Weibull baseline hazard, D covariance matrix for random effects were taken from Rizopoulos et al. [2017] supplementary material file. A seed has been introduced prior to the simulation to have reproducible results.

The simulation was repeated for various scenarios of $\alpha = \{0.04672, 0.8672, 1.01\}$. For all scenarios and simulations we fitted a joint model and an extended Cox model plugin-in the time-varying *blups* as estimated from the linear mixed model. We also fitted a misspecified version for both models, including visit time covariate in a linear form in the LME model instead of modeling its association with the longitudinal outcome by mean of natural cubic splines.

For all scenarios, mean estimates, standard errors, bias ($\sum_{i=1}^{100} \hat{\theta}_i/100 - \theta_{true}$) and mean square error ($\sum_{i=1}^{100}(\hat{\theta}_i - \theta_{true})^2/100$) haven been calculated and reported in table 5.1.

Parameter	True Value	Model	Coeff	Std.Err	Bias	Mse
Alpha	0.4672	joint	0.4749	0.1048	0.0077	0.0118
		twostage	0.4195	0.0992	-0.0477	0.014
		joint.linear	0.3111	0.1019	-0.1561	0.062
		twostage.linear	0.3116	0.1026	-0.1556	0.0596
Gamma	0.48	joint	0.526	0.3178	0.046	0.1307
		twostage	0.5045	0.317	0.0245	0.1123
		joint.linear	0.4451	0.3139	-0.0349	0.1129
		twostage.linear	0.4441	0.3162	-0.0359	0.108
Alpha	0.8672	joint	0.9081	0.1562	0.0409	0.0296
		twostage	0.761	0.1327	-0.1062	0.0387
		joint.linear	0.8225	0.1805	-0.0447	0.067
		twostage.linear	0.7119	0.148	-0.1553	0.0745
Gamma	0.48	joint	0.4938	0.2743	0.0138	0.0751
		twostage	0.4116	0.2617	-0.0684	0.0717
		joint.linear	0.4583	0.2821	-0.0217	0.0943
		twostage.linear	0.3868	0.2647	-0.0932	0.0858
Alpha	1.01	joint	1.0084	0.1704	-0.0016	0.0331
		twostage	0.8781	0.1455	-0.1319	0.0545
		joint.linear	1.0066	0.2071	-0.0034	0.0734
		twostage.linear	0.8535	0.1626	-0.1565	0.0752
Gamma	0.48	joint	0.4635	0.2694	-0.0165	0.0732
		twostage	0.3951	0.2532	-0.0849	0.0749
		joint.linear	0.481	0.2834	0.001	0.1029
		twostage.linear	0.3947	0.2574	-0.0853	0.0836

Table 5.1: Models results from the 100 simulated datasets.

In this simulation example we confirm what reported from literature, so that EM approach is more efficient and less biased than two stages methods for estimating the association between a longitudinal and a time-to-event process.

In particular, we notice that the fitted joint model for all three scenarios gives lower bias and mean square error compared to the two stage model for estimating the association coefficient α. With respect to the estimation of γ parameter for the baseline treatment covariate, mse stays very similar in both approaches, only in the first scenario (with low association effect $\alpha = 0.46$) bias and mse results slightly higher for the joint model. When increasing the effect of association in second and third scenarios ($\alpha = \{0.86, 1.01\}$) the difference in terms of α and γ estimation bias between the two methods increases. We see in the case when models are misspecified ("joint.linear" and "two stage.linear") the EM approach ("joint.linear") gives less biased results than two stage method. Further investigation could be done by increasing or decreasing the variance of the longitudinal outcome, the percentage of censoring and the number of measurements for each subject. Similar simulations have been studied in Wen et al., Ibrahim et al. [2010], Sweeting and Thompson [2011],Murawska et al. [2012].

Chapter 6

Data Application

In this chapter, the methods previously introduced in Joint Modeling techniques are now applied to liver transplantation data. In order to analyze our dataset, we will start by describing pre-existent variables, including percentages of respective missing values and exploring relationships. After a first exploratory analysis we will introduce the modeling approach suitable for the goal of the study, we will then describe results and inherent issues encountered, regarding for example computational aspects. Finally we also tried to build a simulation study with similar characteristics of our real data, in this way we try to better understand, justify and judge the model by also comparing it with alternatives parameterizations and approaches.

6.1 Kidney Research Data

As we introduced in the first chapter of this manuscript, our study has been motivated by the analysis of kidney transplantation data.

Data come from the Australian and New Zealand Dialysis and Transplant Registry (ANZDATA) that collects and reports the incidence, prevalence and outcome of dialysis treatment and kidney transplantation for patients with end stage kidney disease across Australia and New Zealand. The registry gather information on Australian patients starting receiving renal replacement therapy between 1971 and 2014 and transplantation between 1980 and 2014. In total, 16820 incident post-transplant patients were included in the study and prospectively followed. Patients consist only in transplant recipients, this is the case because only for them is possible to collect serum creatine measurements (mol/L) and from this to estimate glomerular filtration rate values (mL/min/1.73m^2).

eGFR acronym stands for "Estimated Glomerular Filtration Rate". The eGFR measures how well the kidneys filter the wastes from the blood and is recognized as the best overall measure of kidney function. It helps to determine if there is any kidney damage, if the filtration rate is low the kidneys are not working properly.

It is difficult to calculate the exact glomerular filtration rate at which patient's kidneys are working, therefore a special formula has been developed to estimate it:

$$eGFR = 175 \times (\frac{SCr}{88.4})^{-1.154} \times age^{-0.203} \times 0.742 \times I(female) \times 1.21 \times I(race) \quad (6.1)$$

This formula uses age, gender and blood level of creatinine (SCr). Creatinine is a waste product made by the muscles, usually removed by the kidneys before passing out in the urine. When the kidneys are not working well more creatinine stays in the blood.

Normal filtration rate in young adults is about 90-100 milliliters every minute, but generally speaking only a value below 60 mL/min/1.73m^2 is said to suggest a sign of kidney loss function and is usually taken as possible cut-off value in clinical studies. The information above are presented in detail in the website `http://kidney.org.au` under the library resources fact sheets, otherwise easy understandable formulas and concepts about eGFR can also be found in Wikipedia `https://en.wikipedia.org/wiki/Renal_function` or in various medical resources.

In our clinical study patients had periodical medical visits after transplant. These visits have been prestablished following medical advice at 1,2,3,6 months and subsequently at 1,2,3,5,7,10,15,20,30,35,40 years. As frequently happens some measurements may be missing for some patients. For example, they may be randomly missing if they skip a visit, or non randomly if they stop going to the hospital if they feel completely fine or in contrast if they die. No biomarker's values are recorded after death.

In the attempt to work with clean and reliable data, we first excluded from our dataset 967 patients with no eGFR measurements and 1604 patients with less than three eGFR measurements, to better catch their longitudinal process over time. After data preprocessing the total number of subjects in the study has been reduced to 15216.

Information available in our dataset were recorded in two different registration dates. In those occasions the current patient status, categorized as alive, graft-failure or death, was registered along with the possible death date or graft-failure date. In our analysis we don't consider intermediate recordings but only outcomes at the latest follow up date for every subject.

We are interested in the patients' survival. We analyze a composite outcome, for which a patient is recognized to have experienced an event if at the end of follow up he or she is died or has experienced graft-failure. Lost to follow-up up is, as usual, considered as censoring event.

6.1.1 Descriptive analysis

In this section we will summarize our data, with descriptive statistics from tables and plots.

In table 6.1 we give a short overview of a selection of variables in the original dataset:

- **n** total number of patients from the original registry
- **txage1** is the patient's age at transplant time and *sex* is the patient's gender
- **egfr_sc1m**, . . . , **egfr_sc30y** are levels of eGFR at the specific visit points (one, six months and one, five, ten, thirty years)

- **txstatus1a** is subjects' status in the intermediate recording date while **fustatus** is the status as recorded at the end of the subject follow-up period.

We see that biomarker means and standard deviations are slightly different at different visit times.

