UNIVERSITY OF GENOVA

Department of Mathematics

PhD in Mathematics and Applications

Graduate Programme in Mathematics and Applications

Final Dissertation

The use of the Joint Models to improve the accuracy of prognostication of death in patients with heart failure and reduced ejection fraction (HFrEF)

Supervisor: Prof. Eva Riccomagno

Candidate: Giacomo Siri

XXXIII Cycle

Contents

Introduction

The work presented in this thesis has been developed during a scholarship at the Scientific Directorate - Unit of Biostatistics of the Galliera Hospital in Genoa under the supervision of Dr. Matteo Puntoni. This scholarship was partially supported by a grant from Ministry of Health, Italy "Bando Ricerca Finalizzata - Giovani Ricercatori" (Project code: GR-2013-02355479) won by Dr. Puntoni for conducting a cancer research study. The main objective of my research was to apply the Joint Model for longitudinal and survival data to improve the dynamic prediction of cardiovascular diseases in patients undergoing cancer treatment. These patients are usually followed after the start of the therapy with several visits in the course of which different longitudinal data are collected. These data are usually collected and interpreted by clinicians but not in a systematic way. The innovation of my project consisted in a more formal use of these data in a statistical model.
The Joint Model is essentially based on the simultaneous modelling of a linear mixed model for longitudinal data and a survival model for the probability of an event. The utility of this model is twofold: on one hand it links the change of a longitudinal measurement to a change in the risk of an event, on the other hand the prediction of survival probabilities using the Joint Model can be updated whenever a new measurement is taken.
Unfortunately, the clinical study on cancer therapy for which the project was thought is still ongoing at this moment and the longitudinal data are not available. So, we applied the developed methods based on Joint Model to another dataset with a similar clinical interest. The case of study presented in the Chapter 6 of this thesis is developed after a meeting between Dr. Puntoni and me and Dr. Marco Canepa of the Cardiovascular Disease Unit of the San Martino Hospital in Genoa. The necessity of the last one was to prove that the longitudinal data collected in patients after a heart failure could be used to improve the prognostication of death and, more in general, the patient management and care with a personalized therapy. The last one could be better calibrated by a dynamic update of the prognosis of patients related to a better analysis of the longitudinal data provided during

each follow-up visit.
The Joint Model for longitudinal and survival data solves the problem of the simultaneous analysis of the biomarkers collected at each follow-up visits and the dynamic update of the survival probabilities each time a new measurements are collected (see Chapter 4). The next step, developed in the Chapter 5, was to find a statistical index that was simple to understand and practical for clinicians but also methodologically adequate to assess and prove that the longitudinal data are advantage in the prognostication of death. To do this, two different indexes seemed most suitable: the area under the Receiver Operating Characteristic Curve (AUC-ROC) to assess the prediction capability of the Joint Model, and the Net Reclassification Improvement (NRI) to evaluate the improvement in prognostication in comparison with other approaches commonly used in clinical studies.
In Section 5.3, a new definition of time-dependent AUC-ROC and time-dependent NRI in the Joint Model context is given. Even if a function to derive the AUC after a Joint Model was present in literature, we needed to reformulate it and implement in the statistical software `R` to make it comparable with the index derived after the use of the common survival models, such as the Weibull Model. Regarding the NRI, no indexes are present in the literature. Some methods and functions were developed for binary and survival context but no one for the Joint Model. A new definition of time-dependent NRI is presented in Section 5.3.2 and used to compare the common Weibull survival model and the Joint Model.
This thesis is divided in 6 chapters. Chapters 1 and 2 are preparatory to the introduction of the Joint Model in Chapter 3. In particular, Chapter 1 is an introduction to the analysis of longitudinal data with the use of Linear Mixed Models while Chapter 2 presents concepts and models used in the thesis from survival analysis. In Chapter 3 the elements introduced in the first two chapters are joined to defined the Joint Model for longitudinal and survival data following the approach proposed by Rizopoulos[80]. Chapter 4 introduces the main ideas behind dynamic prediction in the Joint Model context. In Chapter 5 relevant notions of prediction capability are introduced in relation to the indexes AUC and NRI. Initially, these two indexes are presented in relation to a binary outcome. Then, it is shown how they change when the outcome is the time to an event of interest. Ending, the definitions of time-dependent AUC and NRI are formulated in the Joint Model context. The case of study is presented in the Chapter 6 along with strength and limitations related to the use of the Joint Model in clinical studies.

Chapter 1

Longitudinal Analysis

A set of observations collected repeatedly over time on the same subject is becoming one of most common dataset with which a biostatistician has to work. The spread of information technology in various aspects of the life, including the healthcare and hospital environment, has involved the possibility to collect a huge amount of data related to the clinical history of the patients. In this context, the presence of several measurements over time on the same subject is a very common situation and this is the main feature that distinguishes longitudinal studies from others. A constant monitoring of the subjects among a visit process and the subsequent collection of the data allow a direct assessment of changes in the outcomes of interest in a clinically relevant time window.
Mainly, a longitudinal model is used to investigate two types of effects:

- cross-sectional effects, i.e. the differences among groups at a given a specific time point (e.g. the mean difference between male and female, or between two arms of treatment);
- longitudinal effects, i.e. the effect of the time on the outcome or different time effects among groups of subjects (e.g. the mean trajectory of the mean blood pressure after starting a therapy, or the difference between the trajectory observed in males and females).

From a statistical point of view, assuming that data collected at each visit constitute a record, the main characteristic of the longitudinal study is the presence of a correlation structure behind the records related to the same subject. This situation is in contrast with the assumption of independence among the residuals typical, for example, of the linear model and requires the use of ad hoc statistical models. The main problem is to avoid an underestimation of the variability of each effect which could lead to a more narrow

confidence intervals and to a rejection of the null hypothesis when this is true.
A good model for longitudinal data should be able to account for three sources of variability that characterized this type of data [28]:

- random variability coming from heterogeneity among individual trajectories;
- serial correlation due to residuals close to each other in time are more similar than residuals further apart;
- measurement error to account for small variability unavoidable even from an immediate replication of the measurement (noise variability).

Often the homoschedasticity assumption is also violated by longitudinal data, in fact data collected in different occasions have different variability. Moreover longitudinal data have often an unbalanced number of repeated measurements among subjects and not necessarily taken at fixed time points. This last problem and the other described previously make longitudinal data untreatable with standard multivariate regression techniques.

1.1 A two-stage approach

As mentioned above, a longitudinal dataset has a particular structure that makes it unsuitable to be modeled with classical linear models. A linear mixed model (LMM), that will be defined below, is a parametric linear model for longitudinal data (and also for clustered or hierarchical data) that quantifies the relationships between a continuous dependent variable and various predictor covariates. In this model, the relation between the outcome and the covariates (continuous or categorical) is defined between two groups of effects that the latter cause on the former:

- fixed effects, that describe the mean structure model between covariates and outcome in the whole sample;
- random effects, that are related to random cluster-specific (e.g. subjects, hospitals, ...) variations from the overall mean structure.

The second set of effects is also responsible for managing the correlation among repeated measurements on the same subject. The presence of both fixed and random effects gives the name to the model.
The two-stage approach proposed by Verbeke [108] will be followed to define the LMM: the first stage tries to explain the longitudinal response of

interest for each subject by a vector of a small number of estimated subject-specific regression coefficients; in the second, another regression model will relate the estimates obtained in the first stage to known covariates such as treatment, disease classification, patients' demographics and baseline characteristics. Finally, the combination of the two stages into one statistical model will provide the general formulation of the linear mixed model.
Throughout this work, we assume N subjects, each of whom is measured at n_i time points ($i = 1, \ldots, N$), not necessarily $n_i = n_j$ for $i \neq j$, with $i, j = 1, \ldots, N$. The response of interest is modelled by a dependent random variable indicated with Y, which is assumed continuous (other distributions are possible in the Generalized LMMs context). The fixed effects are modelled by random vectors with both continuous and categorical components and all of them are indicated within the matrix X; the random effects, indicated by the matrix Z, are defined as those factors that, in addition to a fixed effect, can have an effect that varies from subject to subject.

1.1.1 Stage 1

Let Y_{ij} be the random variable denoting the outcome of interest for the ith subject measured at time j, with $i = 1, .., N$ and $j = 1, \ldots, n_i$ and where N is the number of subjects while n_i is the number of repeated measurements for the i-th subject. Then $Y_i = (Y_{i,1}, .., Y_{i,n_i})^T$ is the vector of continuous responses for i-th subject. In the first stage we assume that Y_i satisfies the linear regression model:

$$Y_i = Z_i\gamma_i + \varepsilon_i \tag{1.1}$$

where Z_i is a $(n_i \times q)$ matrix whose rows give the values of q covariates over time for subject i. In the context of longitudinal analysis, one of the q covariates is time, while also other covariates could be functions of time. In this case, the Equation 1.1 models how the response evolves over time for the i-th subject. Therefore, in analogy with Generalized Linear Models, γ_i is the vector of q subject-specific regression coefficients and $\varepsilon_i \sim \mathcal{N}_{n_i}(0, \Sigma_i)$ is the vector of residuals where Σ_i is the n_i-dimensionality covariance matrix and $\mathcal{N}_{n_i}$ stands for a multidimensional Gaussian distribution, taken to be centered in zero.
For the application in the definition and use of the Joint Model (Chapters 3 to 6), time is included as a simple linear relationship and Z has only two columns: the intercept and time. Any other covariates will be included in another specific way.
Note that Equation 1.1 could as well be used for multilevel or cluster data. In these cases, the matrix Z could contain some variables necessary to uniquely

identify the repeated measurements within the same cluster (e.g. patients within the same ward).

1.1.2 Stage 2

In the second stage, the goal is to explain the between-subjects variability, modelling the relationship between the $\gamma_i = (\gamma_1, \ldots, \gamma_q)$ coefficients obtained in the first stage and a set of known covariates contained in a $(q \times p)$ matrix indicated with K_i. The model is given by:

$$\gamma_i = K_i\beta + b_i \tag{1.2}$$

where $b_i \sim \mathcal{N}_q(0, D)$ is a q-dimensional residual vector and D its covariance matrix , β is a p-dimensional vector of unknown regression parameter.
Hence, the core of the model is given by the estimate of the regression parameters γ_i that can be obtained by a sequential fitting of the two models in the two stages. This sequential fitting can be interpreted as the analysis (second stage) of the summary statistics calculated in the first stage and this involves at least two problems [108]:

- the estimated vector of effects $\hat{\gamma}_i$ summarizes the information on the longitudinal response Y_i for the subject i, obtained in the first stage, but it carries with it a loss of information;
- in the second stage, the replacement of γ_i with their estimates $\hat{\gamma}_i$ is another source of variability.

The Linear Mixed Model, which will be presented in the next section, is motivated by the need of addressing these two problems and consists of combining the two stages in a single model with a simultaneous parameter estimation process. Despite these two issues, the two-stage estimation is not computationally expensive and it could be used in practice when convergence problems are encountered with the Linear Mixed Model.

1.2 The Linear Mixed Model (LMM)

In order to obtain a single model, we can replace γ_i of the second stage in the first stage, yielding:

$$\begin{aligned} Y_i &= Z_i\gamma_i + \varepsilon_i \\ &= Z_i(K_i\beta + b_i) + \varepsilon_i \\ &= Z_iK_i\beta + Z_ib_i + \varepsilon_i \\ &= X_i\beta + Z_ib_i + \varepsilon_i \end{aligned}$$

in which the name of linear mixed model is given by the mixed presence of fixed effects β and random subject-specific effects b_i. Therefore, the longitudinal outcome for each subjects can be seen as a linear regression model where there are population-specific effects (i.e. common at the whole group of patients) and subject-specific variations from the mean population. In summary, using the definition by Laird and Ware [48], LMM is defined as follow.

Definition 1. *A linear mixed-effects model is any model which satisfies the following relationship for each subject* $i = 1, \ldots, N$*:*

$$\begin{cases} Y_i = X_i\beta + Z_i b_i + \varepsilon_i \\ b_i \sim \mathcal{N}_q(0, D) \\ \varepsilon_i \sim \mathcal{N}_{n_i}(0, \Sigma_i) \\ b_1, .., b_q, \varepsilon_1, .., \varepsilon_{n_i} independent \end{cases}$$

where Y_i *is the* n_i*-dimensional response vector for subject* i*,* X_i *and* Z_i *are* $(n_i \times p)$ *and* $(n_i \times q)$ *dimensional matrices,* β *is a* p*-dimensional vector,* b_i *is a* q*-dimensional vector and* ε_i *is a* n_i*-dimensional vector of residual components. Finally,* D *is a* $(q \times q)$ *covariance matrix while the matrix* Σ_i *is a* $(n_i \times n_i)$ *covariance matrix.*

For ease of interpretation, for each $i = 1, \ldots, N$

- Y_i is the response vector for subject i,
- X_i and Z_i are matrices whose elements are the known values of covariates for subject i,
- β contains the fixed effects,
- b_i contains the random effects,
- D models the associations among the random factors in Z,
- Σ_i is a subject-specific covariance matrix whose dimension depends on the number of repeated visits done by the i-subject (n_i) and represents the relationship among the residuals for the i-subject.

