USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

UNIVERSIDADE DA CORUÑA

Universidade Vigo

Master's Thesis

Joint Modelling for Longitudinal and Time-to-Event Survival. Applications to Biomedical Data

Arís Fanjul Hevia

Master in Statistical Techniques

2014-2015

Propuesta de Trabajo Fin de Máster

Título en español: Modelos conjuntos para datos longitudinales y análisis de supervivencia. Aplicación a datos biomédicos
English title: Joint modelling for longitudinal and time-to-event survival. Applications to biomedical data
Modalidad: A
Autora: Arís Fanjul Hevia, Universidad de Santiago de Compostela
Directora: Carmen Cadarso Suárez, Universidad de Santiago de Compostela
Breve resumen del trabajo: La intención de este proyecto es ilustrar cómo se pueden combinar el análisis de supervivencia, los datos longitudinales y los riesgos competitivos en un mismo modelo. Además de ver las ventajas que aporta este modelado conjunto, se aplicarán las técnicas estudiadas a unos datos procedentes de un programa de diálisis peritoneal.
Recomendaciones:
Otras observaciones:

Doña Carmen Cadarso Suárez, Profesora de Estadística e Investigación Operativa de la Universidad de Santiago de Compostela informa que el Trabajo Fin de Máster titulado

Joint Modelling for Longitudinal and Time-to-Event Survival. Applications to Biomedical Data

fue realizado bajo su dirección por doña Arís Fanjul Hevia para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 8 de julio de 2015.

La directora:	La autora:
Doña Carmen Cadarso Suárez	Doña Arís Fanjul Hevia

Contents

Abstract

Joint modelling of longitudinal and survival data has received much attention in the last years and is becoming increasingly used in clinical follow-up programs. Such biomedical studies usually include longitudinal measurements that cannot be considered in a survival model with the standard methods of survival analysis. Furthermore, that kind of studies can also present more than one possible endpoint, meaning that they have to cope survival analysis with longitudinal data and in the presence of competing risks. Although some joint models have been adapted in order to allow for competing endpoints, this methodology has not been widely disseminated in medical practice. In this project we aim to show how to combine in the same framework survival analysis, longitudinal data, and competing risk, as well as the advantages of the resulting joint model. All those techniques will be applied in the analysis of a database from a peritoneal dialysis program of the Peritoneal Dialysis Unit of the Hospital Geral de Santo António (Portugal).

Resumen en español

El modelado conjunto de análisis de supervivencia con datos longitudinales ha recibido mucha atención en los últimos años. Su uso se ha ido extendiendo cada vez más en estudios clínicos de seguimiento, ya que en ellos solemos encontrar datos longitudinales para los cuales las técnicas habituales de análisis de supervivencia no siempre resultan adecuadas. Además, en este tipo de estudios también se puede dar más de un evento de fallo, lo que conlleva la necesidad de utilizar riesgos competitivos a la hora de analizar la supervivencia. A pesar de que se han adaptado diversos modelos conjuntos para incluir la presencia de estos riesgos competitivos, se trata de una metodología poco difundida. La intención de este proyecto es ilustrar cómo se pueden combinar el análisis de supervivencia, los datos longitudinales y los riesgos competitivos, así como las ventajas que aporta este modelado conjunto. Las técnicas estudiadas se aplicarán a unos datos procedentes de un programa de diálisis peritoneal, realizado en la Unidad de Dialysis Peritoneal del Hospital Geral de Santo António (Portugal).

Chapter 1

Introduction

Biomedical studies have always been a source of inspiration for the statistic field: they provide with data with specific features that need special caution when doing the analysis, and they keep coming up with situations where new statistical tools have to be developed in order to be able to handle them.

This is what happens in follow-up studies: they include several longitudinally measured responses, not always taken at the same time or even with the same number of measurements, and they may also come with different types of outcomes. In general, this kind of data can not be analyzed with standard statistical procedures.

Over the last few decades, since the World War II motivated the study in the reliability of the military equipment, the survival analysis (Cox, 1972) has been a very important field of research: it studies the time until an event of particular interest occurs, and with it, it answers questions like what kind treatment is better for a certain illness, or what variables have an influence in the recovery of a patient.

On the other hand, when taking in to account longitudinal data that comes from follow-up studies, the survival analysis becomes complicated: time-dependent variables can be related to the failure mechanism under study, and this lack of independence causes many problems. It is then when we include the mixed-effects models (Harville, 1977; Laird and Ware, 1982; Verbeke and Molenberghs, 2000) into the equation. Together, they build the joint modelling approach, a model that has become increasingly popular in clinical studies in the last years.

In addition to this, in the disease or recovery process that are examined in the survival situations , often more than one type of event plays a role. Even if one type of event can be singled out as the event of interest, the others may prevent that specific type of failure from occurring. These kind of events are called competing risks (Beyersmann et

al. 2012), and special caution is needed in these cases, for their presence, if not taken into account, may produce some bias in the estimation of the event of interest.

Joint modelling in competing risk framework, despite not being as widely used in medical context as the basic joint modelling or the standard competing risk model, has recently motivated a serial of studies on the topic. Our goal in this project is then to present a model suited for analyzing longitudinal data in the survival analysis field in the presence of competing risks (Rizopoulos, 2012). In particular, we aim to analyze the data from a peritoneal dialysis program, where the presence of a longitudinal outcome repeatedly registered along the follow-up time and the occurrence of several specific events is common.

The progression of end-stage renal disease patients included in a peritoneal dialysis program is monitored with regular control visit where several clinical parameters are recorded, as well as the time until the occurrence of endpoints. Then, as in many other clinical research areas, peritoneal dialysis patients data present different types of outcomes: apart from the baseline data recorded at the beginning of the study (like the sex or the age of the patient), they also present longitudinal outcome measured at several time points (such as albumin, glucose and phosphorus) and time-to-event outcome, composed of the follow-up time until the occurrence on an event of interest (which in this case will be death, transfer to hemodialysis or renal transplant). As an example, Figure 1.1 gathers the longitudinal profiles of albumin of 16 subjects in a peritoneal dialysis program.

The outline of this work will be as follows. Firstly, in Chapter 2 we will introduce the blocks of longitudinal data analysis and survival analysis, explaining the basic features of both matters. Secondly, a review of competing risks is made in Chapter 3. The joint models approach for longitudinal and time-to-event data is then presented in Chapter 4, along with the final model in which all three blocks (longitudinal data analysis, survival analysis and competing risks) are considered. This structure is summarized in Figure 1.2.

Eventually a detailed analysis of the results obtained by applying all this methodology to the peritoneal dialysis data is shown in Chapter 5, followed by a final chapter where conclusions and future lines of research are discussed.

All the analysis that have been performed in this project have been implemented in the `R` software environment. At the end of Chapter 5 we include a brief discussion on the available software on this matter.

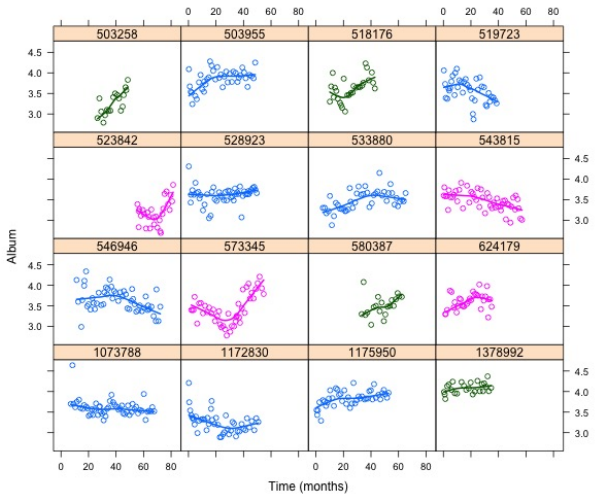

Figure 1.1: Albumin longitudinal profiles of 16 subjects. The different colors show the kind of failure that each of them presented (green for transplant, pink for death or transfer to hemodialysis and blue for the ones that did not suffer any of the above).

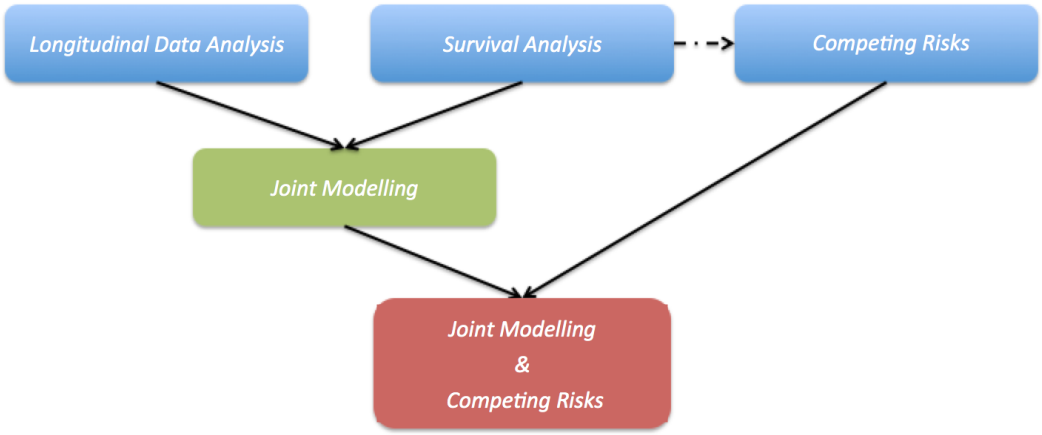

Figure 1.2: Diagram of the methodology that will be introduced in this project.

Chapter 2

Statistical Background

This chapter introduces the two blocks that we need to understand the joint modelling approach, the first step in our path to be able to analyze our data with longitudinal measurements and different causes of failure.

In the first place, we explain the basic concepts of the linear mixed-effects model, the tool that is most frequently used when managing longitudinal data. The second section is dedicated to the survival analysis, where apart from introducing the functions of interest we give special attention to the handling of time-dependent covariates.

2.1. Longitudinal data analysis: the linear mixed-effects model

Our focus on this first section is on *longitudinal data*, which can be defined as the data resulting from the observations of subjects that are measured repeatedly over time. Such data is frequently encountered in health studies, related to human beings, animals or laboratory samples.

For example, in a longitudinal study in which patients are randomly assigned to take different treatments and are followed up over time, we could investigate how treatment means differ at specific time points (cross-sectional effect) or how those means change over time (longitudinal effect). Another example is the recounting of CD4+ cells: this kind of cells are affected by the AIDS virus as their number decreases with the development of the illness. Therefore, their longitudinal study is very important.

Measuring subjects repeatedly through the duration of the study, we expect positive correlation, which means that standard statistical tools (like the t-test and simple regression) that assume independent observation, are not appropriate for this kind of

data analysis.

Thus, we will introduce the linear mixed-effects model for the analysis of continuous longitudinal responses, which constitutes the first block of joint modelling.

2.1.1. Linear Mixed-Effects Models

An intuitive approach for the analysis of longitudinal data is based on the idea that each individual in the population has its own subject-specific mean response profile over time, with a functional form. In Figure 2.1 we can see a graphical representation of this idea: the longitudinal responses of two hypothetical subjects (points), their corresponding linear mean trajectories (dashed lines) and the average evolution (solid line).

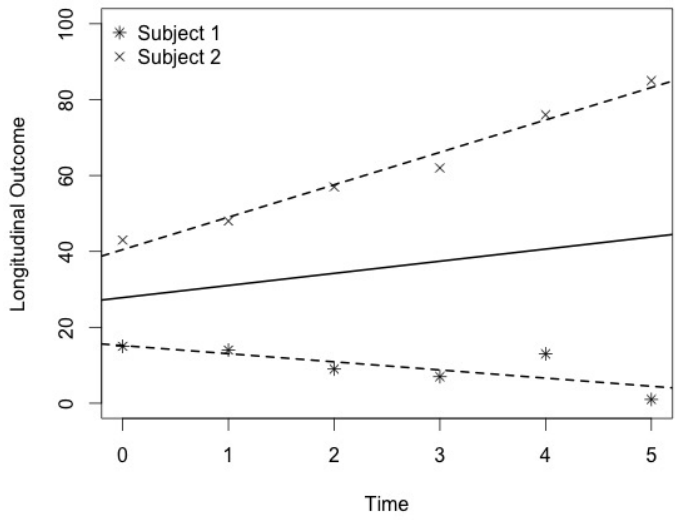

Figure 2.1: Longitudinal responses of two subjects in a simulated longitudinal study.

To introduce this representation, let y_{ij} denote the response of subject i, $i = 1, ..., n$ at time t_{ij}, $j = 1, ..., n_i$. A plausible model for the observed responses y_{ij} in Figure 2.1, taking into account that a linear time effect seems adequate and that different subjects tend to have different intercepts and slopes, could be:

$$y_{ij} = \tilde{\beta}_{i0} + \tilde{\beta}_{i1} t_{ij} + \varepsilon_{ij},$$

where the error terms ε_{ij} are assumed to come from a normal distribution with mean

zero and variance σ^2, and $\tilde{\beta}_{i0}$ and $\tilde{\beta}_{i1}$ are the regression coefficients. It is usually assumed that the distribution of the regression coefficients in the population is a normal bivariate normal distribution with mean vector $\beta = (\beta_0, \beta_1)^T$ and variance-covariance matrix D. We can then reformulate the model as:

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

where $\tilde{\beta}_{i0} = \beta_0 + b_{i0}$ and $\tilde{\beta}_{i1} = \beta_1 + b_{i1}$. The terms $b_i = (b_{i0}, b_{i1})^T$ are called *random effects*, following a bivariate normal distribution $N(0, D)$, and they describe the variability of each individual. On the other hand, the parameters β_0 and β_1 describe the average longitudinal evolution in the population and are called *fixed effects*. Moreover, b_i are assumed to be independent of the error tems ε.

The generalization of this model, allowing additional predictors and regression coefficients to vary randomly, is known as the *linear mixed-effects model* (Laird and Ware, 1982; Harville, 1997; Verbeke and Molenberghs, 2000):

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

where X_i and Z_i are known design matrices for the fixed-effects regression coefficients β and the random-effects regression coefficients b_i respectively, and I_{n_i} denotes the n_i-dimensional identity matrix. The random effects, apart from being assumed to be normally distributed, are taken as independent of the error ε_i, i.e., $cov(b_i, \varepsilon_i) = 0$.

Equivalently, we can express the linear mixed models with this form:

$$\begin{cases} y = X\beta + Zb + \varepsilon, \\ b \sim N(0, D), \\ \varepsilon \sim N(0, R), \end{cases}$$

where X is now a $n \times p$ matrix, with p the number of fixed effects, and Z is a $n \times k$

matrix, with k the number of random effects. Moreover, $R = \sigma^2 I_{n \times n_i}$, with $I_{n \times n_i}$ a $(n \times n_i)$-dimensional identity matrix.