	Overall
n	16820
txage1 (mean (sd))	43.71 (15.34)
egfr_sc1m (mean (sd))	55.40 (49.94)
egfr_sc6m (mean (sd))	57.45 (41.48)
egfr_sc1y (mean (sd))	57.59 (40.33)
egfr_sc5y (mean (sd))	54.78 (27.22)
egfr_sc10y (mean (sd))	54.69 (24.32)
egfr_sc30y (mean (sd))	66.89 (30.19)
timetoend (mean (sd))	10.09 (7.96)
sex = M (%)	10172 (60.5)
txstatus1a (%)	
alive	9025 (53.7)
graft-failure	4542 (27.0)
death	3253 (19.3)
fustatus (%)	
alive	9068 (53.9)
graft-failure	2200 (13.1)
death	5552 (33.0)

Table 6.1: Descriptive statistics for the main variables, before dataset's preprocessing.

Overall distribution of eGFR, for all the patients and their measurements, is reported in figure 6.1 showing an improvement in the skewness of the variable, after the logarithmic transformation.

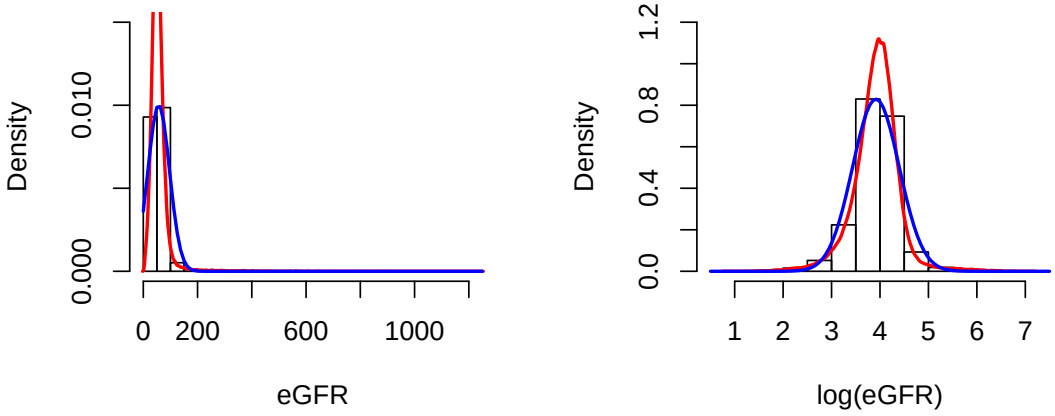

Figure 6.1: Distribution of eGFR variable and log(eGFR), we added a kernel and a normal density curve overlaying the histograms.

In figure 6.2 we report changes of log(eGFR) variable along follow-up years. As we can see, its mean value is increasing from the first visit point, occuring at one

month after transplant, to the latest one. This should suggest that patients with lower eGFR are dying and dropping out of the cohort (non random drop-out).

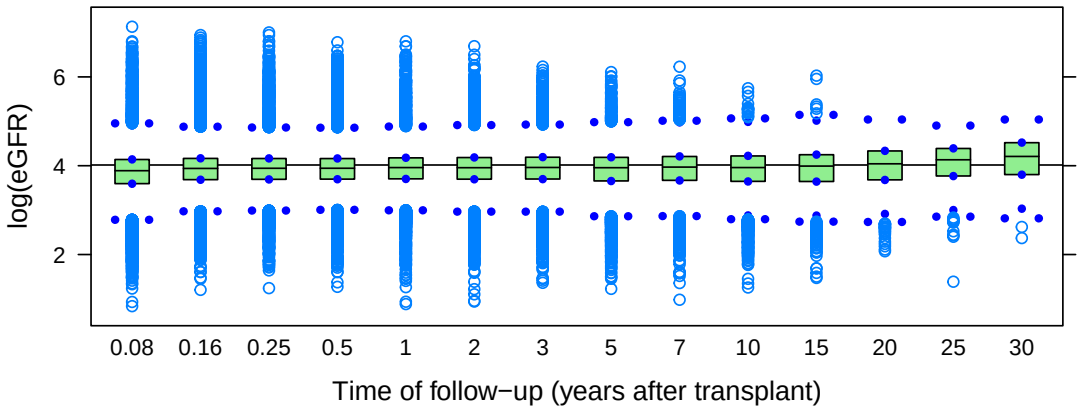

Figure 6.2: Bivariate Trellis Plot showing mean and standard deviation of log(eGFR) among visits.

We investigated the relationship between patients' eGFR levels at each visit time and patients' status at the end of the follow-up period. We notice in figure 6.3 that for the first three measurements points patients classified as deaths, at the end of the follow-up, have lower mean of estimated filtration rate compared to the others. In later years this is the case, instead, for patients who experienced a graft-failure.

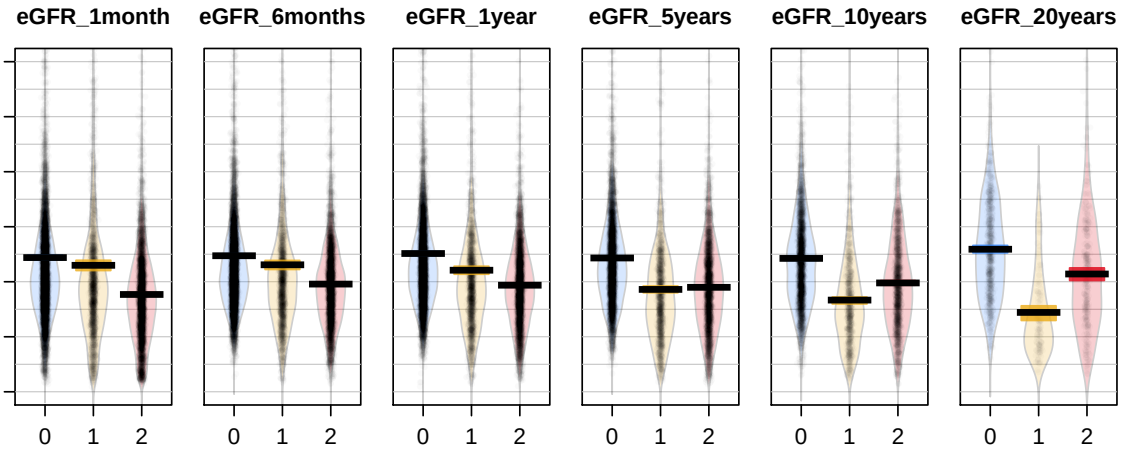

Figure 6.3: Box plots showing relationships between eGFR and status variables at five different visit times. For the status variable level "0" stays for alive, "1" for graft-failure and "2" for death.

During a data preprocessing, the original dataset was transformed into a long format dataset, having a number of rows for each patients that correspond to the number of measurements available. A new subject-specific variable for eGFR values, varying along years of follow up, was created. In the following 6.4 we describe the

visiting process distribution and visualize a sample of subjects trajectories for the longitudinal biomarker measurements.

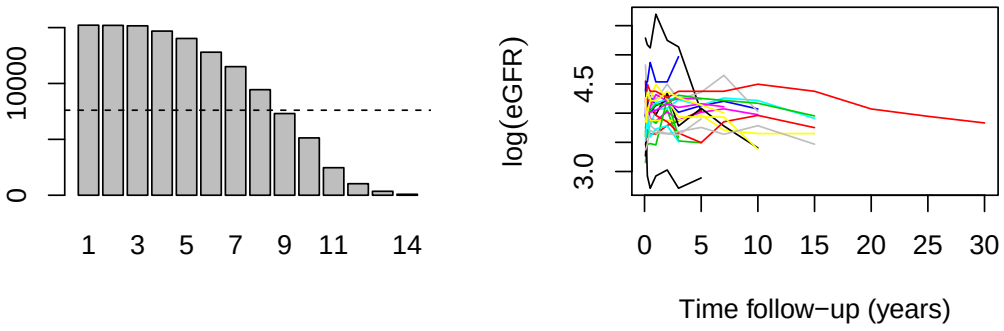

Figure 6.4: On the left: number observations at each visit time. Dashed line rapresent the half level of the sample population. On the right: eGFR trajectories of twentyfour randomly selected patients.