The LMM allows the researchers to estimate all elements of the covariance matrices for the random effects D and the residuals Σ_i or to define a priori a structure for the two matrices.
Starting from the D matrix, the most common approach is set it as a matrix

with only variance components:

$$D = \begin{bmatrix} \sigma^2_{Z_1} & 0 & \dots & 0 \\ 0 & \sigma^2_{Z_2} & \dots & 0 \\ 0 & \dots & \ddots & 0 \\ 0 & \dots & 0 & \sigma^2_{Z_q} \end{bmatrix}$$

where $\sigma^2_{Z_q}$ is the variance for the p-th random effect. In this case, the researcher assumes the independence among the random effects. Another common solution, even if more computationally expensive, is given by the use of an unstructured covariance matrix where all the element of the half matrix are estimated from the data:

$$D = \begin{bmatrix} \sigma^2_{Z_1} & \sigma^2_{1,2} & \dots & \sigma^2_{1,q} \\ \sigma^2_{1,2} & \sigma^2_{Z_2} & & \sigma^2_{2,q} \\ \vdots & & \ddots & \vdots \\ \sigma^2_{1,q} & \dots & \dots & \sigma^2_{Z_q} \end{bmatrix}$$

where $\sigma^2_{a,b}$ is the covariance between the a-th and the b-th random effect, with $a, \neq b$ and $a, b = 1, \dots, q$.
Several other matrices can be defined and are available in the most largely used statistical software.

The considerations done on D can also be done for Σ_i and several structure can be defined for this matrix. In general, the same variance structure is assumed for each $i = 1, \dots, N$ subject. Also in this case, the most common solution is to define a variance components matrix:

$$\Sigma_i = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ 0 & \dots & \ddots & 0 \\ 0 & \dots & 0 & \sigma^2 \end{bmatrix}$$

where σ^2 is the variance of the residuals which is assumed constant over the n_i repeated measurements for each subject $i = 1, \dots, N$. In the context of longitudinal analysis a first-order autoregressive covariance matrix is also used and it assumes the following structure:

$$\Sigma_i = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \dots & \rho^{n_i-1}\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \dots & \vdots \\ \vdots & & \ddots & \rho\sigma^2 \\ \rho^{n_i-1}\sigma^2 & \dots & \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

where the variance σ^2 is assumed homogeneous over the repeated measurements but the correlation ρ declines exponentially with distance. In case of longitudinal data, this means that the variability in the outcome is constant regardless of when the measure is taken but also that two subsequent measurements are more correlated than other two taken in more distant visits.

1.2.1 LMM in clinical trials

In clinical trials, the LMMs are widely used to model the longitudinal trajectories of the endpoint or the change from the baseline value. Depending on the outcome that is chosen to analyze - value or change from baseline - the covariates inserted in the fixed part of the model varies while the random part is usually composed by the random intercept and the random slope for the time effect.
In the first case, when the current value of the endpoint is modelled, the time variable, that can be both a continuous or a discrete variable, starts from 0 and the other covariates are usually kept constant at their baseline value. This approach is often used when the patients are random extracted from the target population and hence no significant differences in the baseline values are expected among the patients. In this case, the equation 1 can be rewritten as:

$$\mathbb{E}[Y_i] = \beta_0 + T_i\beta_t + X_i\beta_p + b_{0,i} + T_i b_{t,i}$$

where the time T_i is modelled both as fixed and random effect and X_i are the $(n_i \times p)$ matrix that contains the other covariates set at their baseline values. In this case, the interpretation of the coefficients are the following:

- β_0 is the intercept that represents the mean value of Y in the whole sample when the other covariates are set in their reference value (if categorical variables) or set to zero (if continuous variables);
- β_t is the mean time effect in the whole sample;
- β_p contains the mean effect of each covariate in X_i;
- $b_{0,i}$ is called *random intercept* and represents the subject-specific variation in the outcome at baseline for the i-th subject;
- $b_{t,i}$ is called *random slope* and represents the subject-specific variation which must to be added to β_t to obtain the mean trajectory of the outcome for the i-th subject.

In case of problem of convergence, it is common to remove, in order, the random slope and then the covariates from least significant to most significant. In the second case, used to study the change from baseline at each following time-points, the time variable starts from the first post baseline assessment while the baseline value of the outcome and its interaction with time are inserted in the model to account for a potential influence of the baseline value which is assumed to lose effectiveness as time goes on.

1.2.2 The population-averaged model under the LMM

The model described in the Definition 1 is referred to as a *subject-specific* model because the random effects are formally used to explain the random variation from a subject to another, or from a cluster to another in the case of a multilevel model.
Starting from Definition 1, it is possible to derive a marginal model to analyse the relationship between the fixed factors and the outcome. It does not explicitly use the random factors in the equation. In this way, it is possible to derive the marginal effect of a covariate, such as time, on the outcome, by modelling a mean trajectory over the entire sample ignoring the subject variations but accounting for them. For this reason, this derived model is also called *population-averaged* model because the random subjects-specific deviations from the mean trajectories are not directly shown. The general formulation of a population-averaged model is given by the following definition.

Definition 2. *A population-averaged model is any model which satisfied the following conditions for each subject* $i = 1, \dots, N$*:*

$$\begin{cases} Y_i = X_i\beta + \varepsilon_i^* \\ \varepsilon_i^* \sim \mathcal{N}(0, V_i^*) \end{cases}$$

where X_i *is the design matrix with dimensions* $(n_i \times p)$*,* β *is the vector of the fixed effects* ε_i *represents a vector of marginal residuals errors and* V_i^* n_i*-dimensional matrix. Furthermore* $V_i = Z_i D Z_i^T + \Sigma_i$ *where* Σ_i *,* Z_i *and* D *are as in Definition 1.*

In synthesis, Definition 2 becomes for each subject

$$Y_i \sim N_{n_i}(X_i\beta, V_i)$$

with

$$V_i = Z_i D Z_i^T + \Sigma_i$$

and in standard vector notation including all subjects it becomes:

$$Y = X\beta + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, V)$ is the vector of residual components and V is a block matrix which contains in the diagonal the matrices $V_i = Z_i D Z_i^T + \Sigma_i$ which explains the relationship among the n_i repeated measures on the same subject i.

1.3 Estimation in LMM

The parameters' estimate for the linear mixed model is based on the criterion of maximum likelihood (ML) applied on marginal density derived in Definition 2 for the i-th subject:

$$Y_i \sim \mathcal{N}_{n_i}(X_i\beta, V_i)$$

$$V_i = V_i(\alpha) = Z_i D Z_i^T + \Sigma_i$$

where:

- α is the column vector of all parameters of the covariance matrix (*variance components*) found in V_i. In α there are at most $\frac{q(q+1)}{2}$ different elements of D and in Σ_i; the actual number of different elements depends on the choice of the shape of the variance-covariance matrices as explained in Section 1.2.
- $\vartheta = (\beta^T, \alpha^T)^T$ is the column vector of all parameters in the marginal model for Y_i.

According to the classical maximum likelihood approach, under independence of the Y_i's, the log-likelihood function is

$$l_{ML}(\vartheta, y) = \sum_{i=1}^{N} \log p(y_i; \vartheta) = \sum_{i=1}^{N} \log p(y_i; \beta, \alpha)$$

where y is the matrix of observed responses, whose column y_i i($i = 1, \ldots, N$) is the observed response vector for the i-th subject. Furthermore,

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

hence

$$l_{ML}(\vartheta, y) = -\frac{1}{2}\log(2\pi)\sum_{i=1}^{N} n_i - \frac{1}{2}\sum_{i=1}^{N}\log|V_i(\alpha)| - \frac{1}{2}\sum_{i=1}^{N}\left[(y_i - X_i\beta)^T {V_i}^{-1}(\alpha)(y_i - X_i\beta)\right]$$

Assuming α known, the ML estimate for β is given by

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

with $V_i = V_i(\alpha)$. In this case, even if V_i is not known, but an estimate of its elements $\hat{\alpha}$ is available, we can estimate β by replacing α with $\hat{\alpha}$ and obtaining $\hat{V}_i = V_i(\hat{\alpha})$. Then, the ML function will be

$$\hat{\beta}_{ML}(\hat{\alpha}) = \left(\sum_{i=1}^{N} X_i^T {\hat{V}_i}^{-1} X_i\right)^{-1} \sum_{i=1}^{N} X_i^T {\hat{V}_i}^{-1} y_i$$

When α is not know, it is possible to estimate it by using a ML estimate of α. It can be obtained by maximizing $l_{ML}(\vartheta, y)$ with respect to α, substituting β with its ML estimate. The drawback of this approach is that it does not estimate α and β simultaneously but β is previously estimated from the data. In this way, the ML estimate of α is biased downward because it does not take into account the loss of degrees of freedom that results from estimating the fixed-effect parameters in β. However, as known from the theory of the asymptotic properties of ML estimators, under certain conditions of regularity, the ML estimate of V_i will be asymptotically unbiased[1]. In a linear regression model $Y = X\beta + \varepsilon$, where Y is an N-dimensional vector, X a $(N \times p)$ matrix of known covariates and $\varepsilon \sim N(0, \sigma^2)$, the ML estimate for σ^2 is given by:

$$\hat{\sigma}^2 = \frac{\sum_i (y_i - x_i^T\hat{\beta})^2}{N}$$

[1]The regularity conditions are the following:

- $X_1, \ldots, X_N$ are independent and identically distributed;
- the support of X_i, $i = 1, \ldots, N$, does not depend on the parameter ϑ;
- the likelihood function is continuous and differentiable.

Under the above mentioned conditions, an ML estimator is consistent, unbiased and asymptotic normal distributed.

biased downward by a factor $\frac{N-p}{N}$. This bias comes from the fact that $\hat{\beta}_{ML}$ is previously estimated on data and it is necessary to correct for $(N-p)$ remaining degrees of freedom.

Patterson and Thompson [62], Harville [35] [36] et other authors proposed the use of the REsidual Maximum Likelihood (REML) to remove the bias in the ML estimates of the covariance parameters. The residual (or restricted, or reduced) maximum likelihood (REML) approach is a particular form of maximum likelihood estimation that does not base estimates on a maximum likelihood fit of all the information, but instead uses a likelihood function calculated from a transformed set of data, so there is no effect related to the nuisance parameters. In the case of variance component estimation like this, the original matrix of data is replaced by a set of contrasts calculated from the data, and the likelihood function is calculated from the probability distribution of these contrasts, according to the model for the complete data set. In other words, the objective the REML is to obtain an estimate for α not depending on β and an estimate for β that depends on α. Moreover, REML can produce unbiased estimates of variance and covariance parameters.
Starting from the Definition 2 it is possible to obtain the REML estimator:

Theorem 1. *Let $Y = X\beta+\varepsilon$ the linear mixed model with $Y_i \sim \mathcal{N}_{n_i}(X_i\beta, V_i(\alpha))$ for $i = 1, \ldots, N$. Then the REML* log*-likelihood function for $\vartheta = (\beta, \alpha)$ can be written as*

$$l_{REML}(\beta, \alpha; y) = C - \frac{1}{2}\log\left|\sum_{i=1}^{N} X_i^T V_i^{-1} X_i\right| + l_{ML}(\hat{\beta}(\alpha), \alpha; y)$$

where C is a constant not depending on α, $V_i = V_i(\alpha)$, $|\cdot|$ denote the determinant of the matrix inside and $l_{ML}(\hat{\beta}(\alpha), \alpha; y)$ is the ML log-likelihood function presented previously.

Because $\left|\sum_{i=1}^{N} X_i^T V_i^{-1} X_i\right|$ does not depend on β, it follows that the REML estimators for α and β can also be found maximizing the so-called REML log-likelihood function:

$$l_{REML} = -\frac{1}{2}\log\left|\sum_{i=1}^{N} X_i^T V_i^{-1} X_i\right| + l_{ML}(\hat{\beta}(\alpha), \alpha; y)$$

with respect to all parameters simultaneously (α and β).

If we could find a matrix A such that $dim(A) = n - rank(X) = n - p$ with columns given by $a_1, .., a_{n-p}$ linearly independent vectors and such that $A^T X = 0$ we can use a transformed set of data $U = A^T y$ to find the maximum

likelihood estimate of ϑ. Let $P_X = X(X^TX)^{-1}X^T$ be a projector on the space generated from X, we can note that $(I - P_X)X = X - P_XX = X - X = 0$ and $(I - P_X)y = y - P_Xy = y - \hat{y}$ is known as error contrast. Because $rank(I - P_X) = n - rank(X) = n - p$, there exist a set of $n - p$ linearly independent rows of $I - P_X$ that can be used to get the matrix A. The name "REML" derives from the fact that if we get a subset of $(I - P_X)$ as columns of the matrix A then we'll use a subset of the elements of residual vector $(I - P_X)y = y - \hat{y}$ as U. be valid $AA^T = I - P_X$ and $A^TA = I_{n-p}$ We will now state two lemmas and then use them to prove the REML theorem.

Lemma 1. *Let A be a matrix defined as previously then*

$$U = A^TY \sim N_{n-p}(0, A^TVA)$$

with $V = V(\alpha)$. Thus, the distribution of U depends on α but not β.