The interpretation of the fixed effects is the same as in a simple linear regression model: assuming we have p covariates in the design matrix X, the coefficient β_j , $j = 1, ..., p$ denotes the change in the average y_i when the corresponding covariate x_j is increased by one unit, while all other predictors are held constant. In the same way, b_i show how a subset of the regression parameters for the ith subject deviates from those in the population.

With the mixed models we are able to estimate parameters that describe how the mean response changes in the population of interest (the fixed-effects) and it is also possible to predict how individual response trajectories change over time (the random-effects). Another advantage is that mixed models can work with unbalanced data: we do not need the same number of measurements on each subject nor that these measurements be taken at the same set of occasions.

Random intercepts and slopes

Depending on the behavior of the response y_i of each subject, we can adjust different kinds of mixed models. In particular, this adjustment will affect the design matrix Z_i, $i = 1, ...n$:

- If Z_i contains a column of 1's, it means we are considering a random intercept b_{i0}. This kind of random effect is used when we observe different departure points of the longitudinal response for each subject.

- In the case where we want to consider a random slope b_{i1}, Z_i has to contain a column with the times of every visit of the patient. This means that each subject has a different temporal slope than the others.

- The rest of the covariates included in Z_i will indicate a random effect that is different for each subject for every covariate.

In Figure 2.2 we can appreciate several models considering different configurations of Z_i. In each graphic the longitudinal response under study is represented separately for every subject.

In the first one, the model includes a random intercept and a null slope: we can appreciate that the response of every subject starts with a different value, and that it does not change much over time. In the second one, we have a random intercept again,

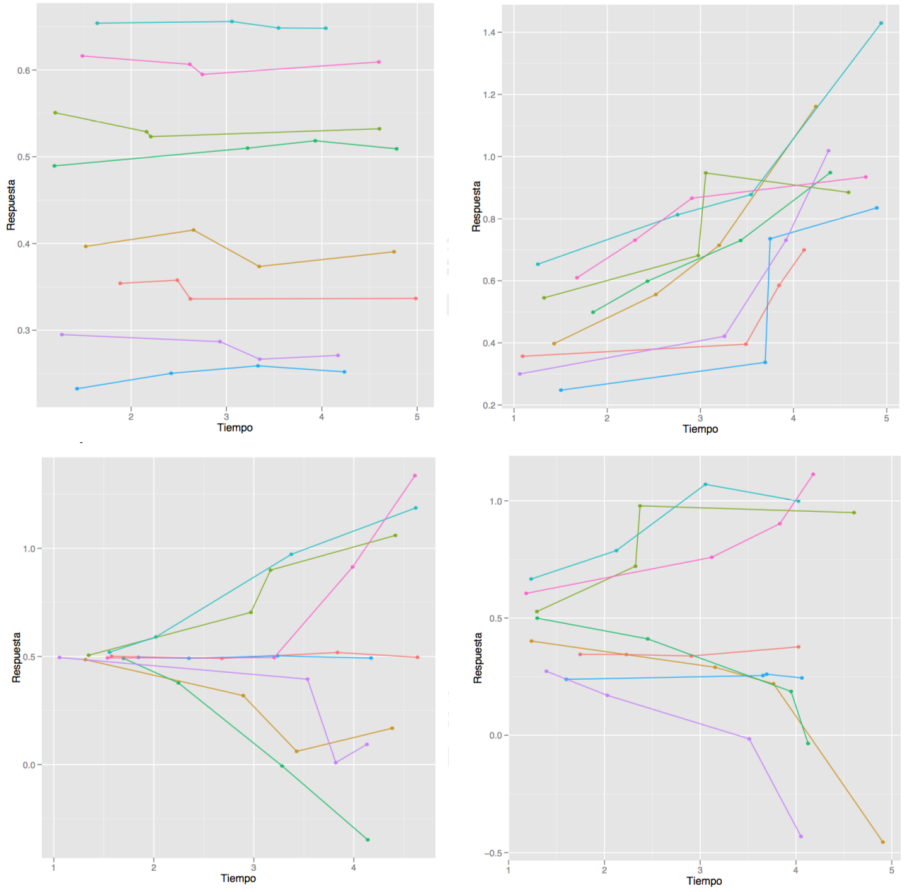

Figure 2.2: Different longitudinal models considering Z_i configuration: the first one has random intercepts and a null slope; the second one has random intercepts too and a non-random positive slope; the third one has random slope but common intercept, and the last one has random intercepts and slopes.

and the slope (though different from zero this time) is still common for all subjects, as their paths stay parallel. The third one is the opposite to the first graphic: here there is no random intercept, as all the responses start at the same level, but there is a random slope: the individuals trajectories differ from each other. The last one shows the case where we have both random intercepts and slopes: we can see that each of them start at a different point and then follow different trajectories.

2.1.2. Estimation of the Linear Mixed-Effects Models

Fixed-effects estimation

One way of obtaining one estimation of β is using the marginal model:

$$y_i = Xi\beta_i + \varepsilon_i^*, \quad \text{with} \quad \varepsilon_i^* = Z_i b_i + \varepsilon_i.$$

This model has correlated errors, with

$$cov(\varepsilon_i^*) = V_i = Z_i D Z_i^T + \sigma^2 I_{n_i}.$$

If we assume that V_i is known, minimizing the function $Q = (y - X\beta)'V^{-1}(y - X\beta)$, we obtain the generalized least squares estimator for β:

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

which, on the other hand, is the same as the maximum likelihood estimator of the fixed-effects vector β.

Random-effects prediction

Being random variables, we can not speak about random-effects estimation, but instead we are able to predict them. There are different ways of obtaining this predictions. One of the best linear unbiased predictor can be yielded using Henderson's mixed model equations (Henderson et al., 2000). It is a procedure that allows us to calculate the best linear unbiased estimator for $X\beta$ and the best linear unbiased predictor for b. It considers the joint density distribution of y and b, and the log-likelihood of the linear model.

Henderson's mixed model equation are

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

and their solutions:

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

We can not forget that we are doing this calculations under the assumption that we know the parameters of the covariance matrix V, but it is usually not the case. Therefore, the next step is to estimate them. The two more common ways of doing this are the maximum likelihood (ML) and the restricted maximum likelihood (REML).

The principal disadvantage of ML is that it is biased for small samples, due to the fact that the ML estimate does not take into account that β is estimated from the data as well as V. On the contrary, the REML estimates the variance components based on the residuals obtained after the estimation of the fixed effects $(y - X\beta)$, and therefore, if the sample is small, it will yield better estimates than the ML.

Neither of those methods have, in general, a close form, so in order to obtain $\hat{V}$ a numerical optimization routine, such as the Expectation-Maximization (Dempster et al., 1977) or the Newton-Raphson algorithms (Lange, 2004) are needed.

2.2. Survival analysis: analysis of time-to-event data

The *survival analysis* is defined as a set of statistical procedures that study non-negative random variables associated with the time span from some time origin until the occurrence of one event of interest. This event is usually called *failure*, as it is associated with death in biological studies, and the random variable is called *failure time*, *survival time* or *event time*.

There are lots of examples of failure times: the time until the death of one patient, the time of convalescence, the time until some new skill is learned... Although it is used in several fields, here we will focus in its applications to the biomedicine, like its use in a clinical follow-up study.

A very important feature of this kind of data is that we don't always know the failure time of every subject: sometimes part of the disease history is unobserved. If the endpoint of interest has not occurred at the end of the observation window (due to lost to follow up or drop out of the study, or if the study ends before an outcome of interest

happens), the event time is *right censored*. This characteristic makes inadvisable the use of the standard statistical tools to analyze this type of data.

There are different classifications for censoring mechanism: we could have either left- or right-censoring (when the survival time is less or greater than the observation time) and interval-censored data (in which the time to the event of interest is known to occur between to certain time points); we could also distinguish between informative censoring (which occurs when the subject withdraws from the study for reasons related to the expected failure time) and non-informant censoring (when those reasons are unrelated to the study).

In this section we will focus on the non-informative right censoring, but informative censoring will play an important role in the next chapter. On the other hand, the left truncated data, in which the individuals have a delayed entry in the study, will be considered.

In addition to this, the subjects under this type of studies are usually heterogeneous. This means that one our goals will be identifying the variables that have an influence in the survival.

2.2.1. Functions of interest

Let T^* denote the non-negative random variable of failure time. In the context of survival analysis, an individual i is represented by the pair (T_i, δ_i), where T_i denotes the observed event time for subject i ($T_i = \text{mín}\{T_i^*, C_i\}$, with C_i the censoring time) and δ_i is a variable that indicates if the individual has experienced the event ($\delta_i = 1$) or not (in this other case the observation is censored and $\delta_i = 0$).

The function that is primarily used to describe the distribution of T^* is the *survival function*. If the event under study is death, it expresses the probability that death occurs after an instant t, that is, the probability of surviving time t. Assuming that T^* is continuous, the survival function is defined as

$$S(t) = P(T^* > t) = 1 - F(t) = \int_t^\infty p(s)d(s), \quad t \geq 0.$$

where $p(\cdot)$ denotes the corresponding probability density function. This density function can be interpreted as the individual probability of observing an event in a certain instant in time. The survival function must be non-increasing as t increases, with $S(t = 0)$ always equal to one.

Another function that plays a prominent role in survival analysis is the *hazard*

function. This one describes the instantaneous risk for an event in the time interval $[t, t + \Delta t)$ provided survival up to t, and is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T^* < t + \Delta t | T^* \geq t)}{\Delta t}, \quad t > 0.$$

The hazard completely describes the survival distribution: it can be derived from the survival function: Likewise, the survival also can be expressed in terms of the risk function.

$$h(t) = -\frac{d \log S(t)}{dt}; \qquad S(t) = \exp\left\{-\int_0^t h(s)ds\right\} = \exp\{-H(t)\}, \quad t > 0,$$

where $H(t)$ is known as the *cumulative risk* (or *cumulative hazard*) function that describes the accumulated risk until time t. It can also be interpreted as the expected number of events to be observed by time t.

2.2.2. Survival estimation

When we are interested in estimating these functions from a random sample (T_1, δ_1), ...,(T_n, δ_n), censoring must be taken into account. The most well-known estimators of both functions are the Kaplan-Meier and the Nelson-Aalen estimator.

- *Kaplan-Meier estimator*

  To introduce this estimator, proposed by Kaplan and Meyer (1958), let $t_1 < ... < t_N$ denote the unique event times in the sample at hand. For each t_i, define d_i to be the number of observed events at t_i, and r_i the number of subjects still at risk at that moment.

  The Kaplan-Meier estimator assumes that the distribution is discrete instead of continuous, with the events only occurring at these observed time points. It considers the conditional probability of failing at t_i, given still alive just before time t_i. This probability can be written as

  $$h(t_i) = P(T = t_i | T > t_{i-1}),$$

  a discretized form of the hazard function given before.

  Under the assumption of uninformative censoring, subjects at risk are representative for all subjects alive just before t_i, so $h(t_i)$ can be estimated simply by the

at risk sample proportion that fail at t_i:

$$\hat{h}(t_i) = \frac{d_i}{r_i}.$$

Using the law of total probability, the probability of surviving at any time t can be written as the product of the conditional probabilities:

$$\begin{aligned} P(T^* > t) &= P(T^* > t | T^* > t-1) P(T^* > t-1) \\ &= P(T^* > t | T^* > t-1) P(T^* > t-1 | T^* > t-2) ... \end{aligned}$$

The probability of surviving up to t_i is then the product of the probability of surviving up to t_{i-1} and the conditional probability of surviving up to t_i given still alive beyond t_{i-1}:

$$\hat{S}(t_i) = (1 - \hat{h}(t_i))\hat{S}(t_{i-1}) = \left(1 - \frac{d_i}{r_i}\right) \hat{S}(t_{i-1}).$$

By repeatedly applying this formula one gets the Kaplan-Meier estimator:

$$\hat{S}_{KM}(t) = \prod_{i:t_i \leq t} \frac{r_i - d_i}{r_i}.$$

The Kaplan-Meier estimator is a step function with discontinuities at the observed event times, coinciding with the empirical survival function if there is no censoring. If the sample size increases, this estimate approaches a continuous distribution.

Its consistency has been proved by Peterson (1977), and Breslow and Crowley (1974) have shown that $\sqrt{n}\left(\hat{S}(t) - S(t)\right)$ converges in law to a Gaussian process with expectation 0 and a variance-covariance function , $\hat{S}_{KM}(t)$, that can be calculated using Greenwood's formula,

$$\hat{Var}(\hat{S}_{KM}(t)) = \hat{S}_{KM}(t)^2 \sum_{i:t_i \leq t} \frac{d_j}{r_j(r_j - d_j)},$$

and using asymptotic normality for $\hat{S}_{KM}$ a confidence interval for $S(t)$ can be derived.

- *Nelson-Aalen Estimator*

  The Nelson-Aalen estimator was developed as an alternative nonparametric esti-

mator for the cumulative hazard function.

$$\hat{H}_{NA}(t) = \sum_{i:t_i \leq t} \frac{d_i}{r_i},$$

where r_i and d_i have the same interpretation as for the Kaplan-Meier estimator. It can be intuitively interpreted as the ratio of the number of deaths to the number exposed. Breslow (1972) suggested then the following estimator for the survival function:

$$\hat{S}_B(t) = \exp\{-\hat{H}_{NA}(t)\} = \prod_{i:t_i \leq t} \exp\{-d_i/r_i\}.$$

To derive a confidence interval for $\hat{S}_B(t)$ we estimate teh variance of $\log \hat{H}_{NA}(t)$ using a formula similar to Greenwood's formula.

The two estimators of the survival function are asymptotically equivalent. However, in general the Breslow estimator has uniformly lower variance than the Kaplan-Meier, though it is biased, especially when $\hat{S}(t)$ is close to zero.

2.2.3. Parametric maximum likelihood

Sometimes it is appropriate to assume that the survival function $S(t)$ behaves as a specific parametric form. In this case, the estimation of the parameters of interest is often based on the maximum likelihood method. In particular, let $\{T_i, \delta_i\}$, $i = 1, ...n$, denote the random sample from a distribution function F, parametrized by θ, with the density function $p(t; \theta)$.

In the construction of the likelihood function we need to account for censoring: when a subject failes at time T_i, it contributes $p(T_i, \theta)$ to the likelihood, whereas for a subject who is censored all we know is that he survived up to that moment, and therefore it contributes $S_i(T_i; \theta)$ to the likelihood.

Thus, combining the information from the censored and uncensored observations, we obtain the likelihood function:

$$L(\theta) = \prod_{i=1}^{n} p(T_i, \theta)^{\delta_i} S_i(T_i; \theta)^{(1-\delta_i)}.$$

Taking the log-likelihood

$$l(\theta) = \sum_{i=1}^{n} \delta_i \log p(T_i, \theta) + (1 - \delta_i) \log S_i(T_i; \theta),$$

and using the relations seen before, it can be rewritten in terms of the hazard function as

$$l(\theta) = \sum_{i=1}^{n} \delta_i \log h_i(T_i, \theta) - \int_0^{T_i} h_i(s;\theta) ds.$$

It is clear that all subjects contribute an amount to the log-likelihood equal to $-H_i(T_i;\theta)$, and the subjects who experienced the event additionally contribute the amount of $\log h_i(T_i, \theta)$. Thus, censored observations contribute less information to the statistical inference than uncensored observation, as it could be expected.