Figure 6.5 exhibits specific indivual smoothed log(eGFR) curves, plotted for a small sample of twelve subjects.

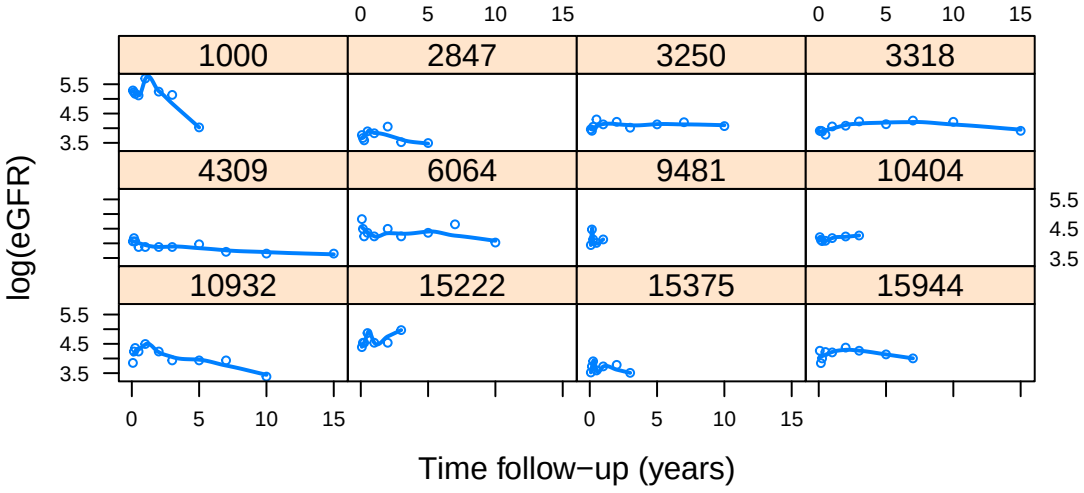

Figure 6.5: Smoothed curves eGFR profiles for twelve patients.

This simple visualization helps the analyst to explore how biomarker trajectories vary across patients and along time of visit for each individual. We clearly see that evolution of eGFR is very different between patients, it can change from having a flat trend in time, to have a very variable trend that cannot be captured with a linear curve.

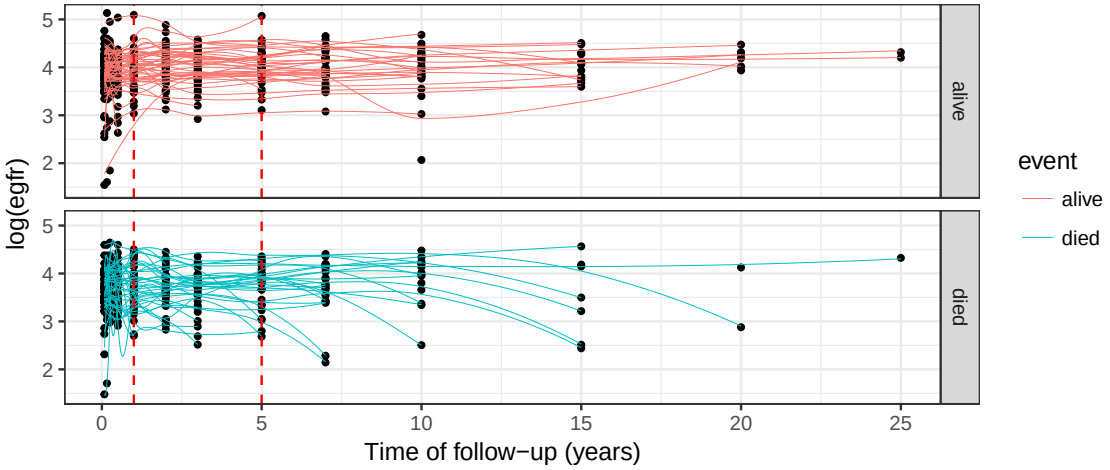

Figure 6.6: Smoothed trajectories eGFR for patients with or without an event at the of follow up.

At the end of the analysis a total number of 6478 events was recorded(43%). Median follow-up time was 8.73 years. Figure 6.7 displays estimates for unadjusted survival probabilities. On the right side, the plot of $\hat{S}_{km}$ stratified by a binary variable for log(eGFR) values at baseline, shows that survival probabilities are indeed lower for those having eGFR < 60 mL/min/1.73m^2 at first visit.

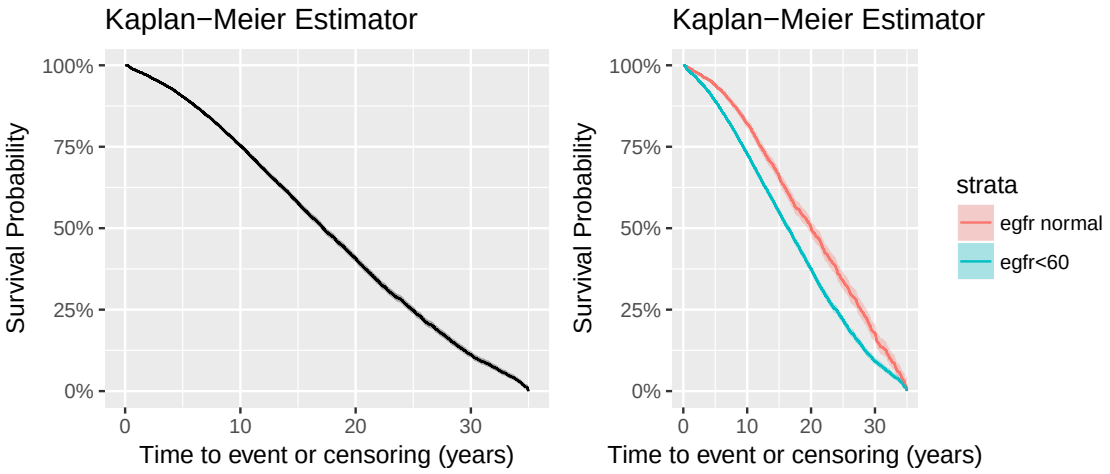

Figure 6.7: Kaplan Meier estimate for the survival function curve. On the right we stratified for eGFR factor at baseline (level "normal" indicates a value for eGFR ≥ 60 mL/min/1.73m^2).

If we were specifically interested in graft-failure after transplant probability, survival analysis should have been taking into account the "Competing Risks" [Putter et al., 2007] context, as reported in Andrinopoulou et al..

6.2 Model Building

In this section we set up theoretical models to adapt them to our data. With joint models we try to explain and measure the effect of repeated eGFR measurements

on time to all-cause death for patients under study.

A linear mixed models (LMM) was specified to estimate individual trajectory changes in time and across individuals in renal functionality. The logarithm of eGFR was chosen as form of the longitudinal response, because it approximately follows a symmetrical, normal distribution with mean equal to3.92 and standard deviation 0.48. In the original eGFR biomarker variable mean was equal to 56.84 and standard deviation to 39.81. On the other hand, the usual Cox proportional hazard model was utilized to investigate independent mortality(or graft-failure) risk factors caused by baseline characteristics (age and sex).

6.2.1 Longitudinal process of glomerular filtration rate

A linear mixed effects model (LMM) was implemented to model the longitudinal process of log(eGFR) along of years of follow-up visits. The basic model we first specify was a random intercept and slope model in the form

- $log(eGFR)_i(t) = \beta_0 + \beta_1 t + b_{0i} + b_{1i} t + \epsilon_i(t)$.

However, we noticed from raw longitudinal plots we noted that it becomes necessary in our model to allow for more flexibility in the specification of the patient-specific longitudinal trajectories. In a second specification natural cubic splines *ns()* and B-splines *bs()* from library *splineDesign* in R were used, varying their degrees of freedom; knots were placed at the corresponding quartiles of the follow-up times.