Lemma 2. *Let A be a matrix defined as previously, and $G = V^{-1}X(X^TV^{-1}X)^{-1}$, then*

(a) $|[A, G]|^2 = \left|X^TX\right|^{-1}$

(b) $\left|(A^TVA)\right| = \left|X^TV^{-1}X\right| |V| \left|X^TX\right|^{-1}$

(c) $A(A^TVA)^{-1}A^T = V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}$

where $|\cdot|$ denotes the determinant of matrix inside.

The proofs of the Theorem 1 and related Lemmas 1 and 2 can be found in the Appendix 1. Because the estimates for α can not be written, in general, in closed form, a numerical optimization procedures is required. One usually uses Newton-Raphson algorithms whose implementation for LMM was thoroughly described in [53].

1.4 Model selection

The choice of the "best" model is hard for researchers who are faced with several competing models and must take into account research objectives, sample size, study design, known predictors and source of confounding or bias. Moreover the best model should be a model that is parsimonious in terms of the number of covariates. The selection of the model that best

fits the longitudinal data is an iterative process that requires to fit several models: at each step, different elements (e.g. the structure for the fixed effects the population average model, the choice of the random effects and the form for the variance/covariance matrices) will be tested and chosen or discarded in order to obtain the best model. At the end of this process it will be necessary to reduce the number of parameters in the model in order to simplify according to the criterion of parsimony to avoid over-fitting. Two strategies are proposed in the literature to guide the model selection process, however there is no single strategy that applies in every situation.

The Top-Down strategy

This strategy of model selection was proposed by Verbeke [108] starting with a model that includes the maximum number of fixed effects arriving at a reduced model:

(a) Definition of the mean structure for the model. In this step, each potential effect is inserted in the model.

(b) Definition of the random-effect structure that can be tested by performing REML-based likelihood ratio test for the associated covariance parameters.

(c) Choice of the covariance structure for the residuals in the model to study the remaining variation in the observed response after the correct definition of the fixed and random effects.

(d) Semplification of the model testing whether some fixed effects can be deleted from the model.

The Step-Up strategy

An alternative to the previous one is given by Snijder and Bosker [94] and Raudenbush and Bryk [76].

(a) The initial model is formed by only the intercept expressing the mean structure, this step also includes random effects associated with the Level 2 units (cluster of longitudinal measurements) that allows the assessment of the variation in the outcome across longitudinal data set without adjusting for the effects of any covariates.

(b) Choice of the covariates and definition of their associated fixed effects for the single measurement (Level 1). In this step we can also include

adding random effects for Level 2 that can improving the fitting of the Level 1.

(c) Eventually iterative steps for the definition of fixed and random effects to the Levels after the Level 2.

1.5 Inference for the marginal model

The goal of the most part of the statistical analysis is to get from a sample an estimate of the effects that can be generalized to the entire population. Therefore most of the tests suggested below are aimed at the best choices for the marginal model and the study of the effects on the population averaged model.

1.5.1 Inference on the fixed effects

As shown in Section 1.4, the two strategies pass through a different definition of the population average model.
The fixed effects vector is defined as:

$$\hat{\beta}(\alpha) = \left(\Sigma_{i=1}^{N} X_i^T V_i^{-1} X_i)\right)^{-1} \Sigma_{i=1}^{N} X_i^T V_i^{-1} y_i$$

where $V_i = V_i(\alpha) = Z_i D Z_i^T + \Sigma_i$ is substituted by its REML estimate.
Under the marginal model $Y_i \sim N_{n_i}(X_i\beta, V_i(\alpha))$, and conditionally on α, $\hat{\beta}(\alpha)$ follows a multivariate normal distribution with mean vector β and with covariance matrix given by:

$$\begin{aligned}\mathbb{V}[\hat{\beta}] &= \left(\Sigma_{i=1}^{N} X_i^T V_i^{-1} X_i\right)^{-1} \left(\Sigma_{i=1}^{N} X_i^T V_i^{-1} \mathbb{V}[Y_i] V_i^{-1} X_i\right) \left(\Sigma_{i=1}^{N} X_i^T V_i^{-1} X_i\right)^{-1} \\ &= \left(\Sigma_{i=1}^{N} X_i^T V_i^{-1} X_i\right)^{-1}\end{aligned}$$

where the second equality holds if and only if the covariance matrix V_i is correctly specified and $V_i = \mathbb{V}[Y_i]$. The first two tests presented below are on the parameters β and are given by an approximate version of the Wald Test, t-Test and F-Test seen for the Generalized Linear Model. The last test for the fixed effects is for the comparison between two nested models and it is given by the likelihood ratio test.

Approximate test for β

In general, for any matrix L, we can perform a test with hypotheses:

$$\begin{cases} H_0 : L\beta = 0 \\ H_1 : L\beta \neq 0 \end{cases}$$

using three possible test.

- **Approximate Wald Test** An approximate Wald test (also called Z-Test) for each parameter of vector β is given by approximating the distribution of $Z = \frac{\hat{\beta}-\beta}{se(\beta)}$ with a standard univariate normal distribution. Given a test of hypothesis and matrix L as previous defined, we have:

$$W = Z^2 = (\hat{\beta}-\beta)^T L^T \left[L \left(\sum_{i=1}^{N} X_i^T V_i^{-1} X_i \right)^{-1} L^T \right] L(\hat{\beta}-\beta) \sim \chi_{rank(L)}$$

- **Approximate t-Test and F-Test** The estimates of the standard errors in the approximate Wald test underestimate the true variability in β because they do not take into account the variability introduced by estimating $V(\alpha)$.
  A possible way around this is given by an approximate $t-$ or $F-$ Test. In general, under H_0 the distribution of $\frac{\hat{\beta}_j-\beta_j}{se(\beta)_j}$ does not follow an exact t distribution, in fact the number of degrees of freedom of the test statistic is not equal to $N-p$ (where p is equal to the number of fixed-effect parameters in the model) but must be calculated using an appropriate approximation method. Analogously in the case of F statistic given by:

$$F = \frac{(\hat{\beta}-\beta)^T * L^T \left[L \left(\sum_{i=1}^{N} X_i^T V_i^{-1} X_i \right)^{-1} L^T \right]^{-1} L(\hat{\beta}-\beta)}{rank(L)}$$

  where the numerator degrees of freedom is equal to $rank(L)$ but the denominator degrees of freedom needs to be approximated with specific methods. Several methods are implemented to estimate the number of degrees of freedom taking into account the presence of random effect and correlations among the residuals such as the Satterthwaite method [85]. Neverthless, for large sample, different estimation methods do not lead to severe differences in the resulting $p-$values.

If we can preserve the validity of inference on β from possible misspecification of the covariance structure for the model, we can use the so-called *sandwich estimator* for $\mathbb{V}[\hat{\beta}]$ [51] obtained by replacing $\mathbb{V}[Y_i]$ by $(y_i - X_i\hat{\beta})(y_i - X_i\hat{\beta})^T$. The resulting estimator, also called robust or empirical variance estimator, is proven to be consistent, as long as the mean structure is correctly specified for the model. Thus, if the interest is in providing good inferential conclusions on the mean longitudinal response then it is worth devoting efforts and paying attention to modelling the covariance structure.

Likelihood Ratio Test

The likelihood ratio test is based on comparing the likelihood functions of two nested model, where a model A is nested in a full model B. LRT requires that both the nested model (under the null hypothesis of some parameters $\vartheta = \vartheta_0$) and full model corresponding to a specified hypothesis are fitted to the same subset of the data. The LRT statistic is calculated as shown above:

$$-2ln\left[\frac{l_{ML}(\hat{\vartheta}_0)}{l_{ML}(\hat{\vartheta})}\right] \sim \chi^2_{df}$$

where l_{ML} denote the maximum likelihood function and $\hat{\vartheta}_0$ and $\hat{\vartheta}$ the ML estimates obtained from maximizing l_{ML} under H_0 and H_a respectively. Under certain regularity conditions, follows asymptotically under H_0, a chi-squared distribution with *df* equal to the difference in number of parameters between the two models.
However this result is not valid if the models are fitted using REML rather than ML estimation ([59], [69], [108]) because both nested and full models must be fitted on the same subset of the data but in this case the mean structure $X_i\beta$ under H_0 is not the same of the full model and this leads to different error contrast $U = A'Y$. The strategy used to avoid this issue is to do a step in which we fit the marginal model under ML estimation to choose the best set of covariates. Then, the model with selected covariates is again fitted under REML estimation [111].

1.5.2 Inference for the Variance Components

From a practical point of view, although in most cases the focus is on the model for the average, an appropriate modeling for the variance/covariance matrix components is useful for interpreting the subject-specific random variation and essential to obtain a valid inference on the parameters of the model for the mean population effects. However, on the one hand, an excessive parametrization of the variance / covariance structure would lead to inefficient estimates and a potentially poor evaluation of the standard errors for the estimates of the fixed effects; on the other hand, an excessively restricted parametrization would risk invalidating the inference. A likelihood ratio test can help as detailed next.

Likelihood Ratio Test

As seen in the previous section, the maximum likelihood ratio test can be used to compare nested models, with different variance-covariance structures.

Unlike the hypothesis test on fixed effects, the likelihood ratio test for the α components of the variance/covariance matrix is valid both when using both ML and REML. It is true because the two compared models have the same model for the mean which leads to the same contrasts $U = A^T Y$ that are necessary in order to obtain two structurally comparable REML functions. However, the distribution of the test statistic under H_0 depends on whether the null hypothesis values for the covariance parameters lie on the boundary of the parameter space for the covariance parameters or not.
An other way is given by the possibility of using the Wald test in analogy to what we have seen for the inference on β.

1.5.3 Tests based on the Information Criteria

All the tests seen so far were concerned with the comparison of nested models, i.e. the model under H_0 could be seen as a particular alternative model or a reduced model compared to the one under H_1. Next, the comparison of two non-nested models is taken into account.
The information criteria provide a way to assess the fit of a model based on its optimum log-likelihood value $l(\hat{\vartheta})$, after applying a penalty for the number of parameters that are estimated in the model. In general, AIC (Akaike Information Criteria) and BIC (Bayes InformationCriteria) are the most used Information Criteria and are given by:

$$AIC = -2l(\hat{\vartheta}) + 2p$$

$$BIC = -2l(\hat{\vartheta}) + p \log N$$

The BIC applies a greater penalty for the models with more parameters than does the AIC, because it multiplies the number of parameters fitting with the natural logarithm of the number of total observations used. The choice of the model follows the rule "smaller is better".
Since the indexes based on the Information Criteria are based on the ML or REML estimates, the limitations set in the case of the Likelihood Ratio Test are valid and can therefore be used to make inference on the average model when using the ML method to estimate the parameters, while if we use REML we can apply them only when the compared models have the same structure for the model of the means and different structure relative to the variance/covariance matrix.

1.5.4 Inference for the need of a LMM

To ask whether it makes sense to use a linear mixed model (intercept and/or slopes) instead of a classical linear model where each single record (each

line of the dataset) is assumed independent of each other, is equivalent to doing a test on the nullity of the random effects. This is possible with a LRT that compares a model with random effects and a linear model without random effects, while holding constant the model for the means. In this case, testing the hypothesis that the variability of random effects is zero leads us to work on the boundary of the α parameters and this makes it necessary to make small changes to the asymptotic distribution of the LRT. In general, the asymptotic null distribution for the likelihood ratio test for nonstandard testing situations is often a weighted mixtures of chi-squared distributions rather than the single chi-squared distribution ([90], [97], [98]).