Once the log-likelihood has been formulated, there exist several iterative optimization procedures (such as the Newton-Raphson algorithm) that can be used to locate the maximum likelihood estimates $\hat{\theta}$.

2.2.4. Regression methods for censored data

The subjects under this type of survival analysis are hardly ever homogeneous. They have several characteristics, such as age at baseline, sex, randomized treatment... that may or may not affect their survival. This makes it necessary to study the effect of this covariates and to determine which ones influence the most.

There are several methods to relate the outcome to predictors in survival analysis, like the *Cox proportional hazards model* or the *accelerated failure time model*. Here, we will focus on the Cox model (Cox 1972).

In its simplest form, the hazard for a subject with covariate values $w_i^T = (w_{i1}, ..., w_{ip})$ is assumed to be

$$h_i(t|w_i) = h_0(t) \exp\{\gamma^T w_i\},$$

where $\gamma^T = (\gamma_1, ..., \gamma_p)$ is the vector of regression coefficients and $h_0(t)$ is the *baseline hazard* or *baseline risk* function, and corresponds to the hazard function of a subject that has $\gamma^T w_i = 0$.

Note that, writing this model in the log scale,

$$\log h_i(t|w_i) = \log h_0(t) + \gamma_1 w_{i1} + ... + \gamma_p w_{ip},$$

the regression coefficient γ_j, for predictor w_{ij}, denotes the change in the log hazard at any fixed time point t if w_{ij} is increased by one unit while all other predictors are held constant. Analogously, $\exp\{\gamma_j\}$ denotes the ratio of hazards for a subject i with

covariate vector w_i compared to subject k with covariate vector w_k is:

$$\frac{h_i(t|w_i)}{h_k(t|w_k)} = \exp\{\gamma^T(w_i - w_k)\}.$$

If all the covariates of both subjects were equals but for one, j, (and that difference was only one unit), then

$$h_i(t|w_i) = \exp\{\gamma_j\}h_k(t|w_k).$$

That is the reason why it is called a proportional hazards model. It is a semi-parametric model that does not make any assumption for the distribution of the event times, but assumes that the covariates act multiplicatively on the hazard rate.

To determine the relation between the covariates and the survival time it is necessary to estimate the coefficients in γ. One way of doing this would be to assume a parametric distribution for the baseline hazard (like the Weibull distribution) and then estimate the regression coefficients by maximizing the corresponding log-likelihood function. However, Cox (1972) showed that the estimation of those coefficients (the primary parameters of interest) can be estimated without specifying $h_0(\cdot)$.

Thus, assuming all event times are distinct, the parameter vector γ is found by maximizing the partial likelihood, which is a product of a quotient that compares the hazard ratio of the individual with the event at t_i to the hazard of all the individuals at risk at t_i (represented by R_i):

$$pL(\gamma) = \prod_{i=1}^{n} \left[\frac{\exp\{\gamma^T w_i\}}{\sum_{k \in R_i} \exp\{\gamma^T w_k\}} \right]^{\delta_i}.$$

Note that the baseline hazard cancels out. Excluding the censoring terms and taking logarithms, the coefficients may be estimated on the partial log-likehood

$$pl(\gamma) = \sum_{i=1}^{r} \gamma^T w_i - \log\{ \sum_{T_j \geq T_i} \exp(\gamma^T w_j)\}.$$

Even though this is not equivalent to a full log-likelihood, it can be treated as such. The maximum partial likelihood estimators are then found by solving their partial log-likelihood score equations, using in the process some iterative optimization procedures such as the Newton-Raphson algorithm.

On the other hand, the estimate $\hat{\gamma}$ is used in Breslow's estimate of the baseline

hazard and of the cumulative hazard:

$$\hat{h}_0(t) = \frac{1}{\sum_{k \in R_i} \exp(\hat{\gamma}^T w_k)} , \qquad \hat{H}_0(t) = \sum_{i:t_i \leq t} \frac{1}{\sum_{k \in R_i} \exp(\hat{\gamma}^T w_k)}.$$

For further discussion on the matter, we refer to Kalbfleisch and Prentice (2002).

Time-Dependent Covariates

In the risk model just presented we assumed that the hazard depends only on covariates whose value is constant during follow-up. However, in some studies it may also be of interest to investigate wether time-dependent covariates are associated with the risk for an event.

A *time-dependent variable* is defined as any variable whose value for a given subject may change over time. We can distinguish two different categories of time-dependent covariates, namely *external* or *exogenous* and *internal* or *endogenous*.

To introduce these two types of covariates, let $y_i(t)$ denote the covariate vector at time t for subject$i$ and $\mathcal{Y}_i(t) = \{y_i(s), 0 \leq s < t\}$ denote the covariate history up to t. Those categories require a different treatment, so it is very important to distinguish them.

- *Exogenous covariates*

  A variable is called an exogenous covariate if its value changes because of causes not related to the subject of the study, 'external' characteristics that affect several individuals simultaneously. They satisfy the relation

  $$P(\mathcal{Y}_i(t)|\mathcal{Y}_i(s), T_i^* \geq s) = P(\mathcal{Y}_i(t)|\mathcal{Y}_i(s), T_i^* = s), \ s \leq t,$$

  which means that $y_i(\cdot)$ is associated with the rate of failures over time, but its future path up to any time $t > s$ is not affected by the occurrence of failure at time s. A exogenous covariate is a predictable process, while the endogenous covariates are not, and they do not satisfy that relation.

  An example of an exogenous covariate is the time of the year, the covariates whose complete path is predetermined from the beginning of the study, or environmental factors. The value of these covariates at any time is not affected by the true failure time. For them we can directly define the survival function conditional on the

covariate path, using its relation to the hazard function:

$$S_i(t|\mathcal{Y}_i(t)) = P(T_i^* > t|\mathcal{Y}_i(t)) = \exp\left\{-\int_0^t h_i(s|\mathcal{Y}_i(s))ds\right\}.$$

- *Endogenous covariates*

  On the other hand, the endogenous covariates are time-dependent measurements taken on the subjects under study, such as biomarkers and clinical parameters. They typically require the survival of the subject for their existence, so their path may carry direct information about the failure time. Besides, failure of the subject at time s corresponds to nonexistence of the covariate at $t \leq s$, which violates the endogeneity condition mentioned above. Because of these characteristics, the hazard function is not directly related to a survival function, so the log-likelihood constructions used before will not be appropriate for this type of covariates.

  Another feature of endogenous covariates is that they are usually measured with error and their complete path up to any time is not fully observed: the clinical parameters of a patient are only known for the specific occasions that this patient visited the study center, and not in between these visit times.

Extended Cox Model

The Cox model presented previously can be extended to handle exogenous time-dependent covariates. The intuitive idea behind this formulation is to think about occurrence of events as the realization of a very slow Poisson process.

The extended Cox model, also known as the Andersen-Gill model (1982), is written as

$$h_i(t|\mathcal{Y}_i(t), w_i) = h_0(t)\exp\{\gamma^T w_i + \alpha y_i(t)\},$$

where, as before, $\mathcal{Y}_i(t) = \{y_i(s), 0 \leq s < t\}$, $y_i(t)$ denotes a vector of time-dependent covariates and w_i denotes a vector of baseline covariates (such as sex or randomized treatment). The interpretation of the regression coefficients vector α is the same as for γ. Thus, assuming there is only a single tie-dependent covariate, $\exp(\alpha)$ denotes the relative increase in the risk for an event at time t that results from one unit increase in $y_i(t)$ at that point. Note that, since $y_i(t)$ is time-dependent, the hazard ratio is no longer constant in time. Estimation of γ and α is again based on the corresponding partial log-likelihood function.

This formulation of the Cox model is quite flexible: it allows time-dependent co-

variates, left truncation, multiple time scales... However, it is not appropriate for the time-dependent endogenous covariates.

This is because the extended Cox model assumes that time-dependent covariates are predictable processes, measured without error, with their complete path fully specified, properties that the endogenous covariates do not have. Besides, the time-dependent covariates the extended Cox model handles are assumed to change value at the follow-up visits and remain constant in the time interval in between these visits, and it is evident that this approximation is unrealistic for many endogenous covariates, in concrete for follow-up studies. As the extended Cox model is not able to work with this kind of data, we need new statistical tools to study the time-dependent endogenous covariates.

Chapter 3

Competing Risks Models

In clinical studies it is usual to have more that one event playing an important role in the survival process. Because of that, the independence between the event and censoring distribution, often assumed without further consideration, may easily fail to be true. Reasons for the occurrence of right censored event can be categorized as:

- End of study. In this case is generally safe to assume that the censoring mechanism is independent of disease progressions.
- Loss to follow-up. This type of censoring time can be negatively or positively correlated with the event time.
- *Competing risks*. A competing risk (Beyersmann et al. 2012) is defined as an event that, if it takes places before the outcome under study, it may prevent it from happening.

The censoring time due to loss to follow-up is negatively correlated with the event time when healthy participants of the study feel less need for medical services and therefore quit.This causes a downward bias of the estimated survival curve: it overestimates the probability to experience the event, since individuals with worse prognosis are assumed to be representative for the censored individuals. Furthermore, if the subjects with advanced disease progression get too ill for further follow-up, the censoring time will be positively correlated with the event time, and censoring this individuals will cause a upward bias of the survival curve.

However, the focus in this chapter will be on the competing risks. They concern the situation where more than one cause of failure is possible, and where only the first of these to occur is observed. For example, in a cancer study, death due to cancer

may be of interest, and death due to other causes (surgical mortality, old age) would be considered as competing risks. Alternatively, one could be interested in time of recovery from certain illness, where death due to any cause would be a competing risk.

One way of handling this situations is to single out one of the events and consider the rest of them as censored, but this procedure has a very important problem: doing this, we will be assuming that upon removal of one cause of failure, the risks of failure of the remaining causes is unchanged. This assumption may be reasonable in the industrial setting, but in human studies it will rarely be true. Fortunately, the theory that has been developed over the past two decades for the analysis of right censored survival data can be applied to competing risks models by adding extra adjustments.

We could also be interested in what happens after a non fatal event, and study the transition between different states. These multi-state models are an extension to the competing risks models, but they will not be discussed in this project.

3.1. Approaches to Competing Risks

The competing risks model is usually represented graphically with an initial state (called alive or event-free) and a number or different end points (that correspond with the different events considered), as shown in Figure 3.1.

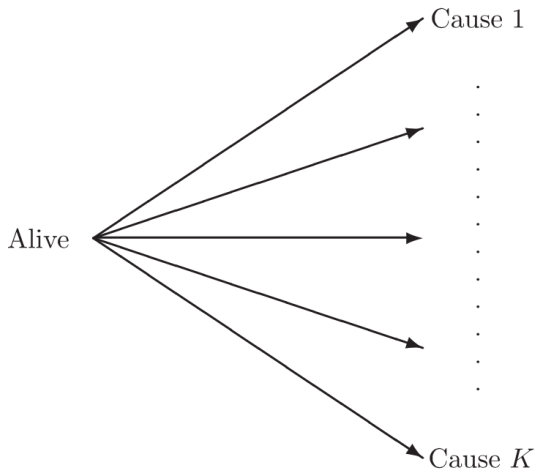

Figure 3.1: A competing risks situation with K causes of failure.

3.1.1. The naive Kaplan-Meier

One way of treating this type of data is to consider the failures from the competing causes as censored observations. The failure probability is then estimated with the Kaplan-Meier estimate, a method called the *naive Kaplan-Meier*. However, this method can be biased: while treating the competing causes as censored, we can be violating one of the assumptions underlying the Kaplan-Meier estimator: the independence of the censoring distribution.

If the competing event time distributions were independent of the distribution of time to the event of interest, this would imply that at each point in time the hazard of the event of interest is the same for subjects that are still under follow-up (alive) as for the subjects that have experienced a competing event by that time. However, a subject that is censored because of failure from a competing risk will not experience the event of interest. The naive Kaplan-Meier will then overestimate the probability of failure (and hence underestimate the survival probability), given that those subjects that will never fail are treated as if they could fail. This bias is greater when the competition is heavier, when the hazard of the competing events is larger.

This is different form censoring due to end of the study or loss to follow-up: in those cases, individuals may still fail at a later point.

3.1.2. Cause-specific Hazard and Cumulative Incidence functions

To handle different failures types we need to extend the notation for the survival process. Assuming K different causes of failure, we let $T_{i1}^*, ..., T_{iK}^*$ denote the true failure times for each one of them. The observed data for the ith subject is composed of the observed event time $T_i = min\{T_{i1}^*, ..., T_{iK}^*, C_i\}$ (with C_i denoting the censoring time) and the event indicator $\delta_i \in \{0, 1, ..., K\}$, where 0 represents the censoring and $1, ..., K$ the competing events.

The fundamental concept in competing risks models is the *cause-specific hazard* function, the hazard of failing from a given cause k in the presence of the competing events (D):

$$h_k(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, D = k | T \geq t)}{\Delta t}.$$

This hazard is estimable from the data, as it is the *cumulative cause-specific hazard*:

$$H_k(t) = \int_0^t h_k(s)ds.$$

We can also define

$$S_k(t) = \exp\{-H_k(t)\},$$

thought it should not be interpreted as a *marginal survival function*, that is, $S^k(t) = P(T_k^* > t)$, which describes the event time distribution in the situation in which there were no competing risks. $S_k(t)$ and $S^k(t)$ only have the same interpretation when the event time distributions and the censoring distribution are independent.

Furthermore, we define

$$S(t) = \prod_{k=1}^{K} S_k(t) = \exp\left(-\sum_{k=1}^{K} H_k(t)\right).$$

This survival function is interpreted as the probability of not having failed from any cause at time t. From this definition we introduce the *cumulative incidence function* of cause k, $I_k(t)$, the probability $P(T \leq t, k)$ of failing from cause k before time t. It can be expressed in terms of the cause-specific hazard as

$$I_k(t) = P(T \leq t, k) = \int_0^t h_k(s)S(s)ds.$$

Note that this is not a proper probability distribution, because the cumulative probability to fail from cause k remains below one, $I_k(\infty) = P(k)$.

On the other hand, observe that, as the events from causes other than k are treated as censored, the naive Kaplan-Meier estimate of the probability of failing from cause k before or at time t is estimating

$$1 - S_k(t) = \int_0^t h_k(s)S_k(s)ds,$$

which is slightly different from the cumulative incidence function: in $I_k(t)$, $S_k(s)$ is replaced by $S(s)$. Since $S(t) \leq S_k(t)$, it is obvious that $I_k(t) \leq 1 - S_k(t)$, with equality at t if there were no competition (i.e. if $\sum_{j=1,j\neq k}^{K} H_j(t) = 0$), showing the bias in the naive Kaplan-Meier estimator that was mentioned before.