The new formulated spline models were :

- $log(eGFR)_i(t) = (\beta_0 + b_{i0}) + \sum_{k=1}^{3} (\beta_k + b_{ik}) B_n(t, d_k) + \epsilon_i(t)$
- $log(eGFR)_i(t) = (\beta_0 + b_{i0}) + (\beta_k + b_{ik})^T B(t, df, 4) + \epsilon_i(t)$;

where $B(t,d_k); k = 1, 2, 3$ denotes a B-spline basis matrix for a natural cubic spline of time with two internal knots placed at the 33.3% and 66.7% percentiles of the follow-up times and $B(t; df = 4, 5; q = 4)$ denotes the matrix for $q-1$ degree splines with $df - q + 1$ internal knots; β_k and b_{ik} are the vectors of fixed and random effects corresponding to the B-splines matrix. The random effects are assumed to have a diagonal covariance matrix.

Model	Variable	Coeff	Std.Error	t-value	p-value
Model0	(Intercept)	3.9254	0.0037	1069.4614	<0.0001
	time	-0.0186	8e-04	-24.1238	<0.0001
Model1	(Intercept)	3.8972	0.0039	999.4395	<0.0001
	ns(time, 3)1	-0.2501	0.0098	-25.5202	<0.0001
	ns(time, 3)2	-0.1416	0.009	-15.6868	<0.0001
	ns(time, 3)3	-0.14	0.0165	-8.4853	<0.0001
Model2	(Intercept)	3.882	0.004	975.5573	<0.0001
	bs(time, 4)1	0.0697	0.0032	21.9533	<0.0001
	bs(time, 4)2	-0.2677	0.0141	-18.9985	<0.0001
	bs(time, 4)3	-0.3385	0.0294	-11.5122	<0.0001
	bs(time, 4)4	-0.3041	0.0456	-6.6703	<0.0001
Model3	(Intercept)	3.8492	0.0041	944.6994	<0.0001
	bs(time, 5)1	0.0906	0.0025	36.0202	<0.0001
	bs(time, 5)2	0.0699	0.0037	19.1382	<0.0001
	bs(time, 5)3	-0.1505	0.0136	-11.0446	<0.0001
	bs(time, 5)4	-0.4366	0.0304	-14.3632	<0.0001
	bs(time, 5)5	-0.2237	0.0453	-4.9346	<0.0001

Table 6.2: Linear mixed models results for the association between the logarithm of estimated filtration rate and visit times.

	df	AIC	BIC
Model0	6.00	55683.39	55741.78
Model1	9.00	50531.36	50618.94
Model2	11.00	41453.91	41560.95
Model3	13.00	38656.01	38782.51

Table 6.3: Information criteria for the fitted linear mixed models.

Tables 6.2 and 6.3 respectively report results for coefficients' estimation in the different linear mixed models formulated along with information criteria (AIC and BIC), these recognize as best model to be kept into the joint formulation Model 3 with B-splines of five degrees of freedom and two internal knots. However, this longitudinal model is also the more complex in terms of random effects specification, resulting therefore in a computationally more demanding estimation procedure for the joint model.

6.2.2 Cox PH model for the event process

For the survival submodel we formulated a Cox proportional hazard model with age and sex baseline covariates and log(eGFR) as time-dependent variable

$$h_i(t) = h_0(t) \exp[\gamma_1 age + \gamma_1 sex + \alpha log(eGFR)_i(t)],$$

with $h_0(t)$ is the baseline risk function, t is the time-to-event and log(eGFR) is the true (unobserved) value of the longitudinal biomarker.

6.3 Results

A crude baseline Cox proportional hazards analysis was first formulated as starting point. The model included individuals values of log(eGFR) at first measurement time as baseline covariate with age and sex characteristics. This resulted in a hazard ratio for death/graft-failure of 0.81(95% CI: 0.78; 0.85) for each unit increase in the logarithm of eGFR level, having fixed the other covariates at baseline. If we want to interpret the outcome in terms of unit decreasing of log(eGFR), the HR is equal to the exponential of the negative coefficient, $\exp\{-\hat{\beta}_{log(egfr)}\}$. In this specific case we would stay that for each unit decrease in log(eGFR) variable there is a 23% increase in the risk of event.

A first attempt to include repeated biomarker measurements in a survival model is the classical time-dependent Cox model (also known as the "Andersen-Gill" model). As previously mentioned in this paper, an extended Cox model is not be adequate for the analysis of a time-varying *internal* covariate. With this in mind we nevertheless fitted the misspecified model to see how it affects estimation results. In the extended Cox model time-dependent covariates are usually encoded using the (start, stop] notation, therefore we further modified the dataset to have for each subject information on the longitudinal process $y_i(t)$ for each specific time interval.

We proceeded to implement joint models and to report here the estimates for the association coefficient as obtained from different formulations. In the following table we present results from the joint model where the linear mixed model specify a non linear association between the logarithm of eGFR and time of follow up visit, modeled by splines (df=4, one knot at one year).

	Event Process				Longitudinal Process		
	Value	Std.Err	p-value		Value	Std.Err	p-value
txage1	0.04	0.00	< 0.0001	(Intercept)	3.88	0.00	< 0.0001
sexM	0.14	0.03	< 0.0001	bs(time, 4)1	0.07	0.00	< 0.0001
Assoct	-0.91	0.02	< 0.0001	bs(time, 4)2	-0.29	0.01	< 0.0001
$\log(\xi_1)$	-2.49	0.10		bs(time, 4)3	-0.30	0.03	< 0.0001
$\log(\xi_2)$	-2.27	0.10		bs(time, 4)4	-0.60	0.03	< 0.0001
$\log(\xi_3)$	-1.96	0.09		$\log(\sigma)$	-1.70	0.00	
$\log(\xi_4)$	-1.68	0.09					
$\log(\xi_5)$	-1.42	0.09		(Intercept)	0.22	0.00	
$\log(\xi_6)$	-1.08	0.09		bs(time, 4)1	0.09	0.00	
$\log(\xi_7)$	-0.37	0.08		bs(time, 4)2	1.70	0.03	
				bs(time, 4)3	2.07	0.07	
				bs(time, 4)4	1.04	0.05	

Table 6.4: Parameter estimates, standard errors and p-values under the joint modeling analysis. D_{ij} denote the ij-element of the covariance matrix for the random effects.

In the results for the survival process, 'Assoct' is the parameter denoted as α in joint models notation and measures the association between the current value of the biomarker $m_i(t)$ (in our case log(eGFR) level at time t) and the risk for death or graft-failure. The parameters $\varepsilon_1, \ldots, \varepsilon_7$ are the parameters for the piecewise-constant baseline risk function assumed. The model finds a strong association between the longitudinal and the time-to-event outcome, with a unit decrease in the marker corresponding to a $\exp(-\alpha) = 2.49$-fold increase in the risk for the event (95% CI: 2.38; 2.6).

An alternative parameterization was also formulated, including the slope of the longitudinal marker at time t, $m_i^{'}(t)$, as covariate in the model:

$$h_i(t) = h_0(t) \exp[\gamma_1 age + \gamma_1 sex + \alpha_1 log(eGFR)_i(t) + \alpha_2 log(eGFR)_i^{'}(t)].$$

A significant effect was found. The association between the current value and the event outcome was now $\alpha_1 = -1.15$ (95% CI: -1.21 ; -1.09) whereas the association between the slope of the marker and the event outcome at time t was $\alpha_2 = 5.32$ (95% CI: 4.4 ; 6.24).

Model comparisons In table 6.5 we list the estimated α parameters under a set of alternatives to the classical joint model. By row, first model appearing in the table is the baseline Cox model where only log(eGFR) level at first visit was included for each patient, the second one is the extended Cox model where measurement error for the time-varying covariate is not taken into account. Finally the last three models denotes two stages estimation's methods, under different splines degrees, that considers the *blups* from the longitudinal models as time-dependent covariate in the survival analysis. Last model used B-splines with five degrees of freedom and two knots are assumed.