1.6 Prediction of random effects

In the Sections 1.5 we have seen that in many applications inference is focused on fixed effects (i.e. the overall change of the response over time) and on the variance components. However, we can also estimate the subject-specific deviation b_i from the mean trajectories over time and we will see from Chapter 3 that this is necessary in order to develop a Joint Model. This estimates are known as Empirical Bayes predictions (EB) [108].
From Section 1.2 we can observe that :

$$Y_i|b_i \sim N_{n_i}(X_i\beta + Z_ib_i, \Sigma_i)$$

with

$$b_i \sim N_q(0, D).$$

In a Bayesian approach the distribution of b_i is called the prior distributions since it does not depend on the data Y_i. Instead, the distribution of $b_i|Y_i = y_i$, where y_i is the observed response for the i-th subject, is called posterior distribution of b_i. Merging these considerations we have that:

$$f(b_i|Y_i = y_i) = \frac{f(y_i|b_i)f(b_i)}{f(y_i)} = \frac{f(y_i|b_i)f(b_i)}{\int f(y_i|b_i)f(b_i)db_i}.$$

Using the theory on general Bayesian linear models ([52], [93]), it can be shown that the previous result is the density of a multivariate normal distribution. Very often, b_i is estimated by the mean of this posterior distribution, called the posterior mean of b_i. This estimate is then given by:

$$\begin{aligned}\hat{b_i}(\vartheta) &= E[b_i|Y_i = y_i] \\ &= \int b_i f(b_i|y_i)db_i \\ &= DZ_i^T V_i^{-1}(\alpha)(y_i - X_i\beta)\end{aligned}$$

and the covariance matrix of the corresponding estimator equals:

$$\mathbb{V}[\hat{b_i}(\vartheta)] = DZ_i^T \left(V_i^{-1} - V_i^{-1} X_i \left(\sum_{i=1}^{N} X_i^T V_i^{-1} X_i \right)^{-1} X_i^T V_i^{-1} \right) Z_i D$$

where the unknown parameters β and α are replaced by their ML or REML estimates.
Note that $\mathbb{V}[\hat{b_i}(\vartheta)]$ underestimates the variability in $\hat{b_i}(\vartheta)] - b_i$ since it ignores the variation of b_i. Therefore, inference for b_i is usually based on

$$\mathbb{V}[\hat{b_i}(\vartheta) - b_i] = D - \mathbb{V}[\hat{b_i}(\vartheta)]$$

as an estimator for the variation in $\hat{b_i}(\vartheta)$ [48].
The resulting estimates for the random effects are called Empirical Bayes (EB) estimates, which we will denote as $\hat{b_i}$. The variability is underestimated in the obtained estimates since they do not take into account the variability introduced by replacing the unknown parameter ϑ by its ML or REML estimate.
Also in this case, inference is based on some test like approximate t-, F- or Wald tests with similar procedures for the estimation of the denominator degrees of freedom as seen in the Section 1.5.
After the estimation process, usually histograms or normal quantile plot of residuals are used to check the normality of the $\hat{b_i}$ and in order to detect potential outliers, i.e. subjects that seem to assume values for the outcome and have a longitudinal trajectory very different from the most part of the other subjects in the data. Observing the linear prediction $\hat{Y_i}$ we can see that:

$$\begin{aligned}
\hat{Y_i} &= X_i\hat{\beta} + Z_i\hat{b_i} \\
&= X_i\hat{\beta} + Z_i D Z_i^T V_i^{-1}(\alpha)(y_i - X_i\beta) \\
&= (I_{n_i} - Z_i D Z_i^T V_i^{-1}) X_i\hat{\beta} + Z_i D Z_i^T V_i^{-1} y_i \\
&= (I_{n_i} - (V_i - \Sigma_i) V_i^{-1}) X_i\hat{\beta} + (V_i - \Sigma_i) V_i^{-1} y_i \\
&= \Sigma_i V_i^{-1} X_i\hat{\beta} + (I_{n_i} - \Sigma_i V_i^{-1}) y_i
\end{aligned}$$

Interestingly, the predicted response trajectory for the i-th individual is a weighted combination of estimated population averaged mean response profile and the observed response profile for each subject. Since weight of the population averaged mean response profile is given by the relationship between the residual covariance matrix Σ_i and the overall covariance matrix V_i, then much weight will be given to the overall average profile (i.e. $X_i\hat{\beta}$) if the

residual variability is large in comparison to the between-subject variability (modelled by the random effects). Whereas more weight will be given to the observed data (i.e. y_i) if the opposite is true. The shrinkage factor is desirable because it only affects individuals that provide little information (small number of repeated measures), borrowing strengths from other clusters.

1.6.1 The Mixed models, a midway between a frequentist and a Bayesian approach

The estimation of the random effects presented in Section 1.6. is based on the posterior distribution of the b_i's. This is necessary to include in the model a latent source of variability imputed to each patient which is intrinsically hidden and hard to understand what to measure. In this way, we are not assuming that the subject variable is a possible confounder in the relationship among the covariates and the response variable. In this way, the Mixed Model can be seen as lying midway between a completely frequentist approach (Linear Model) and a completely Bayesian one.
In the first case, the insertion of the subject IDs as a categorical covariate in the model would have two effects on the estimation of the other parameters related to the covariates of interest: a desired effect is to capture the correlation between repeated measurements on the same subject over time, an another possible effect is to use this variable as a control of confounding, imputing to the subject a fixed effect even if it is not clear the clinical meaning of this effect. Moreover, a big limit of this approach is the presence of a huge number of levels in the ID variable that can be difficult to manage in the implementation of the model itself.
In the second case, if we use a completely Bayesian approach, we should provide a prior distribution for each effect inserted in the model, not only for the subject component, but also for each covariate inserted in the fixed part of the model. I do not dwell on the theoretical implications. I just recall that, in the clinical research, the use of the Bayesian model is largely debated because there is a high risk of introducing subjective elements in the definition of the prior distribution. This could lead to a too random scenario not as solid, robust and clear as the one based only on observed data provided by the selected sample of subjects.

1.7 Diagnostic

The Joint Model that will be presented from Chapter 3 will be made from 2 different sub-models. One of them, regarding the longitudinal data, will be

manage with the LMM both for estimation and diagnostic. So it becomes necessary to explain which is the most common methods to assess the goodness of the model use to manage the longitudinal trajectories.
In contrast with the diagnostic methods for the standard linear models which are well established in the literature, the diagnostic for the Linear Mixed Model is more difficult to carry out and interpret due to the complexity of the model itself that contains also the random effects and several covariance structures. There are different approaches to test the model chosen, more or less heuristic. In particular, we can split the the diagnostic procedures in three parts regarding residuals, influence and random effects respectively. In the following sessions a hint of diagnostic techniques is presented, however, for a more in-depth view refer to the main textbooks on the matter (see [108] and [111]).

1.7.1 Residual Diagnostic

Among the informal techniques, the residual diagnostic is commonly used to check whether or not a specific pattern exists in the residuals. In the context of the standard linear model, plotting residuals against fitted values is used to verify model assumption (e.g. normality, constant variance) and to detect outliers and potentially influential observations.
In the context of the Linear Mixed Model this diagnostic it carry out referring to the marginal residuals and conditional residuals. A marginal residuals ε_{i_m} is the difference between the observed data and the estimated (marginal) mean:

$$\varepsilon_{i_m} = y_i - X_i\hat{\beta}.$$

The conditional residuals are given by the difference between the observed values and the predicted values of the outcome variables for each observation:

$$\varepsilon_i = y_i - (X_i\hat{\beta} + Z_i\hat{b_i})$$

where the name conditional residuals stems from the fact that the quantity $X_i\hat{\beta} + Z_i\hat{b_i}$ is the conditional mean of the vector y_i. Residuals in the form of ε_{i_m} or ε_i are defined as raw residuals and they are usually not used for the diagnostic purpose because even if the true model errors are uncorrelated and have equal variance, conditional residuals will tend to be correlated and their variances may be different for different subgroups of individuals. More often, various standardizations (by scaling for the standard deviation) are applied to overcome this problem. As happens for the linear model, one type of residuals is given by the studentized residuals where the unknown standard deviation is replaced by an its estimate. If this estimate is independent from

the $i-$th observation, the process is termed *external* studentization. This is usually accomplished by excluding the $i-$th observation when computing the estimate of its standard error. If the observation contributes to the standard error computation, the residual is said to be *internally* studentized. So, the marginal studentized residuals are given by

$$\varepsilon_{i_m}^{stud} = \frac{\varepsilon_{i_m}}{\sqrt{\hat{\mathbb{V}}[\varepsilon_{i_m}]}}$$

while the studentized conditional residuals are given by:

$$\varepsilon_{i}^{stud} = \frac{\varepsilon_{i}}{\sqrt{\hat{\mathbb{V}}[\varepsilon_{i}]}}.$$

Other possible scaling choices are possible but they are not relevant to the aim of this context.

1.7.2 Influence Diagnostic

The influence diagnostic regards several formal techniques that allow one to identify unusual observations that may heavily influence the ML estimates of the parameters. The idea of influence diagnostic for a given observation (or subset of observations) is to quantify the effect of omission of those records from the data on the results of the analysis of the entire data set. This diagnostic may be used to investigate various aspects of the model fit: some examples are given by the fixed effects, covariance parameters, precision and predicted values. Usually, a sensitivity analysis is done to assess if there are significant changes in the estimation of the parameters when some observations are excluded from the analysis. More details on influence diagnostic can be found in [108].

1.7.3 Random-effect Diagnostic

The empirical Bayes predictors seen in the paragraph 1.6 is the natural choice to do the diagnostic on the random effects. Tools to diagnose these estimates are usually given by diagnostic plots like histograms, Q-Q plots and scatterplots in order to investigate for potential outliers. Conversely to the diagnosis on the fixed effects, checking for normality is of little importance, because the distribution of the random effects does not necessarily reflect the true distribution of the random effects.

1.8 The problem of Missing Data

The problem of missing data is a major challenge for the analysis of the longitudinal trajectories [26]. Clinical studies are designed to collect data of the patients at specified follow-up visits, however subjects miss some of their planned measurements for several reasons. This involves at least two issues:

- in a clinical trials, where the sample size is defined a priori by setting a threshold for the first type error and the power, the missing data could compromise the study because the reduction in the number of repeated measures would result in a loss of statistical power;
- more generally, the missing data could produce biased estimates requiring a greater caution in interpreting the results themselves.

It becomes necessary to understand why some data is missing. This problem is less relevant when the missingness is at random but it is crucial when there are some unexplained factors which make some subjects more likely to skip a visit. Joint Models are used to model the relationship between a longitudinal evolution of a marker and the time to an event of particular interest. They also can be used to address problems related to the presence of missing values in the longitudinal data. In the next pages, different patterns of missingness are presented to provide a better overview while in the Chapter 3 it will shown how the use of the Joint Models could help with missing data.

Let the following table be an extraction from a dataset containing the longitudinal values collected across 5 subsequent visits:

Subject ID	Follow-up visits				
	1	2	3	4	5
1	x	x	x	x	x
2	x	x	x	?	?
3	?	x	x	x	x
4	x	?	x	?	x

Not all subjects have values for all 5 visits but different types of missingness can be noticed. The first difference is between *monotone* and *non-monotone* missingness. In the case of patients 2 and 3 we talk of monotone pattern of missingness, in particular the patient 2 is *dropout* (i.e. he is withdrawn from the study before it is finished) while patient 3 is a case of *late entry* (i.e. the subject does not provide the first measurement). In all other cases we talk about non-monotone missingness, also called *intermittent* as in the case of patient 4. The missingness involves three problems:

- loss of efficiency: we collect less data than originally planned and therefore the changes in the average longitudinal trajectory are less precisely estimated. This fact involves the sample size also, because we need to enroll more patients to achieve the same levels of power.

- unbalanced dataset: not all subjects have the same number of repeated measurements and this creates complications for methods of analysis that require balanced data (it is not the case of the linear mixed models);

- potential bias: missing data may depend on outcome of interest and introduce bias and thereby lead to misleading inferences.

In general, we assume that the study is designed so that measurements of patient i are collected n_i times. For $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$ we define the missing data indicator r_{ij} as

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

Defining the response vector $y_i = (y_{i1}, \ldots, y_{in_i})^T$, we now obtain a partition of the complete response vector into two subvectors: y_i^o containing the observed data, i.e. the observed values y_{ij} for which $r_{ij} = 1$, and y_i^m corresponding to the missing data, i.e. that y_{ij} which are set equal to a conventional values, e.g. NA, whenever $r_{ij} = 0$. For the remaining of this work, we will focus on dropout, in this case the missing data indicator for the i-th subject is of the form $(1, \ldots, 1, 0, \ldots, 0)$ and therefore can be simplified and replaced by

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

if time is on a discrete scale. To describe the probabilistic relation between the measurement and missingness processes, Rubin [84] has introduced three mechanism thinking the missing data mechanism as a probability model based on the conditional density of the missingness process r_i given the complete response vector $y_i = (y_i^o, y_i^m)$:

$$p(r_i | y_i^o, y_i^m; \vartheta_r)$$

where $\vartheta_r = (\beta_r, \alpha_r)$ denotes the corresponding full parameter vector (see Theorem 1). Below are presented briefly the three types of mechanisms.

MCAR - Missing Completely At Random

The probability that responses are missing is unrelated to both y_i^o and y_i^m:

$$p(r_i|y_i^o, y_i^m; \vartheta_r) = p(r_i; \vartheta_r).$$

An example of MCAR is encountered when subjects leave the study after at least a pre-determined number of measurements has been taken or when the measurements are lost due to a problem. In this case the observed data y_i^o can be considered a random sample of the complete data y_i. Under MCAR we can obtain valid inferences using any valid statistical procedure for complete data, while ignoring the process that has generate the missing values.

MAR - Missing At Random

The probability that responses are missing is related to y_i^o, but is unrelated to y_i^m:

$$p(r_i|y_i^o, y_i^m; \vartheta_r) = p(r_i|y_i^o; \vartheta_r).$$

In other words the probability of missingness depends on the observed values, but is unrelated to the outcomes that should have been obtained. An example of MAR occurs when study protocol requires that patients be removed from the study when response value exceeds a prespecified threshold. In this case the missingness process is under the control of the investigator and is related to the observed values y_i^o only and for this the observed data cannot be considered a random sample from the target population because the distribution of y_i^o does not coincide with the distribution of y_i. The most important consequence of this fact is that the sample moments are not unbiased estimates of the same moments in the target population, thus statistics based on these moments without accounting for MAR may be misleading (e.g. sample marginal evolution instead of sample subject-specific evolutions). On the other hand, under MAR, likelihood based inference continues to be correct ignoring the contributions of r_i because the likelihood contribution of the complete data (y_i^o, y_i^m, r_i) for the i-th subject is factorized as follows:

$$L_i(\vartheta) = L_i(\vartheta_y) \times L_i(\vartheta_r)$$

where ϑ_y and ϑ_r are disjoint and inference for ϑ_y can be based on the marginal observed data density $p(y_i^o; \vartheta_y)$ ignoring the likelihood of the missingness process. This property is also known as *ignorability*. Moreover, mixed models with misspecified correlation structure or marginal residuals are not valid under MAR, while a correct specification of the correlation structure for the mixed model or subject-specific residuals are valid.