Both the cause-specific hazard and the cumulative incidence function are the most used functions for analyzing competing risks. The cumulative incidence function is also used extensively in calculating state and prediction probabilities in multi-state models,

but this will not be discussed here.

3.1.3. Estimation

The estimation of this concepts is based on the same principles as for survival analysis with a single failure type. Let $0 < t_1 < t_2 < ... < t_N$ be the ordered distinct time points at which failures of any cause occur. Let d_{ki} denote the number of individuals failing from cause k at t_i, and let $d_i = \sum_{k=1}^{K} d_{ki}$ denote the total number of failures (from any cause) at t_i. In the absence of ties only one of the d_{ki} equals 1 for a given i, and $d_i = 1$, though the formulas are also valid in the presence of ties. Let n_i be the number of individuals at risk (i.e. that are still in follow-up and have not failed from any cause) at time t_i. The survival probability $S(t)$ at t can be estimated, without considering the cause of failure, by the Kaplan-Meier estimator seen in section 2.2 with

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right).$$

As we have seen previously, we can consider a discretized version of the cause-specific hazard, $h_k(t_i) = P(T = t_i, k | T > t_{i-1})$, which would be estimated by

$$\hat{h}_k(t_i) = \frac{d_{ki}}{n_i},$$

the proportion of subjects at risk that fail from cause k. According to this, the previous expression can be written as

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \sum_{k=1}^{K} \hat{h}_k(t_i)\right).$$

The probability of failing from cause k at t_i, $p_k(t_i) = P(T = t_i, k)$, is the product of the hazard and the probability of being event-free at t_j, which is estimated as

$$\hat{p}_k(t_i) = \hat{h}_k(t_i)\hat{S}(t_{i-1}).$$

Finally, the cumulative incidence $I_k(t)$ of cause k is estimated as the sum of these terms for all time points before t:

$$\hat{I}_k(t) = \sum_{i:t_i \leq t} \hat{p}_k(t_i) = \sum_{i:t_j \leq t} \hat{h}_k(t_i)\hat{S}(t_{i-1}) = \sum_{i:t_i \leq t} \left\{ \frac{d_{ki}}{n_i} \prod_{j:t_j \leq t_j} \left(1 - \frac{d_j}{n_j}\right) \right\}.$$

If there were no censoring or left truncation, the estimate of the cumulative incidence function reduces to a very simple form: at time t, the estimate divedes the cumulative number of events of type k until time t by the total sample size:

$$\hat{I}_k(t) = \frac{\sum_{j:t_j \leq t} d_{kj}}{n}.$$

To illustrate this concepts, we use the peritoneal dialysis data that was introduced in the introduction: we recall that this data had two possible competing events: death/transfer to hemodialysis and renal transplantation. Figure 3.2 shows the estimates of the survival of transplant and the probabilities of death/hemodialysis of the data. The estimates based on the naive Kaplan Meier are in gray, an those based on the cumulative incidence function are in black. We can see the bias we talked about previously: the naive Kaplan-Meier overestimates the probability of failure in both competing risks.

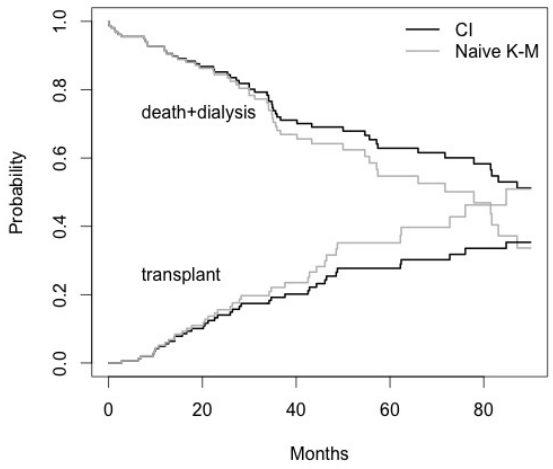

Figure 3.2: Estimates of probabilities of death or dialysis and transplant, based on the naive Kaplan-Meier (grey) and on cumulative incidence (CI) functions (black)

Besides, the naive Kaplan-Meier curves of death and dialysis and transplant cross after 80 months, which means that the estimated probabilities of both of those events sum to more than one, which is clearly impossible, since we are in a competing risk context and they are disjoint events.

Another way of representing this curves is shown at Figure 3.3: the bottom curve shows $\hat{I}_1(t)$ and the top curve $\hat{I}_1(t) + \hat{I}_2(t)$, where $\hat{I}_1(t)$ and $\hat{I}_2(t)$ are the estimates of

the cumulative incidence functions. The distances between adjacent curves correspond to the probabilities of the events.

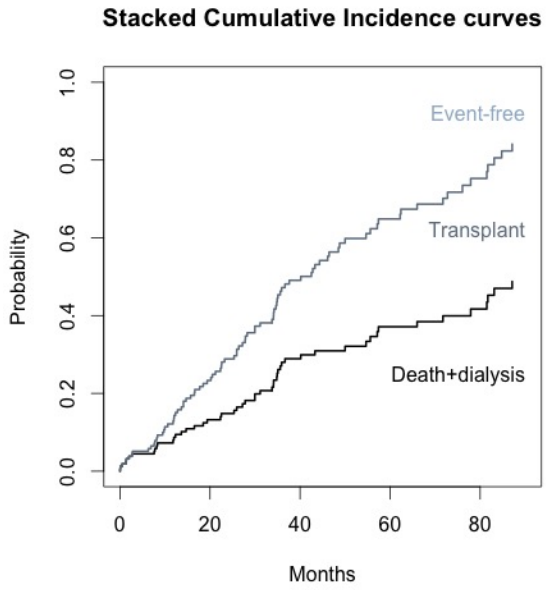

Figure 3.3: Stacked cumulative incidence curves of the two competing events of the peritoneal dialysis data: the bottom curve shows $\hat{I}_1(t)$ and the top curve $\hat{I}_1(t) + \hat{I}_2(t)$. The distances between adjacent curves correspond to the probabilities of the events.

3.2. Modelling and estimating covariate effects

Just like in standard survival analysis, the presence of covariates can affect the different outputs of the model created, so it is very important to add them to the analysis.

If the covariates under study are two binary covariates, there is a log-rank test developed for equality of cumulative incidence curves. Thus, the effect of those covariates is investigated by estimating cumulative incidence curves non-parametrically and testing whether the curves differ or not.

In a more general situation, Prentice and Kalbfleich (2002) proposed to use the classic Cox model to estimate cause-specific hazard functions, with the problem that the coefficients obtained this way do not have a direct interpretation in the cumulative incidence function.On the other hand, Fine and Gray (1999) proposed another model

based on the incidence cumulative function that tries to solve that problem. We will describe both approaches in the next two sections.

3.2.1. Regression on cause-specific hazards

If the covariate is continuous or the simultaneous effect of several covariates on cause-specific failure is of interest, a competing risks analogue of a Cox proportional hazards model is needed. With this model, each cause-specific hazard function is modeled separately, treating the competing risks observations as censored.

We model the cause-specific hazard of a cause k for a subject with covariate vector w_i as

$$h_{ik}(t|w_i) = h_{k,0}(t)\exp\{\gamma_k^T w_i\},$$

where $h_{k,0}(t)$ is the baseline cause-specific hazard of cause k, and the vector γ_k represents the covariate effects on cause k. At each time some person moves to state k, the covariate values of this individual are compared with the covariates of all other individuals still event-free and in follow-up. The $\exp(\gamma_k)$ is called the *cause-specific hazard ratio* for the k event, and it represents the relative risk of failing from that event when the correspondent variable increases one unit its value.

The covariate effects in that model are proportional for the cause-specific hazards, as in the traditional Cox model. In the absence of competing risks this would mean that the survival functions for different values of the covariates were related through a simple formula. If S_1 and S_2 were the survival functions for the covariates w_1 and w_2, then

$$S_2(t) = S_1(t)^{\exp\{\gamma^T(w_2-w_1)\}}.$$

However, in the presence of competing risks, when the effect of the same covariates are also modelled for other causes of failure, this relation does not extend to cumulative incidence functions.

The reason is that the cumulative incidence function for cause k not only depends on the hazard of cause k, but also on the hazards of all other causes. Recall

$$I_k(t) = \int_0^t h_k(s)S(s)ds = \int_0^t h_k(s)\exp\left(-\sum_{k=1}^{K}\left(\int_0^s h_k(r)dr\right)\right)ds.$$

Hence, the relation of the cumulative incidence functions of cause k for two different covariate values not only depends on the effect of the covariate on cause k, but also on the effects of the covariate of all other causes and on the baseline hazards of all

other causes. As a result, the simple effect of a covariate on the cause-specific hazard of cause k can be quite unpredictable when expressed in terms of the cumulative incidence function.

Returning to our example, in Figure 3.4 we can see the cumulative incidence functions estimated for the peritoneal dialysis data (where the main event is the death or the transfer to hemodialysis and the competing risk is the patient receiving a transplant) for both sexes. This estimation is based on the cause-specific hazards.

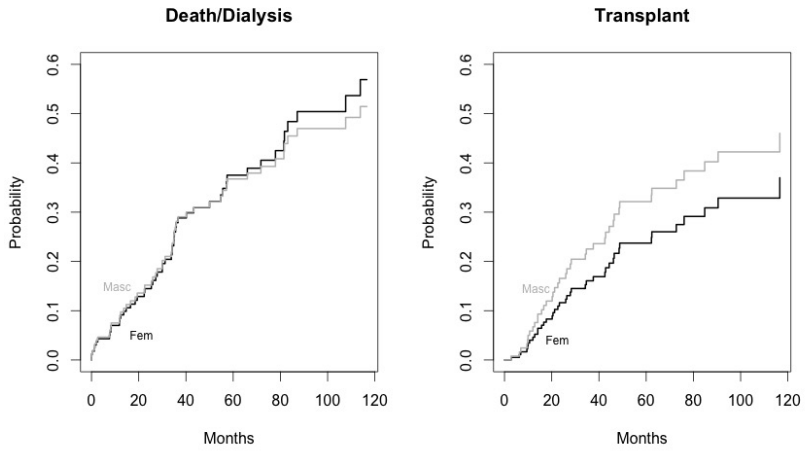

Figure 3.4: Cumulative incidence functions for Death/Transfer and Transplantation for both sexes, based on a proportional hazards model on the cause-specific hazards.

3.2.2. Regression on cumulative incidence functions

Seen the limitations of this previous model, Fine and Gray (1999) introduced a way to regress directly on cumulative incidence functions. In analogy with the relation $h(t) = -\frac{d \log S(t)}{dt}$ seen in the chapter 2 between hazard and survival, they defined a *subdistribution hazard*:

$$\overline{h}_k(t) = -\frac{d \log(1 - I_k(t))}{dt}.$$

At the moment of estimating this quantity, the difference between that and the cause-specific hazard is in the risk set: for the cause-specific hazard, the risk set decreases at each time point at which there is a failure of another cause; for $\overline{h}_k(t)$, the individuals who fail from another cause remain in the risk set.

Fine and Gray (1999) imposed a proportional hazards assumption on the subdistribution hazards:

$$\overline{h}_{ik}(t|w_i) = \overline{h}_{k,0}(t)\exp\{\gamma_k^t w_i\}.$$

An example (once again, with the peritoneal dialysis data) of the cumulative incidence functions estimated this way is in Figure 3.5. They are similar to the previous cumulative incidence functions estimated in the previous section, but here we can see that the effect of the covariate *Sex* is proportional in the cumulative incidence: the separate curves do not cross as they did before.

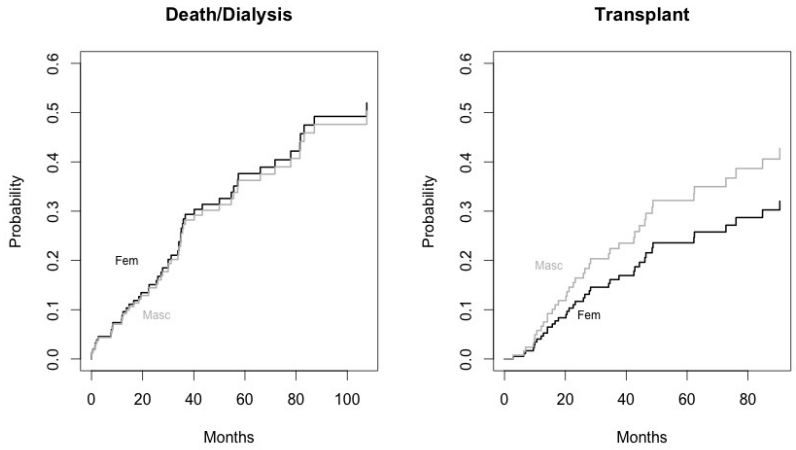

Figure 3.5: Cumulative incidence functions for Death/Transfer and Transplantation for both sexes, based on the Fine and Gray method.

The Fine and Gray method is a way of repairing problems with proportional hazards regression on cause-specific hazards, but there is nothing wrong with that regression. The problem lies in the fact that we are used to interpreting hazard ratios in the standard proportional hazards regression with a single endpoint, implying a similar cumulative effect.

A way of judging the goodnes-of-fit of the two approaches is by comparing the predicted cumulative incidence curves of the regression models with the non-parametric cumulative incidence curves obtained by applying

$$\hat{I}_k(t) = \sum_{j:t_j \le t} \left\{ \frac{d_{kj}}{n_j} \prod_{i:t_i \le t_j} \left(1 - \frac{d_i}{n_i}\right) \right\}.$$

to the subsets of covariates considerated separately. Figure 3.6 shows these model-free

cumulative incidence curves.

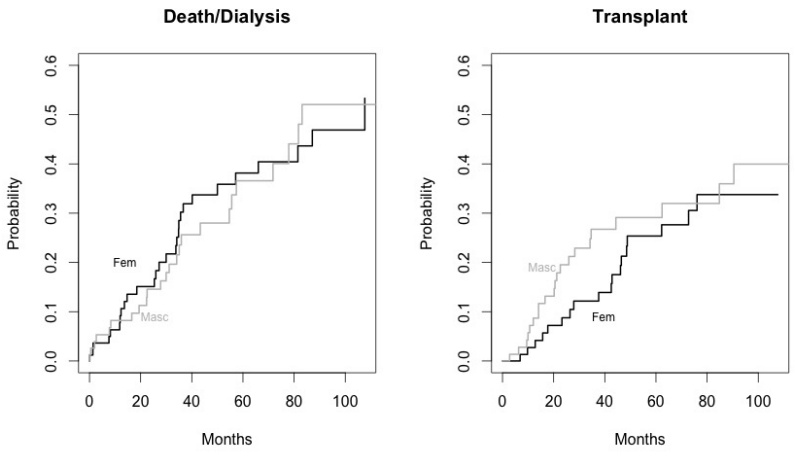

Figure 3.6: Non parametric cumulative incidence functions for Death/Transfer and Transplantation for both sexes.