Model	Coeff	Std.Err	z-value	p-value	AIC	BIC
surv.basecox	-0.2082	0.0228	-9.131	<0.0001	107189.15	107209.48
surv.tdcox	-0.84	0.0181	-46.4153	<0.0001	102694.81	102715.07
surv.twolin	-0.9709	0.0247	-39.285	<0.0001	103167.08	103187.34
surv.twons	-0.9898	0.0239	-41.4389	<0.0001	103042.82	103063.07
surv.twobs	-0.9823	0.0226	-43.437	<0.0001	102871.84	102892.1
surv.twobs2	-0.9674	0.0221	-43.798	<0.0001	102845.37	102865.63

Table 6.5: Extended Cox models scenarios. Results for the estimates of the 'Assoct' parameter that measures association between the log(eGFR) and time-to-event outcome

In table 6.6 instead we compare results for the estimated α coefficient in joint models settings. Last two models differs from the one in 6.4 for the baseline hazard, here respectively assumed as following a Weibull distribution or alternatively modeled by regression splines.

Model	Coeff	Std.Err	z-value	p-value	AIC	BIC
joint.linear	-0.4314	0.0129	-33.4635	<0.0001	105472.72	105594.8
joint.ns	-1.0118	0.0255	-39.611	<0.0001	99759.37	99904.34
joint.bs	-0.9126	0.0225	-40.5828	<0.0001	90338.37	90498.6
joint.bs2	-0.9336	0.0225	-41.5198	<0.0001	87477.23	87652.72
joint.bs_spline	-0.8581	0.0219	-39.1234	<0.0001	90077.29	90252.78
joint.bs_weibull	-0.9008	0.0219	-41.0507	<0.0001	90684.33	90806.41

Table 6.6: Joint models scenarios. Results for the estimates of the 'Assoct' parameter that measures association between the log(eGFR) and time-to-event outcome

As we can notice, most of the models give similar results. Only the baseline Cox model and the joint model with a linear specification seem to largely underestimate the association between log(eGFR) and survival.

Dynamic predictions Here we illustrate how the fitted joint model can be used for individualized predictions for the survival and longitudinal outcomes. We computed conditional survival probabilities $\pi_i(u|t)$ at time $u > t$, for one patient in our dataset survived up to last point t. In figure 6.8 we observe the patient's evolution of log(eGFR) biomarker up to 10 years.

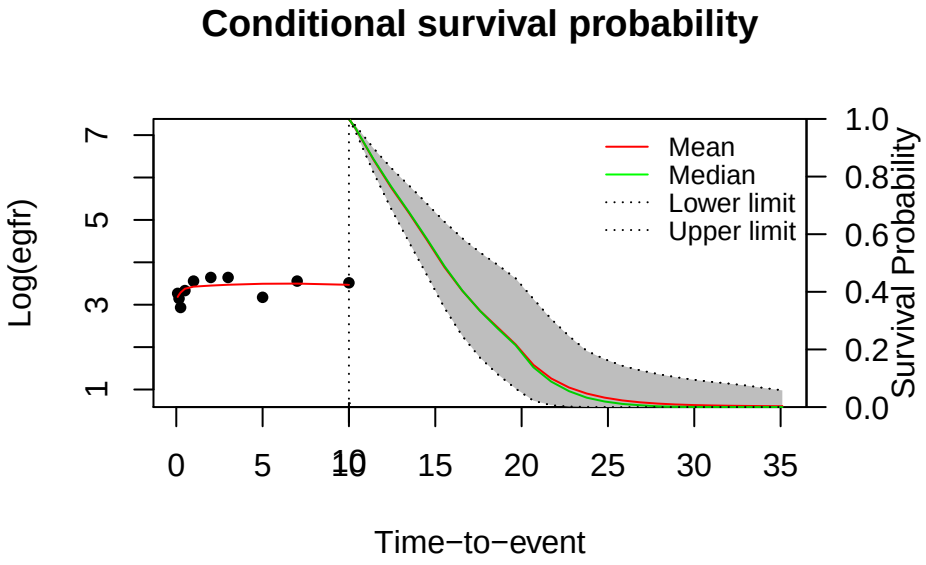

Figure 6.8: Survival probabilities for Patient 16814. The dashed and solid lines correspond to the median and mean estimators, the dotted lines are the corresponding 95% pointwise confidence intervals.

To better understand how the changes in the logarithm eGFR are reflected in changes in the dynamic updates of the survival probabilities in time, we plotted the updated survival curves after the baseline measurement (one month after transplant)and at one, two, five and ten years after transplant. Figure 6.9 captures an increase of the biomarker level at one year after transplant and a further improvement after two years; but the trajectory then becomes flat (stable) till the end of the following period.

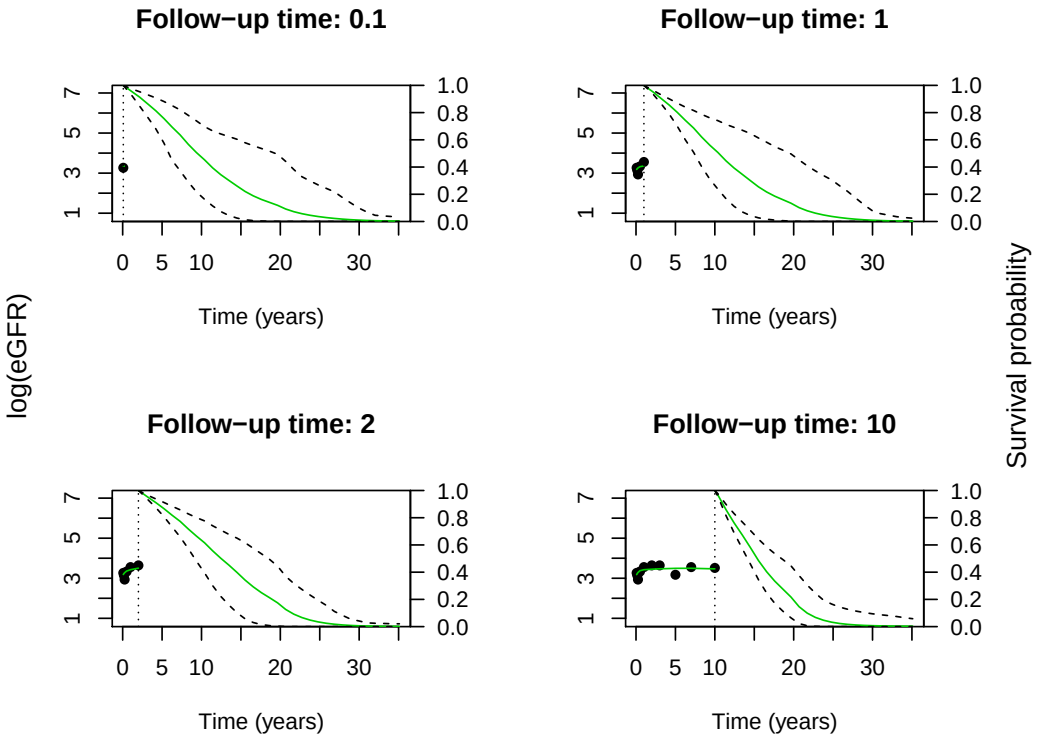

Figure 6.9: Dynamic survival probabilities for Patient 16814 during follow-up. The vertical dotted lines represent the time point of the last marker measurement, on the left of the line the longitudinal trajectory is detected, on the right the solid line represents the median estimator for $\pi_i(u|t)$, and the dashed lines the corresponding 95% pointwise confidence intervals.

Figure 6.10 explains in more detail the process in the subject's survival probability estimation, illustrating how point estimates for median survival along with their 95% confidence intervals change from one visit time to another and increasing the time of prediction u.

After five and even more after ten years, we can easily notice that estimated median survival probability decrease for Patient 16814 and confidence bands become wider.