MNAR - Missing Not At Random

The probability that responses are missing is related to y_i^m, and possibly also to y_i^o:

$$p(r_i|y_i^m;\vartheta_r) \text{ or } p(r_i|y_i^o,y_i^m;\vartheta_r).$$

An example of MNAR occurs in a study of drug addicts where people who return to drugs are less likely than others to report their status. Also in this case the observed data cannot be considered a random sample from the target population. The predictive distribution of y_i^m conditional on y_i^o is not the same ad in the target population, but rather depends on both y_i^o and on $p(r_i|y_i)$. For this fact, only procedures that explicit model the joint distribution $\{y_i^o, y_i^m, r_i\}$ provide valid inference. Note that the type of the missingness mechanism may depend on covariates: if missingness is related to a covariate but not to y_i (i.e. missingness mechanism is MCAR), and in our analysis of the longitudinal trajectories we do not condition on this covariate, then MCAR can no longer be considered valid. We cannot tell from the data at hand whether the missing data mechanism is MAR or MNAR, however we can distinguish between MCAR and MAR.

1.9 Summary of chapter

This chapter presents a review on the use of the Linear Mixed Models to analyse longitudinal data. It constitutes one of the two fundamental parts of the Joint Model. Particular attention is given to practical aspects such as the interpretation of the coefficients and the procedures used to select the best model. An overview of the problems related to the presence of non-random missing data is given in order to better understand how the use of the Joint Models may solve it.

Chapter 2

Survival Analysis

In follow-up studies different types of outcomes are often collected, among them there are the multiple longitudinal responses, that we have seen in the previous chapter, and time-to-event(s) of particular interest (e.g. death, relapse, hospitalization, ...) that will be the focus of this chapter.
Survival analysis includes techniques and models for the study of the time between a clinically relevant starting point, called baseline, and the occurrence of an event of interest. In this contest the outcome is usually defined as failure time, survival time or event time.

2.1 Distribution of the failure times

When it comes to the statistical analysis of failure times, usually denoted with T (continuous or discrete), the first feature that must be taken into account is the shape of their distribution. Event times must be positive and they very often have skewed shapes of distribution. Thus, statistical methods that rely on normality are not directly applicable, and, if used for survival data, may produce invalid results. This often could be easily overcome using a suitable transformation of the event times, such as the logarithm or the square root. We introduce the following basic definitions.

Definition 3. *Let T be a continuous r.v. defined for $t \in [0, +\infty)$ with cumulative distribution function $F(\cdot)$ and probability function $f(\cdot)$. Its survival function is defined as*

$$S(t) = 1 - F(t) = \mathbb{P}[T > t] = \int_t^{+\infty} f(u)du$$

Note that $f(t) \geq 0$ and $\int_0^{+\infty} f(t)dt = 1$ and it gives the density of prob-

ability at time t and for an very small *epsilon* > 0 we have

$$f(t)\epsilon \approx \mathbb{P}[t \leq T < t + \epsilon] = F(t + \epsilon) - F(t)$$

Theoretically, as t ranges from 0 up to infinity, the survival function can be graphed as a smooth curve. The survival function is non-increasing, at the start of the study (i.e. $t = 0$) $S(0) = 1$ and when time tends to infinity and the event of interest is dead, the survival tends to 0.
Another important function necessary to derive the probability density function is the *hazard function*.

Definition 4. *Let T be a continuous r.v. defined for $t \in [0, +\infty)$ with cumulative distribution function $F(\cdot)$ and probability function $f(\cdot)$. The hazard function is defined as*

$$h(t) = \lim_{\epsilon \to 0^+} \frac{\mathbb{P}[t \leq T < t + \epsilon | T \geq t]}{\epsilon}$$

if the limit exists, ϵ is a very small positive quantity.

The hazard function gives the instantaneous potential risk per unit time for the event to occur in the time interval $[t, t + \epsilon)$ given that the individual has survived up to time t.
From the Definition 4 it follows that

$$h(t) = \lim_{\epsilon \to 0^+} \frac{F(t + \epsilon) - F(t)}{\epsilon\,(1 - F(t))} = \frac{f(t)}{S(t)}$$

In particular, the hazard function is always non-negative and it has no upper bound.
The survival function $S(\cdot)$ is more used for analysis than $h(\cdot)$ because it directly describes the survival experience of a study sample. However, the hazard function is also of interest and it may be used to identify a specific parametric model form, such as an exponential, a Weibull, or a log-normal curve. Anyway, the two function can be linked as following. By observing that $dS(t)/dt = -f(t)$, the hazard function can also be also rewritten as

$$h(t) = \frac{f(t)}{S(t)} = -\frac{dS(t)/dt}{S(t)} = -\frac{d}{dt}[\log S(t)]$$

If the above equation is integrated from $t = 0$ to t and assuming the boundary condition for which $S(0) = 1$ (since the event is sure not to have occurred at the baseline), we obtain

$$S(t) = \exp[-\int_0^t h(u)\,du]$$

obtaining a formula for the probability of surviving up to t as a function of the hazard at all durations up to t.
Furthermore, the integral in brackets can be interpreted as the sum of the risks you face going from time 0 to a time t in the future. Formally, it is defined as *cumulative hazard function*.

Definition 5. *Let be $h(\cdot)$ an hazard function and $t > 0$*

$$\mathcal{H}(t) = \int_0^t h(u)\, du$$

is defined cumulative hazard function.

Finally, the probability density function $f(\cdot)$ can be obtained by using the Definition 4 and 5, as

$$f(t) = h(t) \exp[-\mathcal{H}(t)].$$

2.2 Estimating the survival function

In the previous section, the general formulation for the survival, hazard and cumulative hazard functions was provided related to a random variable T that models the failure times. No considerations were done about the reason for which a patient experiences an event that interrupts the observation period. It becomes necessary to introduce the concept of *censoring* that is the main characteristic that distinguishes survival analysis. Censoring occurs whenever one of the following conditions happens and the event time of interest is not fully observed on all subjects under study:

- the study ends, but some patients still have not had the event yet (administrative censoring);
- some individuals drop out or get lost during the study, and all we know about them is the last time they were still free of the event;
- some individuals develop competing events.

In all these cases we talk about right censoring. When a subject withdraws from the study for reasons directly related to the expected failure time, for example, because of a worsening of her prognosis, the censoring mechanism is called *informative* and, unfortunately, very few things can be done to complete his follow-up.
When a subject withdraws from the study for reasons not related to her

prognosis, but for example depending on covariates, the censoring is called *non-informative* or *random* and it is in this case that the theory following explained is developed (see Definition 6).
The main implications of censoring are that the standard tools, such as the sample average, the t-test, and linear regression cannot be used, moreover inferences may be sensitive to misspecification of the distribution of the event times.
When censoring occurs, the outcome can be thought of as comprising two dimensions: an event indicator and a time at risk. With a little variation from the previous section, for each subject $i = 1, \dots, N$ let T_i^* (and not T_i) denote the random variable of the failure times under study and let C_i be a non-negative variable which models the censoring times, then only the random variable $T_i = min\{T_i^*, C_i\}$ is observed due to censoring. In addition to observing T_i we also get to see the event indicator $\delta_i = \mathbb{I}(T_i^* \leq C_i)$. Furthermore, note that under non informative censoring T_i^* and C_i are independent for each i, furthermore $T_1^*, \dots, T_n^*$ are independent and identically distributed (iid), and also $C_1, \dots, C_n$ are assumed to be iid. In general, in survival analysis, we are interested in estimating characteristics of the distribution of T^* using only the available information T_i and δ_i for each $i = 1, \dots, N$.
The survival function can be estimated in two ways:

- by developing an empirical estimate of the survival function, i.e. a non-parametric estimation;
- by specifying a parametric model for $\lambda(t)$ on a particular density function $f(t)$.

If no censoring occurs, an empirical estimator of the survival function is

$$\hat{S}(t) = \frac{\Sigma_{i=1}^{N} \mathbb{I}(T_i > t)}{N} = \frac{\#\ individuals\ with\ T > t}{total\ sample\ size}$$

Under the assumption of independent identically distributed sample, it holds

$$n\,\hat{S}(t) \sim Binomial(n, S(t))$$

and for large sample sizes, by the central limit theorem

$$\hat{S}(t) \simeq \mathcal{N}\left(S(t), \frac{S(t)(1 - S(t))}{n}\right)$$

where $\sim$ means that the random variable on the left side follows the distribution in the right side, and *simeq* that the distribution of the random variable on the left is approximated by the distribution on the right. Otherwise, if there are censored observations, $\hat{S}(t)$ is not a good estimate of the true $S(t)$ and other non-parametric methods must be used to account for censoring process, the most used is described in the Section 2.2.1.

2.2.1 Non-parametric estimation

The most well-known estimator of $S(t)$ when censoring occurs has been proposed by Kaplan and Meier [45] and it is also called the *product limit* estimator. This is a non-parametric estimator that does not make any assumptions on the underlying distribution of the failure times. This estimator is based on the cumulative distribution function $F(\cdot)$ calculated on the observed data:

$$\hat{F}(t) = \frac{\#\text{ of individuals who experienced the event up } t}{\text{total sample size}}$$

This function, that assumes the name of *empirical distribution function* is a useful and simply way to summarize and display survival data. Its plot versus the time t provides full information on the percentiles and the dispersion of T (or of the data which we assume independent instances of T), moreover it is an aid in studying the shape of the distribution necessary in constructing formal tests of hypotheses.
Another way is given by its complementary survival function $\hat{S}(t)$ also called *empirical survivor function*:

$$\hat{S}(t) = 1 - \hat{F}(t).$$

Both $\hat{F}$ and $\hat{S}$ do non take into account censoring. The Kaplan and Meier estimator can account for censoring. As stated in the previous Section, the main problem in using these functions is related to the presence of censoring.

The method proposed by Kaplan and Meier [45] is based on the conditional probability. Suppose:

- $t_0 \leq t_1 \leq \ldots \leq t_j \leq \ldots \leq t_k \leq t < t_{k+1}$ are different failure times in a sample size of N individuals and $t_{k+1} = +\infty$
- d_j is the number of subjects who experience the event at time t_j, $j = 0, \ldots, k$
- m_j is the number of censored subjects in the interval $[t_j, t_{j+1})$
- $r_j = (d_j + m_j) + \ldots + (d_k + m_k)$ is the number of subject at risk at a time just prior to t_j

The probability of failure at t_j given that you are at risk before t_j is

$$\mathbb{P}[T^* = t_j | T^* > t_{j-1}] = F(t_j) - F(t_j^-) = \frac{d_j}{r_j}$$

and the Kaplan-Meier estimator of the survival probability beyond t is given by

$$\begin{aligned}\hat{S}_{KM}(t_k) &= \mathbb{P}(T^* > t_k) \\ &= \mathbb{P}(T^* > t_k \cap T^* > t_{k-1} \cap ... \cap T^* > t_1 \cap T^* > t_0) \\ &= \mathbb{P}(T^* > t_1) \cdot \prod_{j=2}^{k} \mathbb{P}(T^* > t_j | T^* > t_{j-1}) \\ &= \prod_{j=1}^{k} [1 - \mathbb{P}(T^* = t_j | T^* > t_{j-1})] \\ &= \prod_{j=1}^{k} \left(1 - \frac{d_j}{r_j}\right).\end{aligned}$$

It has been proven that the Kaplan-Meier estimator, also in presence of censoring, is consistent and asymptotically normal [13], and it is normally distributed when no censoring occurs with distribution

$$S_{KM}(t) \sim \mathcal{N}\left(S(t), \frac{S(t)(1 - S(t))}{N}\right).$$

Moreover, it is shown that the KM estimator is also a non-paramteric maximum likelihood estimator [23].
Regardin the variance of $\hat{S}_{KM}(t)$, it can be calculated using Greenwood's formula ([33], [44]). To obtain the large sample variance of the KM estimator, we apply the delta method twice and use the logarithm of the survivor function. The KM estimator can be rewritten as:

$$\hat{S}_{KM}(t) = \prod_{t_j < t} (1 - \lambda_j)$$

where $\lambda_j = d_j / r_j$. Since λ_j is just a binomial proportions given the number of subjects at risk r_j, we can observe that

$$\mathbb{V}[\lambda_j] \approx \frac{\lambda_j(1 - \lambda_j)}{r_j}$$

Since $\hat{S}_{KM}(t)$ is a function of λ_j, an estimator of its variance can be found

using the Delta method twice on the logarithm of the $\hat{S}_{KM}(t)$:

$$\begin{aligned}\hat{\mathbb{V}}[\log(\hat{S}_{KM}(t))] &= \sum_{j:t_j \leq t} \mathbb{V}[\log(1-\lambda_j)] \\ &= \sum_{j:t_j \leq t} \left(\frac{1}{1-\lambda_j}\right)^2 \mathbb{V}[\lambda_j] \\ &= \sum_{j:t_j \leq t} \left(\frac{1}{1-\lambda_j}\right)^2 \frac{\lambda_j(1-\lambda_j)}{r_j} \\ &= \sum_{j:t_j \leq t} \frac{\lambda_j}{(1-\lambda_j)r_j} \\ &= \sum_{j:t_j \leq t} \frac{d_j}{(r_j-d_j)r_j}\end{aligned}$$

Now, considering $\hat{S}_{KM}(t) = \exp(\log(\hat{S}_{KM}(t)$ and using the delta method again, we obtain (see [33]):

$$\begin{aligned}\hat{\mathbb{V}}[\hat{S}_{KM}(t)] &= [\hat{S}_{KM}(t)]^2 \cdot \hat{\mathbb{V}}[\log(\hat{S}_{KM}(t))] \\ &= [\hat{S}_{KM}(t)]^2 \cdot \sum_{j:t_j \leq t} \frac{d_j}{(r_j-d_j)r_j}.\end{aligned}$$

2.2.2 Parametric estimation

We consider the estimation of the survival data when one is willing to assume a parametric form for the distribution of survival time T^*. We can draft into service distributions such that for $Y \in \mathbb{R}$ by considering $T^* = e^Y$, so that $Y = \log T^*$ represents the log failure time. The exponential and the Weibull distributions are used largely but other distributions can be used.