In summary, modelling the effect of covariates on cause-specific hazards may lead to different conclusions than modelling their effect on subdistribution hazards and cumulative incidence functions.

The standard Cox model can be used to model the effect of covariates on the cause-specific hazards of the different endpoints, and has the advantage that there is a wealth of theory and software that has been developed for this purpose. The problem is that proportionality is lost, and hence covariate effects on cumulative incidence curves can no longer be expressed by a simple number, as it can be done with the regression on cumulative incidence curves.

Chapter 4

Joint Modelling

Once we have explained the basics of the survival, the longitudinal analysis, and the competing risk, we are in position to build a model that takes into account all three of those blocks.

We start by introducing the joint modelling approach that studies the association between the survival and longitudinal process, without considering competing risks. In the first section we explain its importance and its advantages over the extended Cox model (Andersen and Gill, 1982). In section 4.2 we specify the longitudinal and survival submodels of the model, followed by the estimation of the model's parameters. Next, in section 4.4, inference to the regression coefficients is discussed.

Finally, in section 4.5 we will focus on the inclusion of the competing risks into the basic joint model.

4.1. The Basic Joint Model

As mentioned in section 2.2.4, the extended Cox model can study the association between longitudinal measurements and the survival process, but it has its limitations. Those drawbacks can be expressed by the example showed in Figure 4.1.

In the top panel of that Figure the solid red line illustrates how the hazard function evolves in time, i.e., how the instantaneous risk of an event changes in time. On the other hand, in the bottom panel the asterisk denote the observed longitudinal responses. The green line represents the underlying longitudinal process.

The joint models approach postulates a relative risk model for the event time outcome directly associated with the longitudinal process. That process is approximated using a mixed effects model and the observed data (asterisks). That model contains

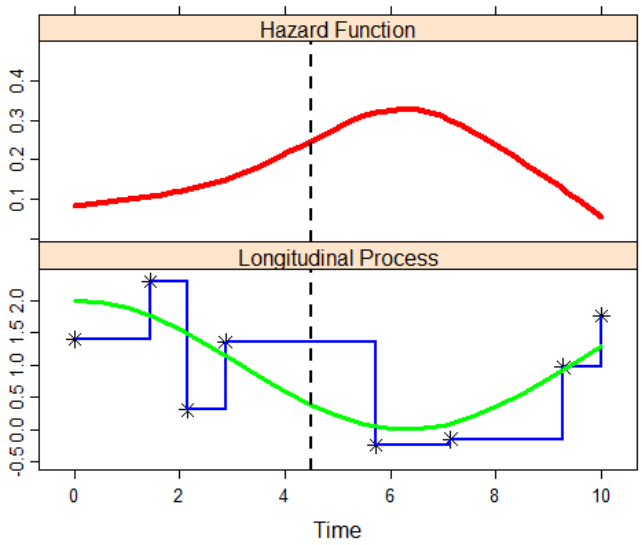

Figure 4.1: Intuitive idea of joint models. In the top panel the solid red line represents the hazard function. In the bottom panel the blue line corresponds to the extended Cox approximation of the longitudinal trajectory, meanwhile the green curve illustrates the underlying longitudinal process.

fixed effects, describing the average longitudinal evolution in time, and random effects that describe how each patient deviates from this average evolution.

In their basic form joint models assume that the hazard function at any particular time point t (in Figure 4.1, the dashed line) is associated with the value of the longitudinal process (green line) at the same point. As for the blue line, it represents the assumption behind the time-dependent Cox model, which postulates that the value of the longitudinal outcome remain constant in between the observation times. Hence, the blue line is staggered.

Through this example we see that using the extended Cox model we would be introducing some error in the estimation of the longitudinal variable included in the model. This is why a joint model approach is preferable, as it can be used to account for both exogenous and endogenous time-depending covariates.

Though they are different proposals of joint approaches, here we will introduce the one proposed by Rizopoulos (2012), where the main goal is to study the subjects'survival. With this in mind, we will firstly specify the two submodels that compose the joint modelling. Then we will discuss the maximum likelihood estimation of the model's parameters, following with the estimation of the random effects and ending

with a brief summary of inference for those parameters.

4.2. Submodels specification

The joint model is composed of two linked submodels: the longitudinal and the survival submodel. The notation here will be similar to the one used in previous chapters; let T_i^* denote the true event time for the ith subject, T_i the observed event time (defined as the minimum of the potential censoring time, C_i, and T_i^*) and $\delta_i = I(T_i^* \leq C_i)$ the event indicator.

Besides, let $y_i(t)$ be the observed value of the time-dependent covariate at time point t, and equivalently, $y_{ij} = \{y_i(t_{ij}), j = 1, ..., n_i\}$. Thus, $m_i(t)$ will denote the true and unobserved value of the respective longitudinal outcome at time t, uncontaminated with the measurement error value of the longitudinal outcome (and, because of this, different from $y_i(t)$).

4.2.1. The Survival Submodel

Our aim is to associate the true an unobserved value of the longitudinal outcome at time t, $m_i(t)$, with the risk for an event. As stated in section 2.2.4, the relative risk model can be written as

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

where $\mathcal{M}_i(t) = \{m_i(t), 0 \leq s < t\}$ denotes the history of the true (unobserved) longitudinal process up to time t. Let $h_0(t)$ denote, as before, the baseline risk function, and w_i the vector of baseline covariates. The interpretation of the regression coefficients is the same as in previous models:

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

Note that this expression depends only on a single value of the time-dependent marker $m_i(t)$. However, this does not hold for the survival function. To take into account the whole covariate history $\mathcal{M}_i(t)$ to determine the survival function we use the relation

$S(t) = \exp\left\{-\int_0^t h(s)ds\right\}$ to obtain

$$\begin{aligned} S_i(t|\mathcal{M}_i(t), w_i) &= P(T_i^* > t|\mathcal{M}_i(t), w_i) \\ &= \exp\left\{-\int_0^t h_0(t)\exp(\gamma^T w_i + \alpha m_i(t))ds\right\}. \end{aligned}$$

We keep in mind that both the survival and the hazard functions are written as functions of a baseline hazard $h_0(t)$. Regardless of the fact that the literature recommends to leave $h_0(t)$ completely unspecified, in order to avoid the impact of misspecifying the distribution of survival times, in the joint modelling framework it can lead to an underestimation of the standard error of the parameter estimates (Hsieh et al., 2006). Thus, we will need to explicitly define $h_0(\cdot)$.

One option is to assume that the risk function corresponds to a known parametric distribution, such as the Weibull, the log-normal or the Gamma. For example, the *Weibull model* assumes that the hazard takes the form

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

where, if $p > 1$ the failure rate increases with time, if $p < 1$ it decreases and remains constant over time if $p = 1$ (also called the *exponential model*).

But it is more desirable to have a more flexible model for the baseline risk function. Among the proposals encountered, we highlight this next two options:

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

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

  where $0 = v_0 < v_1 < ... < v_Q$ denotes a partition of the time scale, with v_Q being larger than the largest observed time, and ξ_q denoting the value of the hazard in the interval $(v_{q-1}, v_q]$.

- The *regression splines model*, where the log baseline risk function $\log h_0(t)$ is given by

$$\log h_0(t) = k_0 + \sum_{d=1}^{m} k_d B_d(t, q),$$

  where $k^t = (k_0, k_1, ..., k_m)$ are the spline coefficients, q denotes the degree of the B-splines basis functions $B(\cdot)$ (de Boor, 1978), and $m = \ddot{m} + q - 1$, with $\ddot{m}$ the

number of interior knots. This is the option that we will be using when applying the joint modelling in the real data in the next chapter.

In both models, the specification of the baseline hazard becomes more flexible as the number of knots increases. In particular, in the limiting case of the piecewise-constant model where each interval contains only a single true event time, this model is equivalent to leaving h_0 unspecified and estimating it using nonparametric maximum likelihood. In both approaches, we should keep a balance between bias and variance and avoid overfitting. Although there is not an ideal strategy, Harrel (2001) gives a standard rule of thumb based on keeping the total number of parameters between 1/10 and 1/20 of the number of events in the sample. After the number of knots has been decided, their location is usually based on percentiles of the observed event times T_i.

4.2.2. The Longitudinal Submodel

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

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

where we notice that the design vectors $x_i(t)$ for the fixed effects β and the $z_i(t)$ for the random effects b_i, as well as the error terms $\varepsilon_i(t)$, are time-dependent. Similarly to section 2.1, we assume that error terms are mutually independent, independent of the random effects and normally distributed with mean zero and variance σ^2.

This mixed model formulation allows to settle that the longitudinal outcome $y_i(t)$ is equal to the true level $m_i(t)$ plus an error term. The main difference from the model in section 2.1 is that, in addition to the random error term $\varepsilon_i(t)$ we incorporate an

additional stochastic term $u_i(t)$. This last term is used to capture the remaining serial correlation in the observed measurements, which random effects are unable to capture. Besides, $u_i(t)$ is considered as a mean-zero stochastic process, independent of b_i and $\varepsilon_i(t)$.

4.3. Estimation

In chapter 2 the estimation of the parameters was based on the maximum likelihood approach for both longitudinal and survival processes. Rizopoulos (2012) has also used the likelihood method for joint models, as it is the most commonly used approach in the joint literature. Though the two-stages approach for the parameters estimation is less complex than those methods in a computationally aspect, the approximations applied with this second approach produces bias.

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

4.3.1. Joint Likelihood Formulation

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

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

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

where $p(\cdot)$ denotes the corresponding probability density function, and

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

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

Under the modelling assumptions presented in previous sections and these above conditional independence assumptions, the joint log-likelihood contribution for the $i-th$

subject has the form

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

where the likelihood of the survival part takes the form

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

with $h_i(\cdot)$ and $S_i(\cdot)$ the ones described in section 2.2. On the other hand, the joint density for longitudinal responses together with the random effects is performed through the following expression,

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

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

Then,the maximization of the log-likelihood with respect to θ for all the observed data , formulated as,

$$l(\theta) = \sum_i \log p(T_i, \delta_i, y_i; \theta),$$

requires a combination of numerical integration and optimization algorithms. Due to the fact that both the integral with respect to the random effects and in the survival function do not have an analytical solution, a numerical integration technique is needed.

Despite some authors have employed standard numerical integration techniques, such as Monte Carlo or Gaussian quadrature, the Expectation-Maximization (EM) algorithm described by Wulfsohn and Tsiatis (1997) has been traditionally preferred. The intuitive idea behind the EM algorithm is to maximize the log-likelihood in two steps: the *Expectation step*, where missing data are filled, so we replace the log-likelihood of the observed data with a surrogate function, and the *Maximization step*, where this surrogate function is then maximized.

Besides these techniques, Rizopoulos et al. (2009) have introduced a direct maximi-

zation of the observed data log-likelihood, which is a quasi Newton algorithm. Therefore hybrid optimization approaches start with EM and then continue with direct maximization.

4.3.2. Estimation of the Random Effects

Until now we have focus our attention on the estimation of the parameters β, γ and α, but in many settings interest may lie in deriving patient-specific predictions for their outcomes. To derive such predictions, an estimate of the random effects vector b_i is required. Since the random effects are assumed to be random variables, it is natural to estimate them using the Bayesian theory.

This is what Rizopoulos (2012) does when estimating the random effects. Assuming that $p(b_i;\theta)$ is the prior distribution, and that $p(T_i,\delta_i|b_i;\theta)p(y_i|b_i;\theta)$ is the conditional likelihood part, the corresponding posterior distribution is,

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

which does not have a closed form solution so it has to be numerically computed. However, as the number of longitudinal measurements n_i increases, this distribution will converge to a normal distribution.

To describe this posterior distribution, standard summary measures (such as the mean and the mode) are often utilized. Thus, two types of estimators typically used are:

$$\begin{cases} \bar{b}_i = \int b_i p(b_i|T_i,\delta_i,y_i;\theta)db_i, \text{ and} \\ \\ \hat{b}_i = \arg \text{máx}_b\{\log p(b_i|T_i,\delta_i,y_i;\theta), \end{cases}$$

and they correspond, respectively, with the mean and the mode.

4.4. Model testing

It has been shown in previous sections that the joint models' parameters can be estimated by maximum likelihood. The next step would be to do some inference tests.

In general, if we are interested in testing the null hypothesis

$$\begin{aligned} H_0 : &\quad \theta = \theta_0, \\ H_1 : &\quad \theta \neq \theta_0, \end{aligned}$$

there are different methods we could use:

- the *Likelihood Ratio Test*, with the test statistic defined as

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

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

- the *Score Test*, with the test statistic defined as:

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

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

- or the *Wald Test*, with the test statistic defined as

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

Under the null hypothesis, the asymptotic distribution of each of these tests is a chi-squared distribution on p degrees of freedom, with p denoting the number of parameters being tested. In particular, the Wald test for a single parameter θ_j is equivalent to $(\hat{\theta}_j - \theta_{0j})/s.e.(\hat{\theta}_j)$, which under the null hypothesis follows an asymptotic standard normal distribution.

Despite of being asymptotically equivalent, the behavior of the tests is different in finite samples. The election of any of these procedures depends on the limitations of each one. Specifically, regarding the computational cost of fitting, the Wald test only requires to fit the model under the null hypothesis, and the score test under the alternative. However, the likelihood ratio test requires to fit the model under both hypotheses, being more computationally expensive. But other issues must be considered,

such as the fact that the Wald test does not take into account the variability introduced by estimating the variance components, apart from ignoring that we need to estimate the survival process. Also, the implementation of the score test needs extra steps to calculate the required components.

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

$$\begin{aligned} AIC &= -2l(\hat{\theta}) + 2p, \\ BIC &= -2l(\hat{\theta}) + p\log(n), \end{aligned}$$

where p denotes the number of parameters in the model.

Apart from these topic procedures to models'comparison, we could also be interested in testing whether an extra random effect should be included in the joint model or not. However, this specific field is forgotten in the joint modelling framework, so it could be an interesting future line of research.

4.5. Extension of the standard joint model: competing risks

Joint modelling of longitudinal and survival data has received much attention in the last years and is becoming increasingly used in clinical studies. Although some joint models were adapted in order to allow for competing endpoints, this methodology has not been widely disseminated.

Despite the fact that there are well-established models that allow to analyze longitudinal and time-to-event outcomes separately, this models are not suitable to analyze data when the longitudinal outcome and survival competing endpoints are associated. In those cases, a joint modelling approach is required.

In this section, we will focus on studying the association between a single endogenous time-dependent covariate and time to different types of failure, for it could be of interest to distinguish between the events and investigate how covariates affect the risk for each one of them. One of the traditional types of analysis in such settings is the cause-specific hazard regression, which postulates separate relative risk models for each of the competing events.