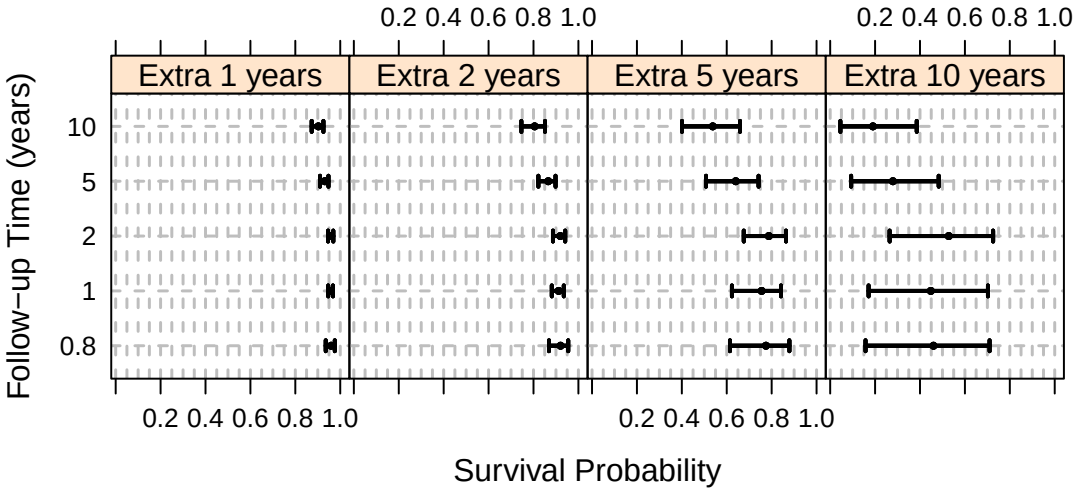

Figure 6.10: Dynamic survival probabilities for Patient 16814. In each panel five estimates and the associated 95% confidence intervals of $\pi_i(u|t)$ are presented, for t that equals the time point the most recent log(eGFR) measurement was collected.

Till now our principal interest was on patient's survival. Nevertheless, joint models also allow us to compute dynamic predictions for the longitudinal outcome. In a similar way as before and for a specific subject, still alive by follow-up time t, we may want to know the expected value $w_i(u|t)$ of his longitudinal outcome at time $u > t$ given the values observed for the same individual up to that time point.

We kept as example a different patient in our dataset and computed dynamic predictions for his log(eGFR) values. As for conditional survival probabilities, these predictions are dynamically updated in time when extra information about the patient is available. We observe that

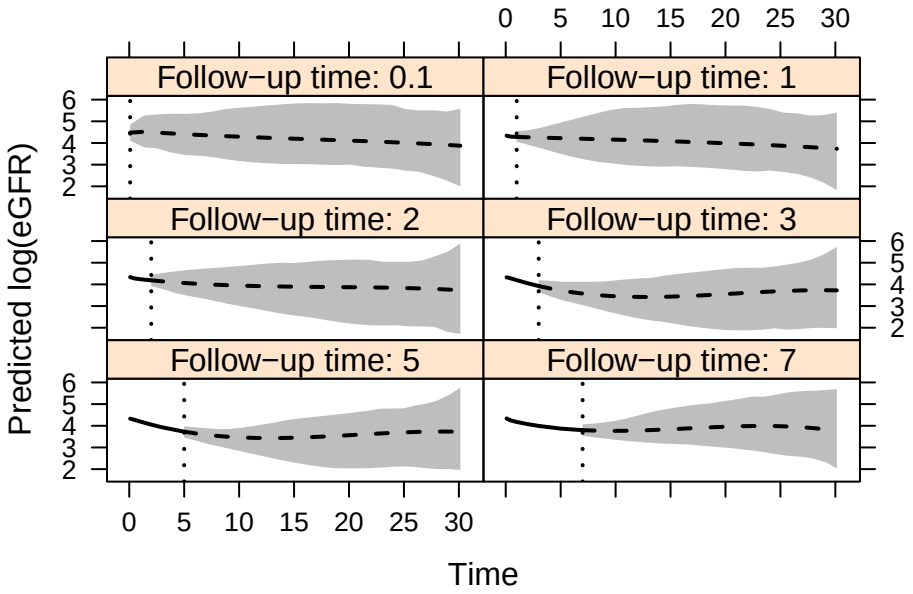

Figure 6.11: Dynamic predictions of longitudinal responses for Patient 14312. In each panel the dotted vertical line denotes the time point of the last observed longitudinal response. The solid line left to the dotted line denotes the fitted longitudinal trajectory prior to the last visit, and the dashed line right to the dotted line denotes the predicted longitudinal trajectory. The grey areas denote the 95% pointwise confidence intervals.

At the end of our analysis we proceeded to assess the eGFR biomarker predictive accuracy. Time-dependent sensitivity and specificity and corresponding ROC curve and AUC have been calculated.

We built a new data frame by considering as representative, a subject with a mean age at baseline and who has provided, for example, five eGFR measurements at one month, three months, one year, three and five years. We are required to specify the lengths Δt of the medically relevant time intervals, we choose six months, one, five and ten years. With R function *rocJM()* we estimated sensitivity and specificity by mean of Monte Carlo simulations and from this constructed the corresponding ROC curve and calculate the AUC.

```
##
## Areas under the time-dependent ROC curves
##
## Estimation: Monte Carlo (500 samples)
## Difference: absolute, lag = 1 (0)
## Thresholds range: (-2.28, 11.69)
##
## Case: 1
## Recorded time(s): 0.08, 0.25, 1, 3, 5
##    dt t + dt    AUC   Cut
##   0.6    5.6 0.6756 3.306
##   1.0    6.0 0.6786 3.306
```

```
##   5.0   10.0 0.7100 3.362
##  10.0   15.0 0.7500 3.474
```

Output provides the time-dependent AUCs under different dt and threshold values for the biomarker that maximize the product of sensitivity and specificity under the same options. The simple prediction rule above is based on the last marker measurement able to discriminate between cases and controls. On the other hand, a composite prediction rule assumes that a patient has higher chance to experience the event within the time interval $(t, t + \Delta t]$ when he shows a 20% decrease in his eGFR levels between two subsequent visits.

```
##
## Areas under the time-dependent ROC curves
##
## Estimation: Monte Carlo (500 samples)
## Difference: relative, lag = 2 (1, 0.8)
## Thresholds range: (-2.28, 11.69)
##
## Case: 1
## Recorded time(s): 0.08, 0.25, 1, 3, 5
##    dt t + dt    AUC Cut.1 Cut.2
##   0.6    5.6 0.6563 4.033 3.226
##   1.0    6.0 0.6595 4.033 3.226
##   5.0   10.0 0.6944 4.088 3.271
##  10.0   15.0 0.7421 4.256 3.405
```

We can see how AUCs and ROCs are very similar between the two prediction rules. A better discrimination is achieved for $\Delta t = 10$ 6.12.

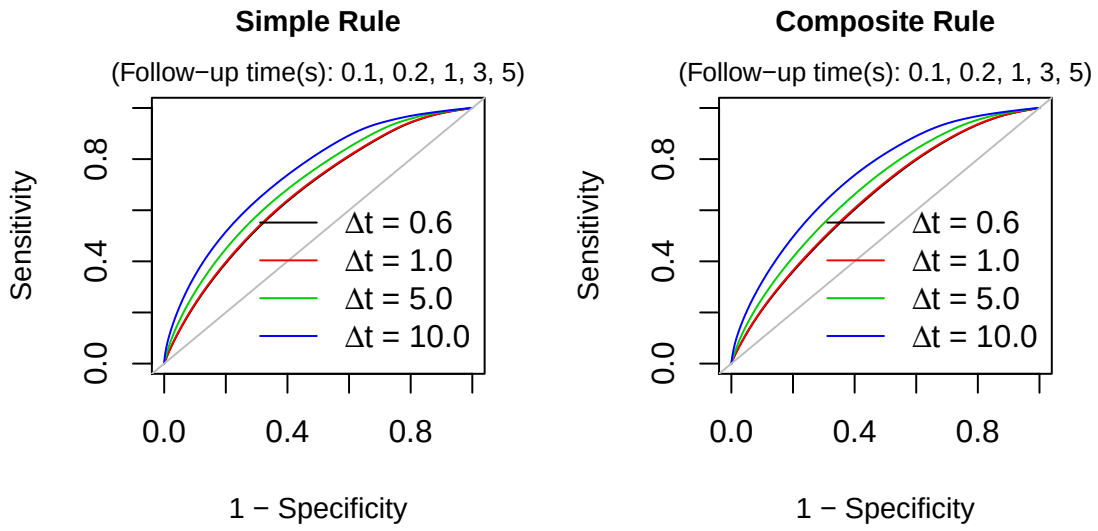

Figure 6.12: ROC curves at time $t = 5$ and four options for Δt under the simple (left) and composite (right) prediction rule, assuming for the second plot a 20% decrease in log(eGFR) levels between visits. The calculation is based on *joint.bs* model for the ANZdataset.