Exponential distribution

The Exponential distribution is used when the hazard function $h(t)$ is constant at $h > 0$. The instantaneous failure rate is independent of t, so that the conditional chance of failure in a time interval of specified length is the same regardless of how long the individual has been on study; this is referred to as the *memoryless property* of the exponential distribution. It holds:

- the density $f(t) = he^{-ht}$;
- the survivor function is $S(t) = \int_t^{+\infty} f(u)du = e^{-ht}$;
- the cumulative hazard function is given by $\mathcal{H}(t) = \int_0^t h(u)du = ht$.

Weibull distribution

The Weibull distribution is a two-parameter distribution and it is an important generalization of the exponential distribution because allows for a power dependence of the hazard on time. The hazard function is given by:

$$h(t) = h\gamma(ht)^{\gamma-1}$$

for $\lambda, \gamma > 0$. This function is decreasing for $\gamma < 1$, increasing for $\gamma > 1$ and reduces to the constant hazard (i.e. the Exponential distribution) if $\gamma = 1$.

- the density is $f(t) = -\frac{d}{dy}S(t) = \gamma h t^{\gamma-1} e^{-ht^{\gamma}}$;
- the survivor function is $S(t) = e^{-ht^{\gamma}}$;
- the cumulative hazard is given by $\mathcal{H}(t) = ht^{\gamma}$.

h is the scale parameter while γ is the shape parameter. The different hazard shapes make the Weibull distribution more convenient.

2.3 Likelihood function for censored data

When the survival function $S(\cdot)$ is assumed to be of a specific parametric form, estimation of the parameters of interest is often based on the maximum likelihood method.
Let T^* be a continuous random variable on $[0, +\infty)$ with cumulative distribution function $F(\cdot)$. Assume that $F(\cdot)$ depends on a parameter ϑ belonging to some sample space. Let C be a censoring random variable with cumulative distribution function $G(\cdot)$. Furthermore, for $i = 1, \ldots, N$, assume:

- $T_1^*, \ldots, T_N^*$ independent copies of T^*, so that $F(t) = \mathbb{P}[T_i^* \leq t]$
- $C_1, \ldots, C_N$ independent copies of C
- $T_i = min\{T_i^*, C_i\}$
- $\delta_i = \mathbb{I}_{(T_i^* \leq C_i)}$

Definition 6. *A censoring mechanism is said to be non-informative or random if*

$$\mathbb{P}(T_i^* > t | C_i = t) = \mathbb{P}(T_i^* > t)$$

for each $t > 0$ *and* $i = 1, \ldots, N$.

Now, consider a patient i with complete observation at time t_i,

$$\{T_i = t_i, \delta_i = 1\} = \{T_i^* = t_i, C_i > t_i\}$$

his contribution to the likelihood is given by:

$$\begin{aligned}
&\lim_{h\to 0} \frac{\mathbb{P}(t_i \leq T_i^* < t_i + h, C_i > t_i)}{h} \\
=&\lim_{h\to 0} \frac{\mathbb{P}(t_i \leq T_i^* < t_i + h)\mathbb{P}(C_i > t_i)}{h} \\
=&f(t_i)(1 - G_i(t_i)).
\end{aligned}$$

Similarly, if the patient i is censored at t_i, his contribution to the likelihood function is

$$g(t_i)(1 - F_i(t_i)).$$

The contribution of the patient i to the likelihood is given by:

$$L_i(\vartheta) = \left[f(t_i)(1 - G_i(t_i))\right]^{\delta_i} \left[g(t_i)(1 - F_i(t_i))\right]^{1-\delta_i}.$$

The overall likelihood is:

$$\begin{aligned}
L(\vartheta) &= \prod_{i=1}^{n} L_i(\vartheta) \\
&= \prod_{i=1}^{n} f(t_i)^{\delta_i}(1 - G_i(t_i))^{\delta_i} g(t_i)^{1-\delta_i}(1 - F_i(t_i))^{1-\delta_i} \\
&= \prod_{i=1}^{n} f(t_i)^{\delta_i}(1 - F_i(t_i))^{1-\delta_i} \times k \\
&= \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \times k
\end{aligned}$$

where k is a multiplicative constant that can be ignored because depends on G which does not depend on the parameter θ of interest.
Using the relation between $S(\cdot)$ and $h(\cdot)$ in the log-scale we have that the log-likelihood for the censored data is given by:

$$\log L(\vartheta) = \sum_{i=1}^{n} \left(\delta_i \log h_i(t_i; \vartheta) - \int_0^{t_i} h_i(s; \vartheta)\, ds \right).$$

All subjects contribute to the log-likelihood through the cumulative hazard function evaluated at the corresponding observed event time t_i. The subjects

who experienced the event additionally contribute an amount equal to the log hazard function evaluated at t_i. Once the log-likelihood has been formulated, iterative optimization procedures (e.g. Newton-Raphson algorithm) can be utilized to give the maximum likelihood estimates $\hat{\vartheta}$. Inference then proceeds under the classical asymptotic maximum likelihood theory paradigm.

2.4 Failure Time Models

We have seen several survival distributions for modelling the survival experience of a population. However, the interest is usually on evaluating whether and how failure time may depend on different explanatory variables. It therefore becomes of interest generalizing models to take into account information on the patients. Considering a failure time $T > 0$ and supposing that a set of covariates X is available for each patients, at baseline ($t = 0$) or relatively little time before (qualitative and/or quantitative variables, e.g information on treatment, biomarkers, age, and so on), we want to model the failure time depending on X. The first aim is to evaluate the effect of some covariates on T, but we include also covariates to account for heterogeneity among the individuals.

2.4.1 Parametric regression model

The Exponential distribution can be generalized to obtain a regression model where the failure rate is a function of a set of covariates X. If the failure rate is assumed to be constant over time and depending on the covariates X, then the hazard function at time t for an individual with covariate X can be written as

$$h(t|X) = h_0(X).$$

The $h(\cdot)$ function may be modelled through a linear combination $\beta' X$:

$$h(t|X) = h_0 c(\beta' X)$$

where the vector β is the set of regression parameters that quantifies the effect of each on the hazard, h_0 here is a constant and c is a specified functional form. There are different specific forms for c, and the most used is the form $c(s) = \exp(s)$ for which the hazard function assumes the form:

$$h(t|X) = h_0 \exp(\beta' X)$$

and the conditional density function of T given $X = x$ becomes:

$$f(t|x) = h \exp(\beta' x) \exp[-ht \exp(\beta' x)].$$

Analogously, hypothesizing a Weibull distribution for the hazard function, we have that the conditional hazard given X:

$$h(t|X) = \gamma(h_0 t)^{\gamma-1} \exp(\beta' X)$$

and the conditional density is:

$$f(t|X) = h_0 \gamma(h_0 t)^{\gamma-1} \exp(\beta' X) \exp[-(h_0 t)^{\gamma} \exp(\beta' X)].$$

The forms of the previous regression models suggest two distinct generalizations. First, the effect of the covariates is multiplicatively on the hazard function and this relationship suggests a more general model called the *Relative Risk Model* or *Cox Model*. Second, both of these models are log-linear models because the covariates have an addictive effect on $Y = \log(T)$ and we can obtain a more general class of log-linear models called the *Accelerated Failure Time Models* (AFT). Next we describe briefly the Cox model to be used in Chapter 3 and following, while the AFT, which is typically used when it is assumed that the effect of the covariates is to accelerate or decelerate the life course of the disease, is not further considered in this thesis.

2.4.2 Relative risk or Cox Model

Let $h(t|X)$ represent the hazard function at time t for an individual with covariates X collected at baseline (i.e. at $t = 0$). The relative risk model [21] assumes that covariates have a multiplicative effect on the hazard for an event, and it is defined as:

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

$$= h_0(t) \exp(\beta' X)$$

where $h_0(t)$ is an arbitrary unspecified *baseline hazard* function and corresponds to the hazard function for a subject for whom $\beta' X = 0$. It's obvious that if $h_0(t) = h_0$ the model reduces to the exponential regression model, while if $h_0(t) = h_0 \gamma(h_0 t)^{\gamma-1}$ then the model is a Weibull model. The conditional density function of T given X becomes:

$$f(t|X) = h_0(t) \exp(\beta' X) \exp\left[- \exp(\beta' X) \int_0^t h_0(u) du \right].$$

The estimation of all parameters in the model, that are the regression coefficients β and the parameters in the specification of the $h_0(t)$, proceeds

by maximizing the corresponding log-likelihood function. However, Cox [21] showed that the estimation of the regression coefficients β can be based on the partial log-likelihood function

$$p\ell = \sum_{i=1}^{n} \delta_i \left[\beta' X_i - \log \left(\sum_{T_j \geq T_i} \exp(\beta' X_j) \right) \right]$$

that does not require specification of $h_0(\cdot)$, that is, without having to specify the distribution of T_i^*. The relative risk model obtained without a specific baseline function is a semi-parametric model because it does not make any assumption for the distribution of the event times, but assumes that the covariates have a multiplicative effect on the hazard rate. The maximum partial likelihood estimators are found by solving the partial log-likelihood score equations:

$$\frac{\partial p\ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left(X_i - \frac{\sum_{T_j \geq T_i} X_j \exp(\beta' X_j)}{\sum_{T_j \geq T_i} \exp(\beta' X_j)} \right) = 0.$$

The solution $\hat{\beta}$ is consistent and asymptotically normally distributed with mean β^0 (i.e. the true values for the parameters in the vector) and variance $[\mathbb{E}[\mathcal{I}(\beta^0)]]^{-1}$ (i.e. the inverse of the expected information matrix). Due to the fact that the computation of the expectation requires knowledge of the censoring distribution, standard error are typically estimated using the observed information $[\mathcal{I}(\hat{\beta})]^{-1}$, where

$$\mathcal{I}(\hat{\beta}) = -\sum_{i=1}^{n} \frac{\partial^2 p\ell_i(\beta)}{\partial \beta' \partial \beta} \bigg|_{\beta=\hat{\beta}}.$$

Extended model

The Cox model is very flexible for the fact that $h_0(\cdot)$ is arbitrary and there are two important generalizations. First, the baseline hazard function $h_0(t)$ can be allowed to vary in specific subset of the data. Suppose that the population is divided into r strata and that the hazard is not proportional among strata, we can consider different hazard function for each stratum $j = 1, \ldots, r$. For each stratum j, the Cox model can be written as

$$h_j(t|X) = h_0 \exp(\beta' X_j)$$

where the baseline hazard function h_{0j} can vary for each $j = 1, \ldots, r$. It is useful in case that some explanatory variable does not appear to have multiplicative effect on the hazard. In fact, we can divide the range of such variable into strata with only the remaining regression covariates contributing to the exponential factor in the model.
The second generalization, treated in the next section, involves the time-dependent covariates.

2.5 Survival analysis with time-varying covariates

So far we assumed that the hazard function $h(\cdot)$ depends only on covariates measured at baseline and we assume that their values are constant during the follow-up, however in many clinical studies the interest is to investigate whether the change of the covariates (e.g. biomarkers) are related with the hazard. These covariates, that can change over time, are called *time-dependent covariates* and could be either *external* (also called *exogenous*) or *internal* (also called *endogenous*) covariates. We need to set up some notations: let $y_i(t)$ denote the covariate vector at time t for the i-th subject, and $\mathcal{Y}_i(t) = \{y_i(s), 0 \leq s < t\}$ denotes the covariate history up to to time t for subject i.