To handle different failure types we need to extend the notation for the survival process. Assuming K different causes of failure, we let $T_{i1}^*, ..., T_{iK}^*$ denote the true failure times for each one of those. The observed data for de ith subject comprise of the observed event time $T_i = \text{mín}(T_{i1}^*, ..., T_{iK}^*, C_i)$, with C_i denoting the censoring time. The event indicator takes values $\delta_i \in \{0, 1, ..., K\}$, with 0 corresponding to censoring, and 1,...,K to the competing events. For each of the K causes, and as mentioned above, we postulate the standard relative risk model

$$h_{ik}(t|\mathcal{M}_i(t)) = h_{0k}(t) \exp\{\gamma_k^T w_i + \alpha_k m_i(t)\},$$

which includes the effects of the baseline covariates w_i and the effects of the current value of the longitudinal marker $m_i(t)$, and where $\mathcal{M}_i(t) = \{m_i(t), 0 \leq s < t\}$ as before.

The specification of the joint model is completed by positing a suitable mixed-effects model for the observed longitudinal responses $y_i(t)$, as the one detailed in the previous section:

$$y_i(t) = m_i(t) + \varepsilon_i(t) = x_i^T(t)\beta + z_i^T(t)b_i + \varepsilon_i(t),\ b_i \sim N(0, D) \text{ and } \varepsilon_i(t) \sim N(0, \sigma^2 I_{n_i}).$$

Estimation of such joint models is based on exactly the same principles as for joint models with a single failure type. The only difference is in the construction of the likelihood part for the event process. In particular, it takes the form:

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

On the other hand, for the estimation of the baseline risk function $h_{0k}(s)$ it is required the use of the regression spline method. For each event k, the log baseline risk function $\log h_0(t)$ is expanded into B-spline basis functions as follows:

$$\log h_0(t) = k_0 + \sum_{d=1}^{m} k_d B_d(t, q),$$

where, as in section 4.1, $k^t = (k_0, k_1, ..., k_m)$ are the spline coefficients, q denotes the degree of the B-splines basis functions $B(\cdot)$, and $m = \ddot{m} + q - 1$, with $\ddot{m}$ the number of interior knots.

Chapter 5

Application to real data

Throughout this past chapters we have seen how to approach a situation where we have interest in the survival of some data with longitudinal measurements and in the presence of competing risks. This chapter presents an application for joint modelling and competing risk process to a real data, that will illustrate the potential benefits of using this techniques.

The data includes patients in the peritoneal dialysis program from the Peritoneal Dialysis Unit, Nephrology Department, Hospital Geral de Santo António - Centro Hospitalar de Porto, Portugal. As explained in the introduction, along the permanence in the peritoneal dialysis program, this data presents different types of information about each patient: baseline characteristics taken at the beginning of the study (like the sex or the age), several clinical parameters measured over time (albumin, calcium and phosphorus score) and the event that forced the patient to abandon the treatment program (Figure 5.1).

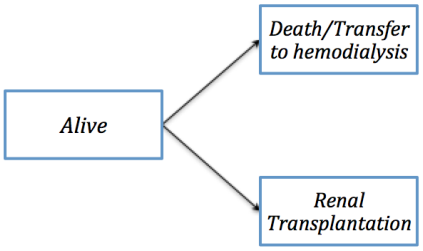

Figure 5.1: Diagram for the competing events of the peritoneal dialysis data.

From the biomedical perspective, low albumin level is usually associated with kidney failure, while calcium levels of the blood tend to drop. As for the phosphorus, it stays

in the body when the kidneys can no longer remove it.

The main objective of the application study is to compare different regression models for the patient behaviors in the peritoneal dialysis program. In order to to so, we first describe the variables contained in the database, then we propose several models to analyze in which ways the survival of the subjects is affected by the available covariates and finally we compare all those models using time dependent AUCs. At the end of the chapter, along with the results obtained we include a brief discussion on the available software used in the analysis.

5.1. Peritoneal Dialysis Data

The data includes the information of 160 patients who started peritoneal dialysis therapy between October 1999 and February 2013.

The different outcomes observed were death (9.37 %), transfer to hemodialysis (21.75 %) and renal transplantation (25 %), though in order to do the competing risks analysis we will distinguish between the subjects that died or were transfered to hemodialysis (the main event we want to study) and the ones that received a renal transplantation (considered as competing risk). The rest of the subjects (41.88 %) were treated as censored. The median of follow-up time was 26.8 months.

The baseline variables that were considered were:

- *Sex*: represents the gender of the patients (51.87 % of women).
- *Age*: age in years of the patient on the moment they started the therapy (mean age 47.86 years, standard deviation s.d.=14.4 years).

As for the clinical parameters that were measured over time, for a total of 3169 observations, (usually recorded once per month), they were

- *Album*: gather the albumin score in g/dL (mean 3.7 g/dL, s.d.=0.4 g/DL).
- *Calc*: calcium score (mean 2.2, s.d.=0.24).
- *Phos*: phosphorus level (mean 1.6, s.d.=0.42).

The number of measures of this parameter varied between patients, with a minimum of 1 observations and a maximum of 60, being 15.3 de median of observations. That means that we are dealing with unbalanced data.

In Figure 5.2 we can see several graphics (barplots and boxplots) of the baseline measures for all the patients, meaning the age, the sex and the first measurements for albumin, calcium and phosphorus separated by the kind of failure they had at the end of the study.

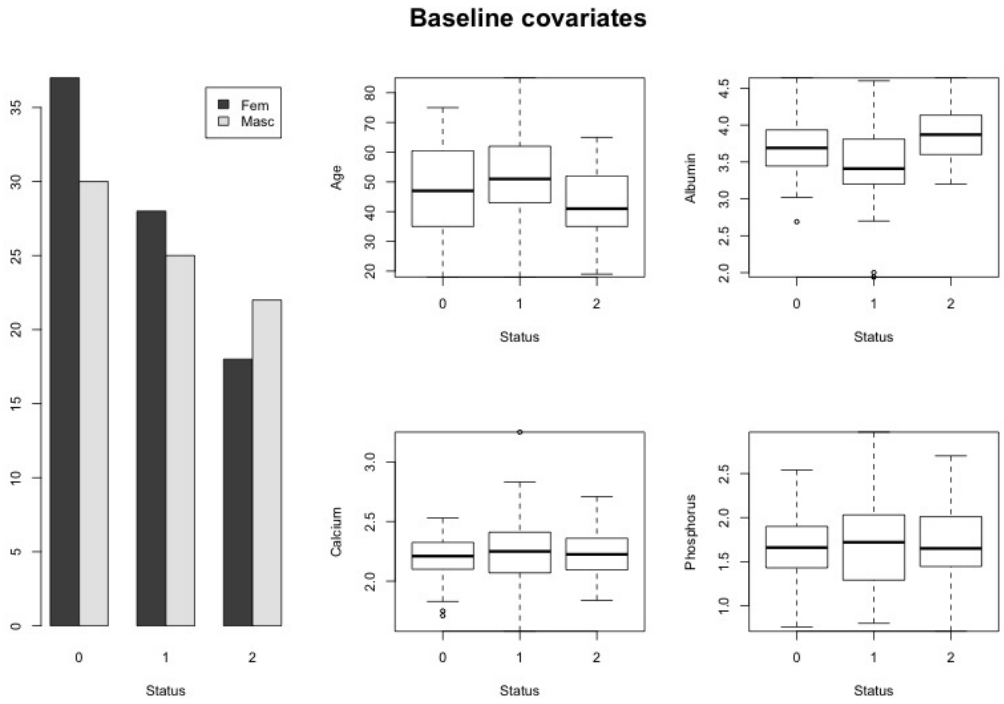

Figure 5.2: Baseline covariates classified by the status: 0 for censored, 1 for death or transfer to hemodialysis and 2 for transplant.

At first sight, we observe differences when talking about the age and the albumin score: the patients who tend to have more possibilities of getting a transplant are the youngest, and the ones with higher albumin levels. As for the sex of the patients, it seems that females tend to experience death, transfer or stay alive more frequently than the males, whom are a majority only in the case of liver transplantation.

As we can observe in Figure 5.2, the subjects who experienced death or transfer to hemodialysis have a lower albumin median, but to be sure of how this covariates affect the different states we will need the statistic tools we have been developing in this project.

5.2. The Statistical Models

In this section several models are presented to analyze the data. In the first place, several linear mixed-effects models are applied to the longitudinal covariates albumin, calcium and phosphorus, leaving aside the survival point of view.

The next models will do the opposite: they are competing risks models where the different kind of failures are taken into account,only including the baseline measurements of the longitudinal process which are taken at the begging of the study.

Finally, the last models use both techniques, applying a joint model with competing risks. The only disadvantage of using this kind of approach is that we are not yet able to include more than one longitudinal covariate in the analysis, so we will have to fit one model for each one of these variables.

5.2.1. Linear Mixed-Effects Models

As mentioned before, in this section we will perform a mixed model analysis to describe the evolution in time of the albumin, calcium and phosphorus score. In Figure 5.3 we can see the scores for those variables of the subjects over time, with the overall trajectories adjusted with a p-spline method.

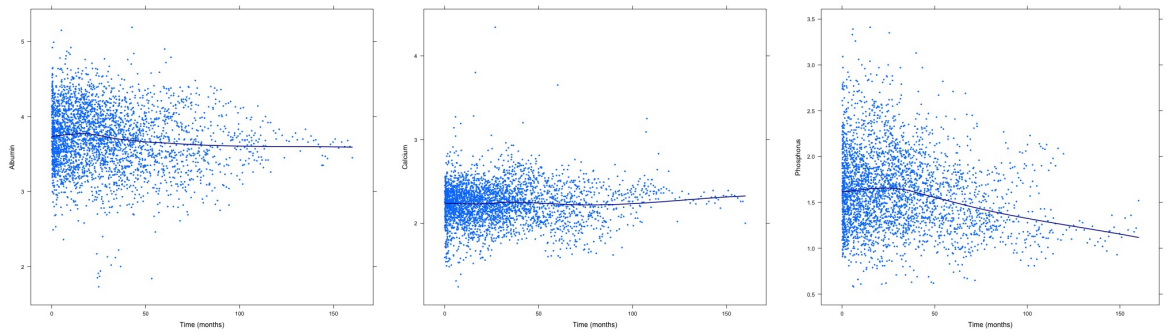

Figure 5.3: Longitudinal scores showing the individual progression of the longitudinal variables

The linear mixed-effect models that we will be applying to these covariates is the

same we expose in chapter 2:

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

Of course, there will be a separate model for each longitudinal variable, but we will consider the same fixed and random effects for all of them. As we expect an evolution in time different for each patient and with different average effects per sex and age, the models that we will consider are:

- $y_{A,i} = \beta_{A,0} + \beta_{A,1} Sex_i + \beta_{A,2} Age_i + \beta_{A,3} t_i + b_{A,0i} + b_{A,1i} t_i + \varepsilon_{A,i}$,
- $y_{C,i} = \beta_{C,0} + \beta_{C,1} Sex_i + \beta_{C,2} Age_i + \beta_{C,3} t_i + b_{C,0i} + b_{C,1i} t_i + \varepsilon_{C,i}$,
- $y_{F,i} = \beta_{F,0} + \beta_{F,1} Sex_i + \beta_{F,2} Age_i + \beta_{F,3} t_i + b_{F,0i} + b_{F,1i} t_i + \varepsilon_{F,i}$,

with the assumption that both the random effects and the error terms come from a normal distribution. That is, we are considering the time, the sex and the age as fixed effects, as well as a random-intercepts and random-slopes model, assuming that the rate of change in those longitudinal variables is different from patient to patient.

In Table 5.1 we synthesize the estimated coefficients for the fixed effects of these models.

These results suggest that the longitudinal scores remains constant along the time. Additionally, age and sex were identified as statistically significant predictors of albumin and phosphorus: in both cases the estimated coefficients indicate that male patients present higher average level of albumin and phosphorus, and that older patients are expected to have lower average levels of those variables. As for the calcium, sex and age can also be considered as significant predictors for its score, but in this case they are the males and the youngest who are expected to have lower calcium levels. In Figure 5.4 we represent again the different scores over time, but this time distinguishing the males (pink) and the females (blue). Note that the overall trajectories are presented with a p-spline method in Figure 5.3 seems to indicate that phosphorus scores tend to decrease over time, whereas the coefficients we have just estimated indicate quite the opposite notion: this is because with the p-spline method we do not take into account the individual trajectories, so it may lead to misinterpretations.

		Coef	s.d.	p-value
Albumin	$\beta_{A,0}$ (Intercept)	3.8848	0.1069	0.0000
	$\beta_{A,1}$	0.2404	0.0619	0.0002
	$\beta_{A,2}$	-0.0053	0.0021	0.0142
	$\beta_{A,3}$	-0.0013	0.0011	0.2703
Calcium	$\beta_{C,0}$ (Intercept)	2.1674	0.0573	0.0000
	$\beta_{C,1}$	-0.0823	0.0324	0.0122
	$\beta_{C,2}$	0.0023	0.0011	0.408
	$\beta_{C,3}$	0.0002	0.0007	0.7464
Phosphorus	$\beta_{F,0}$ (Intercept)	2.1013	0.0887	0.0000
	$\beta_{F,1}$	0.1123	0.0513	0.0301
	$\beta_{F,2}$	-0.0104	0.0017	0.0000
	$\beta_{F,3}$	0.0027	0.0012	0.0248

Table 5.1: Fitted values for the linear mixed-effects models for the different longitudinal variables, with their standard deviations (se) and the p-values.

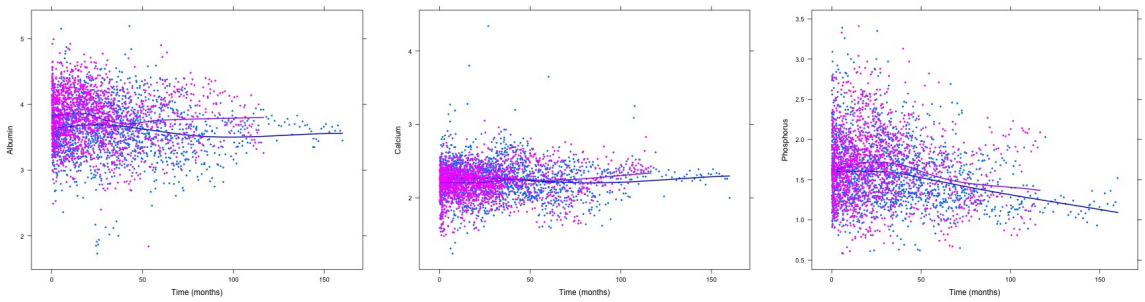

Figure 5.4: Longitudinal scores showing the individual progression of the longitudinal variables by sex (blue for female and pink for male).

These same models will be considered for the longitudinal submodels when applying the joint modelling approach.