To investigate the influence of the joint model parameterization, on the predictions for the survival outcome, we explored discrimination power for the alternative slope model we formulated before.

```
##
## Areas under the time-dependent ROC curves
##
## Estimation: Monte Carlo (500 samples)
## Difference: relative, lag = 2 (1, 0.8)
## Thresholds range: (-2.28, 11.69)
##
## Case: 1
## Recorded time(s): 0.08, 0.25, 1, 3, 5
##    dt t + dt    AUC Cut.1 Cut.2
##   0.6    5.6 0.4935 4.815 3.852
##   1.0    6.0 0.4912 4.815 3.852
##   5.0   10.0 0.4692 4.815 3.852
##  10.0   15.0 0.3484 4.927 3.942
```

We observe that the parameterization that combines both the current value term $m_i(t)$ and the slope term $m_i^{'}(t)$ seems to have worst discrimination compared to first model presented. More surprisingly, in this case we see that AUCs tend to decrease increasing dt lengths.

Chapter 7

Conclusions

7.1 Discussion

In this thesis we reviewed joint models and their extensions as methods for inference on the possible association between a longitudinal and an event process. In contrast to classical survival models, joint models arise as appropriate technique to model time-dependent covariates with informative dropouts. This is achieved by jointly maximizing a likelihood constructed using both the longitudinal and time-to-event data.

In our work we were interested in the association between the eGFR biomarker an the composite survival event, graft-failure or death, in kidney transplant recipients. We fitted a series of suitable models to investigate if a significant association was present. We started with classical survival approaches, that have been demonstrated to give biased results in most of the analysis when a time-varying *internal* covariate, as a biomarker, is taken into account. We also fitted different versions from the joint model we have assumed. We varied, for example, the parameterization of the association between log(eGFR) and its evolution in time of follow-up in the longitudinal submodel, the baseline hazard parameterization in the survival submodel and also formulated a joint model including as covariate the slope of longitudinal outcome in addiction to its current value.

Results of both the two stages approach and joint analyses were consistent and very close each others. A significant strong association was found between the processes. Also, the eGFR biomarker was found to be accurate for predicting survival probabilities in kidney diseases. These results could advice and motivate medical operators so that transplanted patients may benefit from early intervention and dialysis planning. With this in mind and in order to improve kidney research, eGFR could also be used to identify patients most likely to have a decline in kidney function and possibly target them as participants in interventional trials.

An advantage of our study was the vastness of the cohort sample size and of data information, about subjects' marker measurements, we were able to include. As opposite, we had disadvantage in dealing with random shared models where high-dimensional random effects were assumed. This has become a computational issue in the joint model framework.

7.2 Computational issues

Maximum likelihood joint modeling is computationally costly. It involves a combination of a double numerical integration and optimization, requiring lot of machine time to be computed. The R *jointModel()* function locates the maximum likelihood estimates, starting with the EM algorithm and going on eventually with a quasi-Newton algorithm, until convergence. Double optimization and numerical integration required in this setting can be complex and can lead to convergence problems. Likelihood evaluations number increases exponentially with the number of random effects and frailty terms, till the point where the Gauss Hermite Quadrature method is too much computationally expensive for dealing with high-dimensional integration problems [Hof et al., 2017]. In our case, in the linear mixed effects submodel, we considered an high-dimensional vector of functions of time t expressed in terms of splines. Hessian matrix computation and inversion resulted largely demanding, so that high-performance-computing was required. Even that procedure involved lot of time to complete the calculations.

GHQ integration to calculate the individual log-likelihood

$$l_i(\theta) = log \int \big\{ \prod_{j=1}^{n_i} p(y_{ij}|b_i) \big\} \big\{ h_i(T_i|b_i;\theta)_i^{\delta} S_i(T_i|b_i;\theta) \big\} p(b_i;\theta) db_i, \tag{7.1}$$

can be avoided implementing the Bayesian approach as alternative to the classical maximum likelihood estimation.

In the Bayesian framework we write the posterior distribution as

$$p(\theta, b) \propto \prod_{i=1}^{n} \prod_{j=1}^{n_i} p(y_{ij}|b_i,\theta) p(T_i, \delta_i|b_i,\theta) p(b_i,\theta) p(\theta). \tag{7.2}$$

In the formula for $p(T_i, \delta_i|b_i,\theta)$, the integral in the definition of the survival function

$$S_i(t|M_i(t), w_i) = \exp\, - \int_0^t h_0(s;\theta) \exp\big\{\gamma^T w_i + \alpha m_i(s)\big\} ds, \tag{7.3}$$

still doesn't have a closed form and it requires to use numerical methods such as e Gauss-Kronrod and Gauss-Legendre quadrature rules. Bayesian approach is based on Markov chain Monte Carlo (MCMC) algorithms where each Markov chain iteration depends on the previous one. Even with this method computational costs cannot be cut down, however papers using Bayesian estimation for joint models with high-dimensional random effects are very frequent in literature (see [Brown et al.], [Köhler et al.], [Andrinopoulou et al., 2017]).

7.3 Outlook

Our work could extended in various ways. First of all we could investigate the Competing Risk setting for graft-failure outcome alone. Bayesian estimation efficiency can also be analyzed as compared to maximum likelihood. A more flexible model can be formulated in the forms presented in Köhler et al. as for example testing a

time-dependent association effect $\alpha(t)$ in joint models or by including automatic non linear or functional formulation for both the longitudinal process and the association between the two processes. An easier extensions could be to incorporate possible confounding covariates in the model and eventually considering a multivariate longitudinal outcome composed by multiple biomarkers. As a further step in general joint models computation, we would like to encourage the implementation in package **JM** of utilities of *s()* function for smooth terms,from **mgcv** library and a test to possibly determine whether an extra random effects should be or not included in the model.

Bibliography

Odd Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726, 1978.

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

Eleni-Rosalina Andrinopoulou, D Rizopoulos, Johanna JM Takkenberg, and E Lesaffre. Combined dynamic predictions using joint models of two longitudinal outcomes and competing risk data. *Statistical Methods in Medical Research*, 26(4):1787–1801, 2017. doi: 10.1177/0962280215588340. PMID: 26059114.

Eleni-Rosalina Andrinopoulou, Dimitris Rizopoulos, Johanna J. M. Takkenberg, and Emmanuel Lesaffre. Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in Medicine*, 33(18):3167–3178. doi: 10.1002/sim.6158. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6158`.

Jessica Barrett, Peter Diggle, Robin Henderson, and David Taylor-Robinson. Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):131–148, 2015. ISSN 1467-9868. doi: 10.1111/rssb.12060. URL `http://dx.doi.org/10.1111/rssb.12060`.

Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723, 2005.

M Berk. sme: Smoothing-splines mixed-effects models. *R package version 0.8. h. See https://CRAN. R-project. org/package= sme*, 2013.

Paul Blanche, Cécile Proust-Lima, Lucie Loubère, Claudine Berr, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*, 71(1):102–113, 2015. ISSN 1541-0420. doi: 10.1111/biom.12232. URL `http://dx.doi.org/10.1111/biom.12232`.

Elizabeth R. Brown, Joseph G. Ibrahim, and Victor DeGruttola. A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61(1):64–73.

David R Cox. Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer, 1992.

Michael J Crowther and Paul C Lambert. Simulating biologically plausible complex survival data. *Statistics in medicine*, 32(23):4118–4134, 2013.

Carl De Boor, Carl De Boor, Etats-Unis Mathématicien, Carl De Boor, and Carl De Boor. *A practical guide to splines*, volume 27. Springer-Verlag New York, 1978.

Victor De Gruttola and Xin Ming Tu. Modelling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics*, pages 1003–1014, 1994.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Peter Diggle. *Analysis of longitudinal data*. Oxford University Press, 2002.