Definition 7. *A covariate is called exogenous if the future path of the covariate up to any time $t > s$ is not affected by the occurrence of an event at time point s, more formally if*

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

Examples of exogenous covariates are the time of the day or the season of the year, rather than stochastic processes that are external to the subjects under study. In general, an exogenous covariate is a predictable process, meaning that its value at any time t is known infinitesimally before t. The survival function conditional on the covariate path is given by:

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

On the other hand, the endogenous covariates are the ones that do not satisfy the definition 7, in particular, their existence is directly related to failure status, they are measured with error and the complete history is not available

(e.g. biomarkers, clinical parameters, ...). The first important feature of an endogenous covariate is that it typically require the survival of the subject for its existence: thus, if the failure event is death, the trajectory of the biomarker carries direct information about the failure time. In particular, provided that $y_i(t - ds)$ with $ds \to 0$ exists, the survival function satisfies

$$S_i(t|\mathcal{Y}_i(t)) = \mathbb{P}(T_i^* > t|\mathcal{Y}_i(t)) = 1$$

that is, clearly it is the certain event and failure of the subject at time s involves the non-existence of the covariate at time $t \geq s$. Moreover, a direct consequence is that, contrary to exogenous covariates, the hazard function is not directly related to a survival function (for more details, see [44], Section 6.3).
The second feature stays in the measurement error that primarily refers to the biological variation induced by the patient, in fact we do not expect to observe exactly the same value for an endogenous covariate if we measure the patient twice in the same day. Finally, we can observed only measurements that patient provides when he is visited, and not between these visit times.
An extension of the Cox Model (also known as the Andersen-Gill model, [5]), allows to handle exogenous time-dependent covariates but is not suitable for endogenous covariates.

$$h_j(t|X) = h_{0j} \exp[\beta' X + \alpha y_i(t)]$$

As we can seen in the previous model, the time-dependent covariate $y_i(\cdot)$ is assumed to change value at each follow-up visits while it remains constant in the time interval from this visit and the next. This approach deals the longitudinal trajectory of the time-dependent covariate as a step-function and postulates that the hazard for an event, at any time point t, is associates with the extrapolated value of the covariate at the same time point. This approach is obviously unrealistic for a endogenous covariate such as a biomarker because this approach carries forward the last value, ignoring the previous history. Moreover, the parameter estimates and their standard errors can be severely biased [71]. Anyway, this approach remains one of the most commonly used in clinical practice when the interest lies in the relationship between the longitudinal change of a covariate and the onset of an event of interest.

2.6 Diagnostic

For failure time models, a multitude of diagnostics is available for model misspecification, outliers, influential points or others. In particular, two aspects

must be investigated: the first regards the proportional hazard assumption while the second regard the residual diagnostic as usually for regression models. As in the section on the Linear Mixed Model diagnostic in Chapter 1, in the following sections there is a brief presentation of the techniques used to assess the goodness-of-fit of the survival regression model.

2.6.1 Checking for Proportional Hazard assumption

As previous seen, the main assumption for the Cox models is that each covariate has a multiplicative effect on the hazard function. Under this assumption the survival or hazard curve under different levels of the same variable are parallel i.e. that the proportional hazard assumption (PH) is verified.
There are several options for checking the assumption of proportional hazards, mainly grouped into two categories. The first category includes different graphical methods, the second one includes more formal goodness of fit tests. For categorical variables, the methods are used to test the proportionality among each level of the covariate of interest, in case of continuous variable to check the proportional hazards assumption we need to categorize the variable (e.g. using percentiles).

Graphical methods

Among the graphical methods, usually the following are adopted:

- plots of survival estimates for each level to check whether if the estimated survival curves are fairly separated or converge or cross;
- plots of $\log[-\log(\hat{S}(t)]$ for each level against the $\log(t)$;
- plots of Schoenfeld residuals against time to check the absence of a trend.

The first plot is basic and it is very exploratory while the second is one of the most used and it is described below.
Considering a Cox model with a categorical covariate X we have:

$$h(t|X) = h_0(t)\exp(\beta' X)$$

then

$$S(t|X) = S_0(t)^{\exp(\beta' X)}$$

By applying the logarithm to the previous equation, we obtain

$$\log[S(t|X)] = \log[S_0(t)] \cdot \exp(\beta' X)$$

that is a negative quantities. Then, applying the logarithm again to the negative of the previous one, we obtain

$$\log[-\log[S(t|X)]] = \log[-\log[S_0(t)]] + \beta' X$$

where we can see that the quantity $\beta' X$ translates the survival curve up or down for each level of X. Thus, to assess if the hazard at different time are actually proportional to each other over time we can calculate the survival curve for each level of X, compute the $\log[-\log[\hat{S}(t;X)]]$ and plot against the logarithm of the time to see if they are parallel over the time. In case of the multiple failure model we can fit several multiple model adjusting for other covariates, one for each level of the variable for which we want to test the PH assumption.
To assess the PH assumption is important because if the truth is non-PH and we fit a PH model we must pay attention in the interpretation of the results, in fact we are in some fashion estimating an average logarithm of the hazard ratio. In case of PH assumption is not satisfied, possible solutions could be the transformation of the covariate, doing a stratified analysis or trying other models.

2.6.2 Residual diagnostic

In the survival context , residuals are somewhat different than for other types of models, manly due to the censoring. In general we can identify three types of residuals:

- generalized (Cox-Snell) [20];
- Schoenfeld [88];
- martingale.

Before defining the different types of residuals, it is necessary to recall the cumulative hazard function $\mathcal{H}(\cdot)$ seen in the section 2.1.
Let be T_i the survival time for the i-th individual and $S_i(\cdot)$ the relative survival function, then the transformed random variable $S_i(T_i)$ should have a Uniform distribution on $[0, 1]$, and hence

$$\mathcal{H}_i(T_i) = -\log[S_i(T_i)] \sim \text{Exp}(1)$$

In the next paragraphs we dwell on the Cox-Snell and martingale residuals.

Generalized (Cox-Snell) Residuals

The implication of the last result is that if the model is correct, the estimated cumulative hazard for each individual at the time of their death or censoring, $\hat{\mathcal{H}}_i(T_i)$, for $i = 1, ..., n$ should be like a censored sample from a unit Exponential distribution. Given the cumulative hazard function $\hat{\mathcal{H}}_i(T_i|X_i)$, the quantity

$$r_i = \hat{\mathcal{H}}_i(T_i|X_i) = \hat{\mathcal{H}}_0(T_i)\exp(\beta' X_i),\ i = 1, ..., n$$

is called generalized or Cox-Snell residual for the i-th individual [20]. To assess whether Cox-Snell residuals come from a unit Exponential distribution, we plot the $\log[\hat{\mathcal{H}}_i(r_i)]$ against $\log(r_i)$ to obtain a straight line through the origin with slope of 1. This type of residuals are mainly used to assess the overall goodness-of-fit of the model.

Martingale Residuals

Martingale residuals are defined for the i-th individual as:

$$M_i = \delta_i - \hat{\mathcal{H}}_i(T_i) = \delta_i - \hat{\mathcal{H}}_0(T_i)\exp(\beta' X_i).$$

The residuals M_i can be viewed as the difference between the observed number of deaths (0 or 1) for each subject between time 0 and T_i and the expected numbers based on the fitted model. Differently from the Cox-Snell residuals, martingale residuals are used to check the best functional form $f(X)$ of the covariates in the model. To find the best transformation of X, we need to plot the martingale residual against the covariate for which we want to assess the functional form and overlay this with a smoothed curve (e.g. LOWESS): this curve should suggest the transformation of X. In this way, if the plot is linear we do not transform the covariate, if there is a threshold we can discretize the variable, if it assumes other forms we can try with other transformations.

2.7 Summary of chapter

In this chapter there is a brief review of the most widely used techniques to analyse time-to-event data. This chapter, like the previous one, introduces basic concepts necessary for the definition of the Joint Model. In particular, parametric models and the Cox proportional hazards model are described for the analysis of survival data.

Chapter 3

Joint Modelling

Personalized medicine has gained more interest in the last years; however tools that help clinicians to monitor the history of patient are still not widespread and used. Repeated measures of clinical and lab exams are always collected in hospital database but rarely are used in order to improve prediction of patients' prognosis. The main goal of this research is to show how the longitudinal variation of some biomarkers (e.g. systolic blood pressure, haemoglobin, heart rate, ...) that are endogenous variables, could be informative to predict the event of interest (e.g. death) while the most popular Cox model in Chapter 2 uses only the measurements at the baseline or fixed at some time point (see also its generalization in Section 2.5 which includes time varying biomarkers, even if it is affected by the bias in presence of endogenous time-varying covariates). The Joint Model (JM) for longitudinal and time-to-event data allows the study of the association structure between several measured biomarkers collected over a series of visits and time until an event of interest occurs. We are interested in deriving a dynamic individualized prediction of the either longitudinal and survival process based on JMs. This section starts with a naive approach upon which we want to improve a naive approach to the problem.

3.1 A naive two-stage model

In the survival analysis with a time-varying covariate, seen in the Chapter 2, we are assuming that the longitudinal covariate is observed error-free and its value changes only at each observation time point. When we model the longitudinal trajectory of a biomarker using a linear mixed model, we are creating a model for the response at each follow-up visit. In a JM instead of using the observed biomarker values, in the first stage we fit a linear mixed

model and use it in order to obtain subject-specific predictions of the true and unobserved values to be used in the second stage as if they were the the longitudinal response at the observation times. In practice, for each subject $i = 1, ..., N$ we define a linear mixed model for the longitudinal response as follows

$$Y_i(t) = m_i(t) + \varepsilon_i(t)$$

with $\varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$ and

$$m_i(t) = \beta' X_i(t) + b_i' Z_i(t)$$

in the notation of Chapter 1. We obtain the subject-specific predictions $\hat{m}_i(t)$ and use these in the second stage as time-varying covariate in the next model:

$$h_i(t|X) = h_0 \exp[\gamma' W_i + \alpha \hat{m}_i(t)].$$

As in the case of survival analysis with time-varying covariates, also in this case we are still assuming the longitudinal values, collected at each subsequent visit, do not change between measurement time-points. This remains always unrealistic. Moreover the uncertainty in the estimates from the first stage is not carried through to the second stage.
Certainly this approach is good in term of computational efficiency and quite simple to use, moreover reduces the bias of the association parameters α in comparison with a time-varying Cox model. However, it is not as optimal as modelling both the longitudinal and survival process simultaneously. There are at least two issues: firstly, the uncertainty in the estimates from the first stage is not carry through to the second stage [99]; secondly, we do not keep into account the change of value of the longitudinal outcome between each visit.

3.2 The Joint Model formulation

Let N be the number of subjects and let $\mathcal{D}_N = \{T_i, \delta_i, \mathcal{Y}_i(t); i = 1, \ldots, N\}$ denote a sample from the target population, where $T_i = min(T_i^*, C_i)$ is the observed event time for the i-th subject, with T_i^* being the random variable of the failure times and C_i a non-negative censoring variable. In addition to observing T_i we also get to see the event indicator $\delta_i = \mathbb{I}(T_i^* \leq C_i)$. We focus on the endogenous time-dependent covariate $\mathcal{Y}_i = \{\mathbf{y_i}(s), 0 \leq s \leq t\}$ where $\mathcal{Y}_i(t)$ is the vector of n_i observed repeated measurements $\mathbf{y_i}$ of a biomarker collected up time t for the i-th subject. As in the previous section and in the application in Chapter 5 and 6, we assume that $\mathcal{Y}_i(t)$ is univariate, while the generalisation to the multivariate case is treated in

Section 3.6. In particular, we note that for the subject i the number of observations n_i could not be set a priori, and neither the time points t_{ij}, $j = 1, \ldots, n_i$. Furthermore, we might not observe $\mathbf{y_i}(s)$ at each predefined time point $0 \leq s \leq t_{n_i}$ but only in which occasion when the measurement was taken. Thus, for the i-th subject, the observed longitudinal trajectory consists of the measurements $\mathcal{Y}_{ij} = \{\mathbf{y_i}(t_{ij}), j = 1, \ldots, n_i\}$. The Joint Model formulation can be decomposed in a two sub-models to be estimated jointly.

3.2.1 Survival sub-model

Firstly, we define a relative risk model in order to assess the relationship between a set of covariates, including the longitudinal marker, and the risk of an event of interest. In particular we define a set of covariates as made of both the W_i design matrix of baseline covariates and the value $m_i(t)$ that denotes the true and unobserved value of the longitudinal marker at time (t). Let's define a relative risk model for subject i

$$\begin{aligned} h_i(t|\mathcal{M}_i(t), W_i) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T_i^* < t + \delta t | T_i^* \geq t, \mathcal{M}_i(t), W_i)}{\delta t} \\ &= h_0(t) \exp(\gamma' W_i + \alpha m_i(t)) \end{aligned} \tag{3.1}$$

where $\mathcal{M}_i = \{m_i(s), 0 \leq s < t\}$ is the whole longitudinal and unobserved trajectory, while $m_i(t)$ is a time-varying covariate. Finally, γ is a vector of regression coefficients for each covariate in W_i (i.e. it contains the log-hazard ratios for one unit increase in the relative covariate) while parameter α quantifies the effect of the underlying longitudinal response to the risk for an event and it should be interpreted as 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.
In the model 3.1 we can observe that the risk for an event at time t depends only on the current value of the time-dependent marker $m_i(t)$. Only observing the survival function we can note that the whole history of longitudinal response influences the survival function:

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

This fact is very important in the phase of the estimation of the joint model because the survival function is a part of the likelihood of the model in analogy to what we saw in Chapter 2. Unlike the Cox model, in the joint model

formulation the baseline hazard function $h_0(t)$ needs to be explicit in order to avoid underestimation of the standard errors of the parameter estimates ([40]. Usually, a way is given by the use of a risk function corresponding to a known parametric distribution (e.g. Weibull, log-normal or Gamma). We can choose a parametric but flexible specification of the baseline risk function using a piecewise-constant or regression splines approaches (for more details see [80]).