5.2.2. Competing Risks

Following the procedures explained in section 3.1.3 we can estimate the cumulative incidence curves for the two competing events, taking into account only the failure times and the cause of failure of the data. Graphically represented in Figure 5.5, those cumulative incidence functions give us an insight of how those events evolve over time.

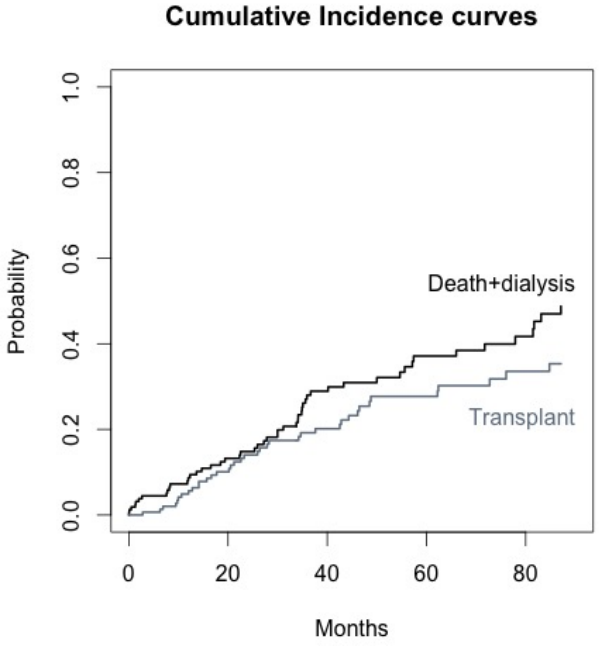

Figure 5.5: Cumulative incidence curves of the two events of failure.

The first model that we present in the context of competing risks model is a complete model which includes all the baseline measurements of the longitudinal markers (albumin, phosphorus,, calcium) and competing risks (death/transfer to hemodialysis and renal transplantation) with baseline covariates (age, gender).

Two cause-specific relative risks models are assumed:

$$\begin{cases} h_{i1}(t) = h_{01}(t)\exp\{\gamma_{11}Sex_i + \gamma_{12}Age_i + \gamma_{13}Album_i + \gamma_{14}Calc_i + \gamma_{15}Phos_i\}, \\ h_{i2}(t) = h_{02}(t)\exp\{(\gamma_{11}+\gamma_{21})Sex_i + (\gamma_{12}+\gamma_{22})Age_i + (\gamma_{13}+\gamma_{23})Album_i \\ \qquad +(\gamma_{14}+\gamma_{24})Calc_i + (\gamma_{15}+\gamma_{25})Phos_i\}. \end{cases}$$

The parameters γ_{11}, γ_{12}, γ_{13}, γ_{14} and γ_{15} denote the effects on sex, age, albumin, calcium and phosphorus, respectively, on the risk for death/transfer to hemodialysis, and the parameters γ_{21}, γ_{22}, γ_{23}, γ_{24} and γ_{25} denote the additional effects of sex, age, albumin, calcium and phosphorus on the risk for renal transplantation. As this model is quite complex, as it includes a lot of parameters, we apply a variables selection method. The model that we obtain is the following:

$$\begin{cases} h_{i1}(t) = h_{01}(t)\exp\{\gamma_{11}Age_i + \gamma_{12}Album_i + \gamma_3 Phos_i, \} \\ \\ h_{i2}(t) = h_{02}(t)\exp\{(\gamma_{11}+\gamma_{21})Age_i + (\gamma_{12}+\gamma_{22})Album_i + \gamma_3 Phos_i\}. \end{cases}$$

Note that the sex and the calcium have been excluded completely from this second model, and that phosphorus is not assumed to have any additional effect on the risk for renal transplantation. The parameters estimates and their standard errors are presented in Table 5.2.

	Coef	Exp(Coef)	Std. Error	p-value
Event of interest (D/HD)				
γ_{11}(Age)	-0.0047	0.9952	0.0097	0.6263
γ_{12}(Album)	-0.6811	0.5060	0.2948	0.0208
γ_3 $(Phos)$	0.51911	1.6805	0.2376	0.0289
Competing risk (T)				
γ_{21} $(Age:CR)$	-0.0306	0.9698	0.0152	0.0450
γ_{22} $(Album:CR)$	2.0260	7.5844	0.5276	0.0001

Table 5.2: Fitted values for the competing risk model.

Considering those results, significantly γ_{12} estimate indicates that individuals who have a lower albumin level tend to have a worse survival, meaning that a unit decrease in albumin score corresponds to $\exp(-(-0.68)) = 1.98$ increase in the risk for death/transfer to hemodialysis. As for the phosphorus, it is just the opposite: a unit increase in phosphorus score correspond to a 1.68 increase in the risk for death/transfer to hemodilaysis. There was also found an association between albumin and age and the risk of renal transplantation: the younger patients have a statistically significant higher

risk of getting a renal transplant, as well as the patients with a more elevated albumin score.

5.2.3. Joint Modelling & Competing Risks

With the purpose of evaluating the relationship between the longitudinal scores and death/transfer to hemodialysis in the presence of the competing risk renal transplantation, three different joint models were analyzed, each of one including a different longitudinal covariate. In Figure 5.6 we can observe once again the longitudinal scores of those covariates, but this time separated by the kind event that occurs. This approach is recommended when the focus of the research is on the survival outcome, and it allows to evaluate the impact of a longitudinal covariate (one at a time).

We will now present the three models, each of one considering one of the longitudinal variables. All of them include a longitudinal and a survival submodel. The first submodel is similar at the one explained in the section 5.2.1 of this chapter: it is a linear mixed-effect model were the fixed effects included are the sex, the age and the time, and we consider the random intercept and random slope effects.

As for the survival submodel, it is a competing risk model where the essential difference with a extended Cox Model where we could handle time-dependent covariates is that we are considering $m_i(t)$, the true and unobserved value of the longitudinal outcome $y_i(t)$.

Next, we expose the three models that we will be considering. The basic difference between each one of them is the true value of the longitudinal outcome, that is $m_{A,i}(t)$ for albumin, $m_{C,i}(t)$ for calcium and $m_{F,i}(t)$ for phosphorus.

Albumin

- Longitudinal Submodel:

$$y_{A,i}(t) = \beta_{A,0} + \beta_{A,1}Sex_i + \beta_{A,2}Age_i + \beta_{A,3}t + b_{A,0i} + b_{A,1i}t + \varepsilon_{A,i}(t).$$

- Survival Submodel:

$$\begin{cases} h_{A,i1}(t) = h_{A,01}(t)\exp\{\gamma_{A,11}Sex_i + \gamma_{A,12}Age_i + \alpha_{A,1}m_{A,i}(t)\}, \\ h_{A,i2}(t) = h_{A,02}(t)\exp\{(\gamma_{A,11} + \gamma_{A,21})Sex_i \\ \qquad\qquad +(\gamma_{A,12} + \gamma_{A,22})Age_i + (\alpha_{A,1} + \alpha_{A,2})m_{A,i}(t)\}. \end{cases}$$

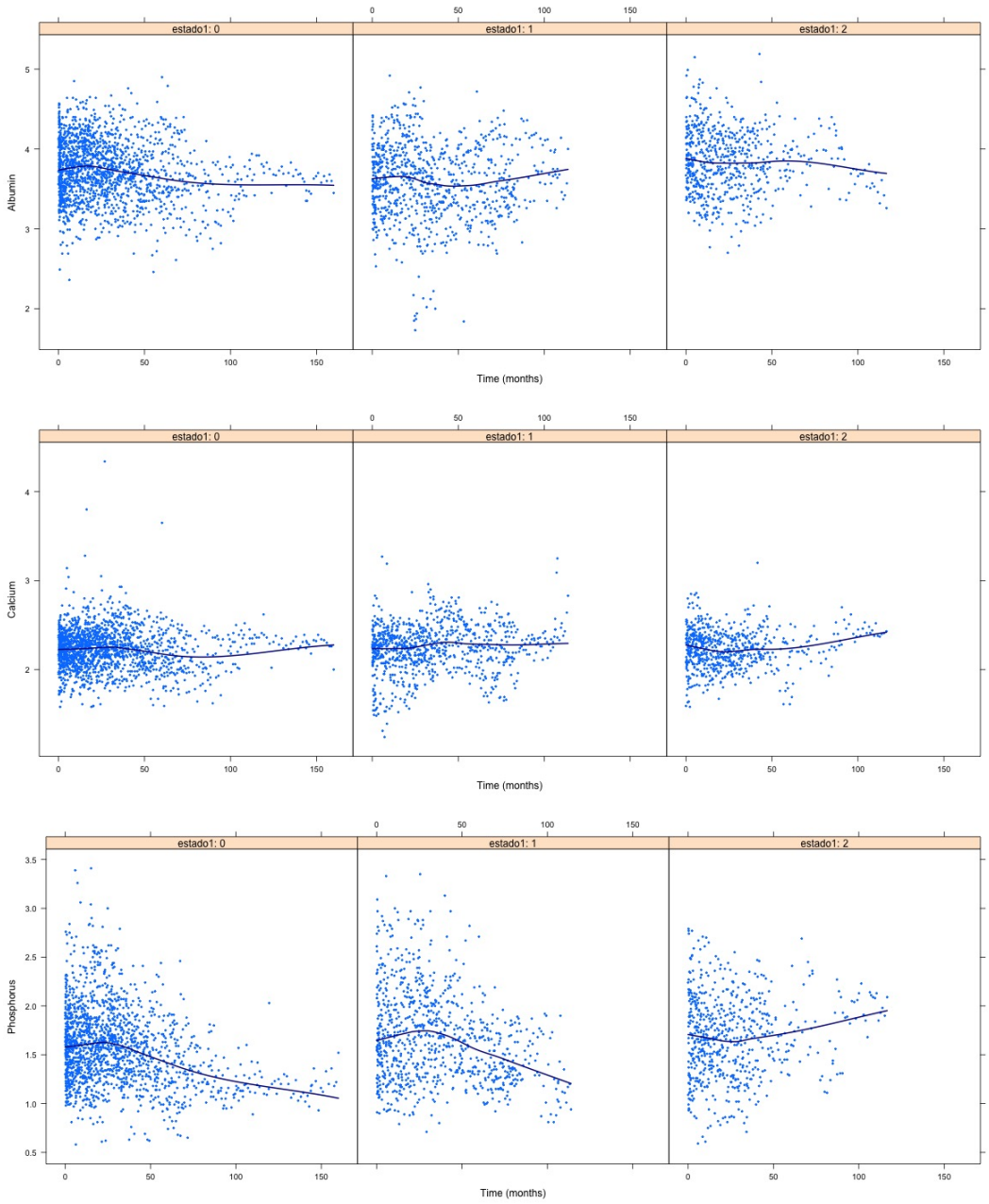

Figure 5.6: Longitudinal scores showing the progression of the albumin, calcium and phosphorus variables separated by the different events that arrived to the patients (0 censored, 1 death or transfer and 2 transplant).

Calcium

- Longitudinal Submodel:

$$y_{C,i}(t) = \beta_{C,0} + \beta_{C,1} Sex_i + \beta_{C,2} Age_i + \beta_{C,3} t + b_{C,0i} + b_{C,1i} t + \varepsilon_{C,i}(t).$$

- Survival Submodel:

$$\begin{cases} h_{C,i1}(t) = h_{C,01}(t)\exp\{\gamma_{C,11}Sex_i + \gamma_{C,12}Age_i + \alpha_{C,1}m_{C,i}(t)\}, \\ h_{C,i2}(t) = h_{C,02}(t)\exp\{(\gamma_{C,11} + \gamma_{C,21})Sex_i \\ \qquad +(\gamma_{C,12} + \gamma_{C,22})Age_i + (\alpha_{C,1} + \alpha_{C,2})m_{C,i}(t)\}. \end{cases}$$

Phosphorus

- Longitudinal Submodel:

$$y_{F,i}(t) = \beta_{F,0} + \beta_{F,1}Sex_i + \beta_{F,2}Age_i + \beta_{F,3}t + b_{F,0i} + b_{F,1i}t + \varepsilon_{F,i}(t).$$

- Survival Submodel:

$$\begin{cases} h_{F,i1}(t) = h_{F,01}(t)\exp\{\gamma_{F,11}Sex_i + \gamma_{F,12}Age_i + \alpha_{F,1}m_{F,i}(t)\}, \\ h_{F,i2}(t) = h_{F,02}(t)\exp\{(\gamma_{F,11} + \gamma_{F,21})Sex_i \\ \qquad +(\gamma_{F,12} + \gamma_{F,22})Age_i + (\alpha_{F,1} + \alpha_{F,2})m_{F,i}(t)\}. \end{cases}$$

The results of these joint models are gathered in Table 5.3. We must emphasize that each model has been fitted separately.

The longitudinal parameter estimates obtained are very similar to the ones calculated in section 5.2.1. for the three models: the albumin/calcium/phosphorus score remains almost constant along time, and age and sex have an influence in each variable sometimes it increases its average level, the other it decreases it).

When considering the survival submodel, not all the models have the same interpretations.

In one hand, we found a strong association between albumin and the risk of death or transfer to hemodialysis ($\hat{\alpha}_{A,1}$ = -1.20, p-value=0.0045), meaning that a unit decrease in albumin score corresponds to a $\exp(-(-1.20)) = 3.32$ increase in the risk for death/transfer to hemodialysis. We observe also an association between albumin and the risk of renal transplantation: it shows that a unit increase in albumin score correspond to a $\exp(1.72)) = 5.58$ increase in the risk for renal transplantation. Considering the factor age, younger patients have a statistically significant higher risk of getting a renal transplant rather than dying/being transfer to hemodialysis (hazard ratio for one year decrease in age equals $\exp(-(-0.03)) = 1.03$).

As for the second model, it appears that the calcium does not have a statistically significant influence on the survival. The only factor that seems to have a significant repercussion in it is the age, with a similar interpretation than in the previous case: younger patients have a higher risk for getting a renal transplant.