Alessio Farcomeni and Sara Viviani. Longitudinal quantile regression in the presence of informative dropout through longitudinal–survival joint modeling. *Statistics in Medicine*, 34(7):1199–1213, 2015. ISSN 1097-0258. doi: 10.1002/sim.6393. URL `http://dx.doi.org/10.1002/sim.6393`.

Cheryl L Faucett and Duncan C Thomas. Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach. *Statistics in medicine*, 15(15):1663–1685, 1996.

Peter J Green and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press, 1993.

Lyle C Gurrin, Katrina J Scurrah, and Martin L Hazelton. Tutorial in biostatistics: spline smoothing with linear mixed models. *Statistics in medicine*, 24(21):3361–3381, 2005.

Robin Henderson, Peter Diggle, and Angela Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480, 2000.

Graeme L. Hickey, Pete Philipson, Andrea Jorgensen, and Ruwanthi Kolamunnage-Dona. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Medical Research Methodology*, 16(1):117, Sep 2016. ISSN 1471-2288. doi: 10.1186/s12874-016-0212-5. URL `https://doi.org/10.1186/s12874-016-0212-5`.

M. H. Hof, J. Z. Musoro, R. B. Geskus, G. H. Struijk, I. J. M. ten Berge, and A. H. Zwinderman. Simulated maximum likelihood estimation in joint models for multiple longitudinal markers and recurrent events of multiple types, in the presence of a terminal event. *Journal of Applied Statistics*, 44(15):2756–2777, 2017. doi: 10.1080/02664763.2016.1262336. URL `https://doi.org/10.1080/02664763.2016.1262336`.

Fushing Hsieh, Yi-Kuan Tseng, and Jane-Ling Wang. Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics*, 62(4):1037–1043, 2006. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2006.00570.x. URL http://dx.doi.org/10.1111/j.1541-0420.2006.00570.x.

Joseph G Ibrahim, Haitao Chu, and Liddy M Chen. Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 28(16):2796, 2010.

Nicholas P Jewell and Jens P Nielsen. A framework for consistent prediction rules based on markers. *Biometrika*, 80(1):153–164, 1993.

John D Kalbfleisch and Ross L Prentice. Relative risk (cox) regression models. *The Statistical Analysis of Failure Time Data, Second Edition*, pages 95–147, 2002.

Meike Köhler, Nikolaus Umlauf, Andreas Beyerlein, Christiane Winkler, Anette-Gabriele Ziegler, and Sonja Greven. Flexible bayesian additive joint models with an application to type 1 diabetes research. *Biometrical Journal*, 59(6):1144–1165. doi: 10.1002/bimj.201600224. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201600224.

Kenneth Lange. Elementary optimization. In *Optimization*, pages 1–17. Springer, 2004.

Roberto Marcén, José María Morales, Ana Fernández-Rodriguez, Luis Capdevila, Luis Pallardó, Juan José Plaza, Juan José Cubero, Josep María Puig, Ana Sanchez-Fructuoso, Manual Arias, et al. Long-term graft function changes in kidney transplant recipients. *NDT plus*, 3(suppl_2):ii2–ii8, 2010.

Marlena Maziarz, Patrick Heagerty, Tianxi Cai, and Yingye Zheng. On longitudinal prediction with time-to-event outcome: Comparison of modeling options. *Biometrics*, 73(1):83–93, 2017. ISSN 1541-0420. doi: 10.1111/biom.12562. URL http://dx.doi.org/10.1111/biom.12562.

Magdalena Murawska, Dimitris Rizopoulos, and Emmanuel Lesaffre. A two-stage joint model for nonlinear longitudinal response and a time-to-event with application in transplantation studies. *Journal of Probability and Statistics*, 2012, 2012.

R. L. Prentice. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69(2):331–342, 1982. doi: 10.1093/biomet/69.2.331. URL +http://dx.doi.org/10.1093/biomet/69.2.331.

Cécile Proust-Lima, Pierre Joly, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Joint modelling of multivariate longitudinal outcomes and a time-to-event: A nonlinear latent class approach. *Computational Statistics Data Analysis*, 53(4):1142 – 1154, 2009. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2008.10.017. URL http://www.sciencedirect.com/science/article/pii/S0167947308004830.

Hein Putter, Marta Fiocco, and Ronald B Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine*, 26(11):2389–2430, 2007.

Feiyou Qiu, Catherine M Stein, Robert C Elston, and Tuberculosis Research Unit (TBRU). Joint modeling of longitudinal data and discrete-time survival outcome. *Statistical methods in medical research*, 25(4):1512–1526, 2016.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

D. Rizopoulos. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Chapman & Hall/CRC Biostatistics Series. CRC Press, 2012. ISBN 9781439872871. URL `https://books.google.com.au/books?id=E__RBQAAQBAJ`.

D. Rizopoulos, M. Murawska, E.-R. Andrinopoulou, G. Molenberghs, J. J. M. Takkenberg, and E. Lesaffre. Dynamic Predictions with Time-Dependent Covariates in Survival Analysis using Joint Modeling and Landmarking. *ArXiv e-prints*, June 2013.

Dimitris Rizopoulos, Geert Molenberghs, and Emmanuel M.E.H. Lesaffre. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal*, 59(6):1261–1276, 2017. ISSN 1521-4036. doi: 10.1002/bimj.201600238. URL `http://dx.doi.org/10.1002/bimj.201600238`.

Josefina Santos and La Salete Martins. Estimating glomerular filtration rate in kidney transplantation: Still searching for the best marker. *World journal of nephrology*, 4(3):345, 2015.

Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 03 1978. doi: 10.1214/aos/1176344136. URL `https://doi.org/10.1214/aos/1176344136`.

Steve Self and Yudi Pawitan. Modeling a marker of disease progression and onset of disease. In *AIDS epidemiology*, pages 231–255. Springer, 1992.

Michael J Sweeting and Simon G Thompson. Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, 53(5):750–763, 2011.

An-Min Tang, Xingqiu Zhao, and Nian-Sheng Tang. Bayesian variable selection and estimation in semiparametric joint models of multivariate longitudinal and survival data. *Biometrical Journal*, 59(1):57–78, 2017.

A. A. Tsiatis, Victor Degruttola, and M. S. Wulfsohn. Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association*, 90

(429):27–37, 1995. doi: 10.1080/01621459.1995.10476485. URL http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476485.

Anastasios A Tsiatis and Marie Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004.

Hans van Houwelingen and Hein Putter. *Dynamic prediction in clinical survival analysis*. CRC Press, 2011.

Ye Wen, Lin Xihong, and Taylor Jeremy M. G. Semiparametric modeling of longitudinal measurements and time-to-event data–a two-stage regression calibration approach. *Biometrics*, 64(4):1238–1246. doi: 10.1111/j.1541-0420.2007.00983.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2007.00983.x.

Simon Wood. *Generalized Additive Models (Texts in Statistical Science)*. Chapman & Hall/CRC, 2006. ISBN 1584884746.

Michael S. Wulfsohn and Anastasios A. Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53(1):330–339, 1997a. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2533118.

Michael S Wulfsohn and Anastasios A Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339, 1997b.

Hok Pan Yuen and Andrew Mackinnon. Performance of joint modelling of time-to-event data with time-dependent predictors: an assessment based on transition to psychosis data. *PeerJ*, 4:e2582, 2016. ISSN 2167-8359. doi: 10.7717/peerj.2582.

Yingye Zheng and Patrick J. Heagerty. Partly conditional survival models for longitudinal data. *Biometrics*, 61(2):379–391, 2005. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2005.00323.x. URL http://dx.doi.org/10.1111/j.1541-0420.2005.00323.x.

R version and packages used to generate this report:

R version: R version 3.3.1 (2016-06-21)

Base packages: splines, stats, graphics, grDevices, utils, datasets, methods, base

Other packages: dynpred, lcmm, JM, MASS, ICEbox, sfsmisc, nlme, lattice, ggplot2, biostatUZH, survival, reporttools, xtable, RColorBrewer, stringr, magrittr, dplyr, readr, data.table, sas7bdat, knitr

This document was generated on maggio 21, 2018 at 09:51.