	Event Process				Longitudinal Process		
Albumin							
	Value	s.d.	p-value		Value	s.d.	p-value
$\gamma_{A,11}$ (sex)	0.41	0.30	0.1719	$\beta_{A,0}$	3.88	0.10	< 0.0001
$\gamma_{A,12}$ (age)	-0.01	0.01	0.3611	$\beta_{A,1}$ (sex)	0.24	0.06	0.0001
$\gamma_{A,21}$ (sex:CR)	0.05	0.46	0.9181	$\beta_{A,2}$ (age)	-0.01	0.00	0.0118
$\gamma_{A,22}$ (age:CR)	-0.03	0.02	0.0494	$\beta_{A,3}$ (time)	-0.00	0.00	0.1202
$\alpha_{A,1}$ (Assoc)	-1.20	0.42	0.0045				
$\alpha_{A,2}$ (Assoc:CR)	1.72	0.69	0.0119				
Calcium							
	Value	s.d.	p-value		Value	s.d.	p-value
$\gamma_{C,11}$ (sex)	0.05	0.28	0.8492	$\beta_{C,0}$	2.17	0.06	< 0.0001
$\gamma_{C,12}$ (age)	-0.00	0.01	0.8914	$\beta_{C,1}$ (sex)	-0.08	0.03	0.0107
$\gamma_{C,21}$ (sex:CR)	0.62	0.44	0.1597	$\beta_{C,2}$ (age)	0.00	0.00	0.0366
$\gamma_{C,22}$ (age:CR)	-0.05	0.02	0.0025	$\beta_{C,3}$ (time)	0.00	0.00	0.7663
$\alpha_{C,1}$ (Assoc)	-0.95	0.78	0.2240				
$\alpha_{C,2}$ (Assoc:CR)	1.80	1.22	0.1389				
Phosphorus							
	Value	s.d.	p-value		Value	s.d.	p-value
$\gamma_{F,11}$ (sex)	-0.04	0.29	0.8866	$\beta_{F,0}$	2.10	0.08	< 0.0001
$\gamma_{F,12}$ (age)	0.00	0.01	0.7426	$\beta_{F,1}$ (sex)	0.11	0.05	0.0241
$\gamma_{F,21}$ (sex:CR)	0.59	0.44	0.1759	$\beta_{F,2}$ (age)	-0.01	0.00	< 0.0001
$\gamma_{F,22}$ (age:CR)	-0.04	0.02	0.0059	$\beta_{F,3}$ (time)	0.00	0.00	< 0.0001
$\alpha_{F,1}$ (Assoc)	1.16	0.51	0.0236				
$\alpha_{F,2}$ (Assoc:CR)	-0.42	0.75	0.5718				

Table 5.3: Parameter estimates, standard errors and p-values for the three joint models considered above.

Finally, there is also an association between phosphorus and the risk of death/transfer, though in this case we do not found an association with the risk of renal transplantation. A unit increase in phosphorus score corresponds to a $\exp(1.16) = 3.18$ increase in

the risk for death/transfer. As for the factor age, in this last model we found that in also indicates that younger subjects have a higher risk of receiving a transplant.

5.3. Model comparison

Once we have fitted different models to the data, the next step would be to determine which one of them is more appropriate to describe the survival.

In order to carry out the comparison between the different models that have been displayed here, we use the linear predictors at time t to compute the ROC curves and the Area Under Curve (AUC) for each time point (Heagerty et al. 2005). This curves will give us an insight of the predictive performance of each model.

We have to take into account that these are AUCs curves for competing risks. Thus, we will have two different curves for each model: one for the event death/transfer and another one for the competing risk renal transplantation.

Firstly, we compare the joint models that we have discussed in the previous section with some competing risks models where we only take into account the baseline value of the longitudinal covariate considered at each model. The time-dependent AUCs computed for each pair of models (the joint model vs. the competing risk model) are shown in Figure 5.7.

In both the albumin and the phosphorus cases, the AUCs curves show that the joint model is preferable to the competing risk where only the baseline scores of albumin/phosphorus were considered, at least in the case of death/transfer. This does not occur for the calcium models: the curves do not differ significantly. This may be due to the fact that both albumin and phosphorus scores were detected as significant covariates in the joint models for the event of death/transfer, and calcium was not. In any case, it seems appropriate to say that the joint models turn out to be a better way to analyze this data than the basic competing risk approach.

On the other hand, we could ask ourselves the question of which one of those models we would recommend to a doctor who wants to understand the survival aspects of these kind of peritoneal dialysis data. To answer this question, we compare the AUCs curves for the three joint models adjusted and also the curve for the competing risk model explained in section 5.2.2, the one who included as covariates both the albumin and the phosphorus scores. This curves are compared in Figure 5.8.

From what it showed in that Figure, the joint model that uses the phosphorus as longitudinal covariate is the best model to explain the survival due to death/hemodialysis.

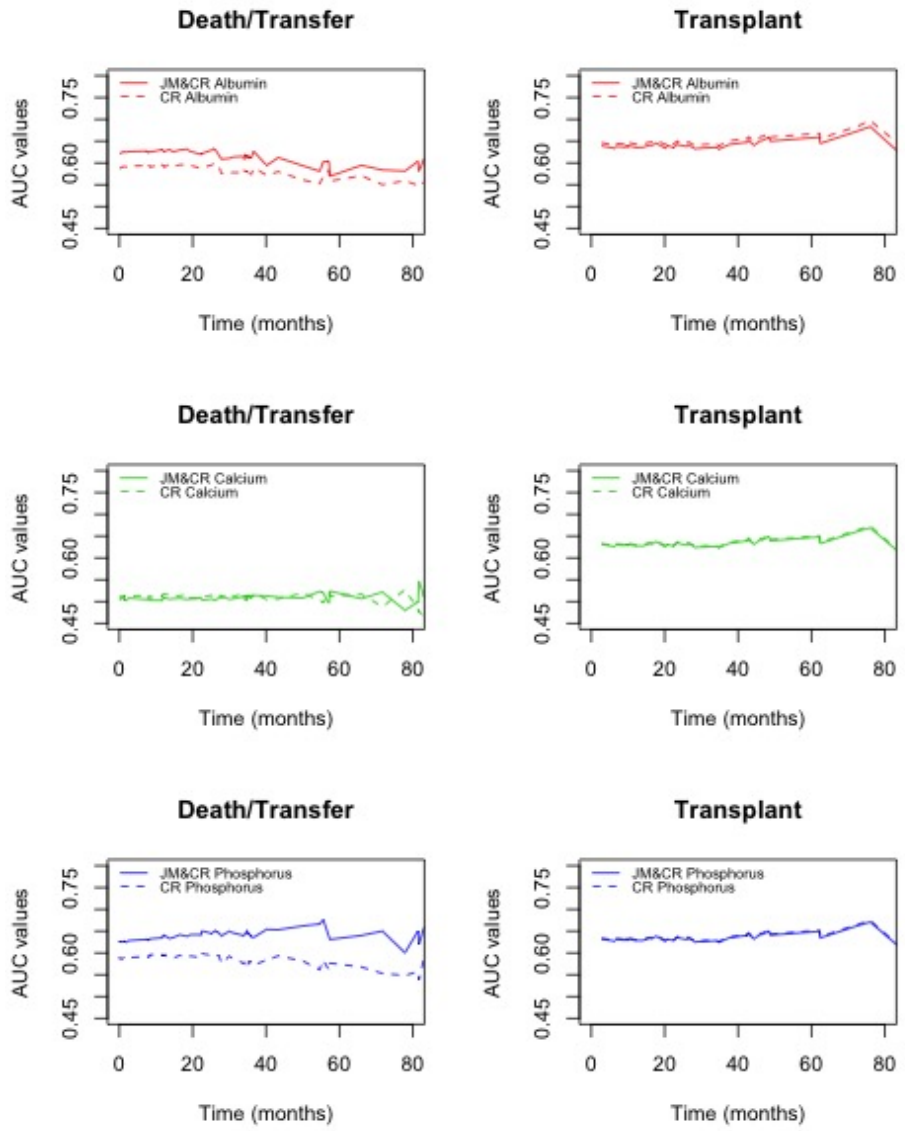

Figure 5.7: Time-dependet AUCs for each pair of models, each one of the pairs considering only one longitudinal covariate: albumin, calcium or phosphorus.

However, it seems that it is the competing risk simple model the one that behaves better when analyzing the renal transplantation. Though this is a simpler model, with no longitudinal measurements, it has the advantage that it can combine the phosphorus and the albumin in the same model, something that is yet under study for the joint modelling. As both longitudinal covariates appear to influence on the survival, it would be very interesting to be able to adjust a model with both their scores along time.

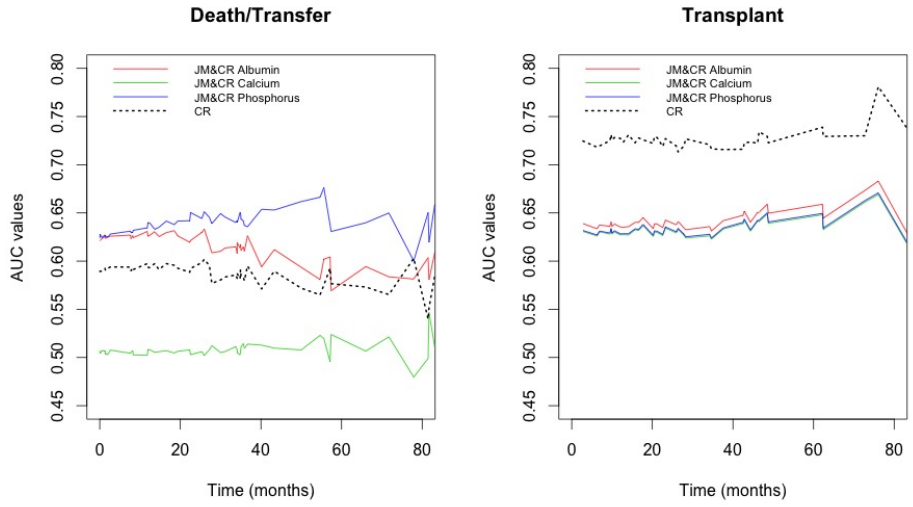

Figure 5.8: Time-dependent AUCs for the joint models with competing risks and the competing risk model considered in 5.2.2.

Apart from this, there are also methods to validate the assumptions behind mixed models and relative risk models when these are fitted separately, but this area is still difficult to cover when considering the three blocs of longitudinal data, survival data and competing risks all together. Rizopoulos (2012) propose the multiple imputation residuals for fixed visit times for the joint model, but when we add the competing risks it gets more complicated, so we do not include that validation here.

5.4. Results

Throughout this chapter we have seen different ways to approach the analysis for the peritoneal dialysis data. Although in general it seems that a joint model considering competing risk is the best option to do the analysis, depending on the event of interest we could also analyze the data without taking into consideration the longitudinal measurements.

We have found that what medicine tell us traditionally is true: a decreasing albumin level is associated with a higher risk to die/receive a transfer to hemodialysis, both of the events related with kidney failure, and the opposite happens with the phosphorus: when it rises, the risk of dying or being transfer to hemodialysis is higher.

Apart from this, younger subjects are shown to have a higher risk of getting a renal transplant. This is to be expected, as a renal transplant could improve significantly

their survival and it is a more permanent solution to their disease than any dialysis.

When analyzing the adjustment of the longitudinal covariates , the factors sex and age have proven to have a statistical significant influence in their behavior.

As we can observe in the context of predictive performances of the different possible models to study the longitudinal markers with competing risks, the joint modelling approach improves the regression model if the longitudinal marker has a significant effect on the competing risks. The baseline measurements of the longitudinal biomarkers may not be enough to explain their time-varying effect on the competing risks. Therefore it would be appropriate to use joint modelling approaches to explain the relationship between the longitudinal and competing risks process.

5.5. Software

In this section we present a brief discussion of the existing software that can be used to perform the analysis explained in this project.

All the analysis were implemented in the `R` software environment. More specifically, the packages that were used were `lattice` (for graphical purposes), `nlme` (for the linear mixed-effects models), `mstate` and `cmprsk` (for the competing risk adjustment), `JM` (for the joint modelling) and `risksetROC` (for the implementation of the AUCs curves).

Of all these packages, we would like to highlight the `JM` package: it was developed by Rizopoulos (2010) and it constitutes a useful tool for the joint modelling of longitudinal and time-to-event data. It contains all the joint modelling methodology explained above: the functions `jointModel()` is the one that fittes the joint models. It uses as arguments the outputs of the function `lme()` (that models the mixed-effects) and `coxph()` (that comes from the package `survival`). It has also the argument `CompRisk` (`FALSE` by default), which is the one that allows the model to consider the competing risk situation.

The result of using the function `jointModel()` is an object of class `jointModel`, which can be used to obtain the general results through functions like `summay()` or `print()`. Though the function `predict()` can be applied in this type of object, its use is restricted to the case in which there are no competing risks. However, Saha and Heagerty (2010) are developing a code to trate this time-dependent predictive accuracy in the presence of competing risks.

Chapter 6

Conclusions

It is very common to found clinical studies with both longitudinal measurements and event times, where this measures are recorded on subjects during follow-up. Joint models arise as an appropriate technique when interest lies in the association between a longitudinal covariate measured with error in a survival analysis. Several simulation studies have shown that joint model could be substantially more efficient than the separate analysis, given that these models use information from both outcomes.

The presence of informative censoring, on the other hand, is also quite common in this kind of studies. Thanks to the joint models that take into account the possible competing risks, it is possible to evaluate the association between the two processes, contributing for a better knowledge of the data. Though these competing risks can be approached from a simpler point of view (taking into account only the baseline covariates) by doing this we could be loosing important information about how the longitudinal covariate affects the different events.

Based in this procedures, in this project we have studied data from a peritoneal dialysis program. We have seen how different longitudinal measurements affected the survival. In particular, we have proven that albumin and phosphorus levels play an important role in the risk for death/transfer to hemodialysis, while we did not find any reasons to think that the calcium score is associated with it or with the renal transplantation event.

Accordingly with the results obtained, it has been observed a better predictor performance of the joint model through the time dependent AUCs curves, though a competing risk model where we consider both the effect of the baseline albumin and phosphorus levels can also explain the data satisfactorily.

Like in any medical study where statistical tools are needed, further progress in

this area is needed to continue to develop new tools that allows us to make a better analysis of the survival with longitudinal data and in presence of competing risk. One line of research would be including more than one longitudinal covariate at a time in the joint models. Others will be to make dynamics predictors to illustrate how all the available information helps to produce predictions of the survival probabilities. Though some of this concepts already exist in the joint modelling area, not a lot of them can be extended to the case in which we have to deal with competing risks.

Bibliography

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

[2] Beyersmann, J., Allignol, A., Schumacher, M. (2012). Competing Risks and Multistate Models with R. *Springer*.

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

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

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

[6] de Boor, C. (1978). A Practical Guide of Splines. *Springer*, Berlin.

[7] Dempster, A., Laird, N., Rubin, D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39: 1- 38.

[8] Fine, J.P., Gray, R.J. (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association* 94(446): 496-509

[9] Harrel, F. (1982). Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. *Springer-Verlag*, New York.

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

[11] Heagerty, P. J., Saha, P. (2010). Time-dependent predictive accuracy in the presence of competing risks. *Biometrics* 66(4):999-1011.

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

[13] Henderson, R., Diggle, P., Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* 1: 465-480.

[14] Hsieh, F., Tseng, Y. K., Wang, J. L. (2006). Joint modelling of survival and longitudinal data: likelihood approach revisited. *Biometrics* 62: 1037-1043.

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

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

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

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

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

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

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

[22] Rizopoulos, D. (2012). Joint Models for Longitudinal and Time-to-Event Data with Applications in R. *CRC Press.*

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

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

List of figures

List of tables