3.2.2 Longitudinal sub-model

We have seen that in the survival sub-model there is a time-dependent covariate $m_i(t)$ denoting the true and unobserved value of the longitudinal response at time point t. Because the longitudinal information is collected intermittently at specific occasions t_{ij} and moreover it is collected with a measurement error, we need to estimate $m_i(t)$ and successfully reconstruct the complete longitudinal history $\mathcal{M}_i(t)$ for each subject. To accomplish this we define a mixed-effects model in order to describe not so much the average longitudinal profile for the sample but the subject-specific trajectory for the response of interest. The following mixed model is developed for a normally distributed longitudinal outcomes, but generalizations to other model are possible [57]. Using a similar notation as in Chapter 1, we have:

$$\begin{cases} Y_i(t) = m_i(t) + \varepsilon_i(t) \\ m_i(t) = \beta' X_i(t) + b_i' Z_i(t) \\ \varepsilon_i(t) \sim \mathcal{N}_{n_i}(0, \Sigma_i) \\ b_i \sim \mathcal{N}_q(0, D) \\ b_1, .., b_q, \varepsilon_1, .., \varepsilon_{n_i} \text{ independent} \end{cases} \tag{3.2}$$

where Y_i is the vector of n_i observed repeated measurements for subject i, X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ dimensional matrices of known covariates, β is a p-dimensional vector containing the fixed (i.e. population-specific) effects, b_i is the q-dimensional vector containing the random (i.e. subject-specific) effects and ε_i is a n_i-dimensional vector of residual components. Finally, D and Σ_i are the covariance matrices. In our case, for each subject i, X_i contains the fixed intercept and time while Z_i the random intercept and time. No other covariates are included in the model 3.2 even if it may be possible. The mixed model accounts for the measurement error (see Section 2.5) postulating that the observed level of the longitudinal outcome $Y_i(t)$ equals the true level $m_i(t)$ plus a random error term, moreover using the random intercept and slope for the time effect we can reconstruct the complete path of the time-dependent process $\mathcal{M}_i(t)$.

As already seen, the survival function depends on the whole history of the true marker levels $\mathcal{M}_i(t)$, and therefore, for an accurate estimation of $S_i(t)$ it is important to obtain a good estimate of the longitudinal trajectory. Therefore, it becomes very important to develop an elaborate specification of the time structure both in the fixed part $X_i(t)$ and subject-specific part $Z_i(t)$. Several structures have been developed for this aim: for example, in applications in which subjects show highly non-linear longitudinal trajectories, it is advisable to consider flexible representations for the time variable using polynomials structures or splines [24].

3.3 ML estimation of fixed effects

Two methods were manly proposed in order to estimate the parameters of the joint model: the first approach, treated in this section, consists of a semi-parametric maximum likelihood [40][38][113], the second method involves Bayesian techniques with MCMC [18][14][112][114]. A further third approach is implemented by Tsiatis and Davidian [102] that proposed a conditional score approach in which the random effects are treated as nuisance parameters. In the next sections the first two approaches are explored, in particular the maximum likelihood estimation when we are interested in a single longitudinal trajectory while the Bayesian approach when we have two or more longitudinal response.

3.3.1 Semi-parametric ML estimation

Maximum likelihood estimation for joint models is based on the maximization of the log-likelihood corresponding to the joint distribution of the time-to-event $\{T_i, \delta_i\}$ and longitudinal outcomes $\{Y_i\}$ in 3.2. We assume that the vector of time-independent random effects b_i is underlying both the longitudinal and survival processes. In other words, the random effects enclose the association between the longitudinal response and event of interest, in addition to the correlation between the repeated measurements on the same subject in the longitudinal process (conditional independence) as seen with the Linear Mixed Model.
Following we use the standard notation for which $p(A)$ and $p(A|B)$ is the probability density function of a random variable A and of A given B respecively.
Combining the two models 3.1 and 3.2, we can define a model for their joint distribution as follows.

Theorem 2. *Let b_i the random variable for the random effects of a linear mixed model for the longitudinal data Y_i, and let T_i and δ_i be two random variables which define the survival time for the subject $i = 1, \ldots, N$. The joint distribution of T_i, δ_i and Y_i, given b_i is*

$$p(T_i, \delta_i, Y_i|b_i; \vartheta) = p(T_i, \delta_i|b_i; \vartheta)p(Y_i|b_i; \vartheta)$$

where:

$$p(T_i, \delta_i|b_i; \vartheta) = p(T_i)^{\delta_i}\, S(T_i)^{1-\delta_i}$$

$$p(Y_i|b_i; \vartheta) = \prod_j^{n_i} p(\mathbf{y_i}(t_{ij})|b_i; \vartheta)$$

and $\vartheta = (\vartheta_t^T, \vartheta_y^T, \vartheta_b^T)$ denotes the whole parameter vector made of the components parameters for the time-to-event, longitudinal and random-effects covariance matrix respectively, while n_i is the number of repeated measurements for the i-th subject.

This situation in which the vector of subject-specific random effects b_i explains all interdependencies is referred to as *full conditional independence*. In fact, given b_i the longitudinal outcome is independent of the time-to-event outcome and also the repeated measurements in the longitudinal outcome are independent of each other.
Other two assumptions regard the censoring and visiting processes that are assumed non-informative. In other words, decisions to withdraw from the study or to turn up for the next visit may depend on observed past history up to time t but there is no additional dependence on underlying, latent subject characteristics associated with prognosis. When either of the two processes depends on the random effect b_i we have a violations of these assumptions, however evaluating the non-informativeness of the two processes requires external information on the study sector, because the data do not contain enough information on this.
Under these assumptions, we can define the log-likelihood contribution for the i-th subject by integrating out the random effect as follows :

$$\begin{aligned} l_{ML}(\vartheta; T_i, \delta_i, Y_i) &= \log\, p(T_i, \delta_i, Y_i; \vartheta) \\ &= \log \int p(T_i, \delta_i, Y_i, b_i; \vartheta)\, db_i \\ &= \log \int p(T_i, \delta_i|b_i; \vartheta_t, \beta)p(Y_i|b_i; \vartheta_y)p(b_i; \vartheta_b)\, db_i \end{aligned}$$

The first factor under the integral sign regards the conditional density of the survival time defined by the random variables T_i and δ_i. It depends on the whole longitudinal history and takes the following form:

$$\begin{aligned} p(T_i, \delta_i | b_i; \vartheta_t, \beta) &= h_i(T_i | \mathcal{M}_i(T_i); \vartheta_t, \beta)^{\delta_i} \cdot S_i(T_i | \mathcal{M}_i(T_i); \vartheta_t, \beta)^{1-\delta_i} \\ &= [h_0(T_i) \exp(\gamma' W_i + \alpha m_i(T_i))]^{\delta_i} \cdot \\ &\quad \cdot \exp\left(- \int_0^{T_i} h_0(s) \exp(\gamma^T W_i + \alpha m_i(s))\, ds \right) \end{aligned}$$

where $h_0(\cdot)$ can be any positive function of time as seen in Chapter 2. Because the true unobserved value of the longitudinal response $m_i(\cdot)$ is also estimated from the longitudinal sub-model (represented by β), particular attention must be paid to the definition of the mixed model. A model misspecification may determine a biased estimate of the association parameter α (see [80] for further details).
The joint density for the longitudinal response together with the random effects is given by:

$$\begin{aligned} p(Y_i | b_i; \vartheta_y) p(b_i; \vartheta_b) &= \prod_j p(y_i(t_{ij}) | b_i; \vartheta_y) p(b_i; \vartheta_b) \\ &= (2\pi\sigma^2)^{-\frac{n_i}{2}} \exp\left(- \frac{(Y_i - X_i\beta - Z_i b_i)^T (Y_i - X_i\beta - Z_i b_i)}{2\sigma^2} \right) \cdot \\ &\quad \cdot (2\pi)^{-\frac{q_b}{2}} |D|^{-\frac{1}{2}} \exp\left(- \frac{b_i^T D^{-1} b_i}{2} \right) \end{aligned}$$

where q_b denotes the dimensionality of the random-effects vector (see Section 1.6). $p(\cdot)$ is a density function and $S(\cdot)$ is a survival function.

Maximization of the log-likelihood function with respect to ϑ is a computationally challenging task. This is mainly because both the integral with respect to the random effects, and the integral in the definition of the survival function do not have an analytical solution, except in rare cases. Standard numerical integration techniques such as Gaussian quadrature and Monte Carlo have been successfully applied in the joint modelling framework [95][38][113]. Furthermore, Rizopoulos, Verbeke, and Lesaffre [77] have recently discussed the use of Laplace approximations for joint models, that can be especially useful in high-dimensional random effects settings (e.g., when splines are used in the random effects design matrix). For the maximization of the approximated log-likelihood the Expectation-Maximization (EM) algorithm has been traditionally used in which the random effects are treated as 'missing data'. The main motivation for using this algorithm is the closed-form M-step updates for certain parameters of the joint model. However, a serious drawback of the EM algorithm is its linear convergence rate

that results in slow convergence especially near the maximum. Nonetheless, Rizopoulos et al. [77] have noted that a direct maximization of the observed data log-likelihood, using for instance, a quasi-Newton algorithm [50], requires very similar computations to the EM algorithm. Therefore hybrid optimization approaches that start with EM and then continue with direct maximization can be easily employed.

The score function assumes the following form:

$$
\begin{aligned}
U_\vartheta &= \sum_i \frac{\partial}{\partial \vartheta^T} \log \int p(T_i, \delta_i | b_i; \vartheta) p(Y_i | b_i; \vartheta) p(b_i; \vartheta)\, db_i \\
&= \sum_i \frac{1}{p(T_i, \delta_i, Y_i; \vartheta)} \frac{\partial}{\partial \vartheta^T} \int p(T_i, \delta_i | b_i; \vartheta) p(Y_i | b_i; \vartheta) p(b_i; \vartheta)\, db_i \\
&= \sum_i \frac{1}{p(T_i, \delta_i, Y_i; \vartheta)} \int \frac{\partial}{\partial \vartheta^T} [p(T_i, \delta_i | b_i; \vartheta) p(Y_i | b_i; \vartheta) p(b_i; \vartheta)]\, db_i \\
&= \sum_i \frac{1}{p(T_i, \delta_i, Y_i; \vartheta)} \int \frac{\partial}{\partial \vartheta^T} A(\vartheta)\, db_i \\
&= \sum_i \frac{1}{p(T_i, \delta_i, Y_i; \vartheta)} \int \left[\frac{\partial}{\partial \vartheta^T} \log A(\vartheta) \right] A(\vartheta)\, db_i \\
&= \sum_i \int \left[\frac{\partial}{\partial \vartheta^T} \log A(\vartheta) \right] \frac{A(\vartheta)}{p(T_i, \delta_i, Y_i; \vartheta)}\, db_i \\
&= \sum_i \int \left[\frac{\partial}{\partial \vartheta^T} \log p(T_i, \delta_i | b_i; \vartheta) p(Y_i | b_i; \vartheta) p(b_i; \vartheta) \right] \frac{p(T_i, \delta_i | b_i; \vartheta) p(Y_i | b_i; \vartheta) p(b_i; \vartheta)}{p(T_i, \delta_i, Y_i; \vartheta)}\, db_i \\
&= \sum_i \int \left[\frac{\partial}{\partial \vartheta^T} \log p(T_i, \delta_i | b_i; \vartheta) p(Y_i | b_i; \vartheta) p(b_i; \vartheta) \right] p(b_i | T_i, \delta_i, Y_i; \vartheta)\, db_i \\
&= \sum_i \int C(\vartheta, b_i) p(b_i | T_i, \delta_i, Y_i; \vartheta)\, db_i
\end{aligned}
$$

where

$$
\begin{aligned}
C(\vartheta, b_i) &= \frac{\partial}{\partial \vartheta^T} \log p(T_i, \delta_i | b_i; \vartheta) p(Y_i | b_i; \vartheta) p(b_i; \vartheta) \\
&= \frac{\partial}{\partial \vartheta^T} \log p(T_i, \delta_i, Y_i, b_i; \vartheta)
\end{aligned}
$$

is the complete score function. Hence, the observed score function U_ϑ is expressed as the expected value of the complete score function with respect to the posterior distribution of the random effects $p(b_i | T_i, \delta_i, Y_i; \vartheta)$.
If the score equation is solved with respect to ϑ, with $p(b_i | T_i, \delta_i, Y_i; \vartheta)$ fixed at the ϑ value of the previous iteration, then this corresponds to an EM algorithm, whereas if the score equation is solved with respect to ϑ and considering $p(b_i | T_i, \delta_i, Y_i; \vartheta)$ also a function of ϑ, then this corresponds to a direct

maximization of the log-likelihood $\log p(T_i, \delta_i, Y_i; \vartheta)$. More details regarding the specification of the steps of the EM algorithm for joint models are given in Rizopoulos [80] Appendix B.
This last fact also facilitates a straightforward calculation of the standard errors for the parameter estimates. In particular, even if we have estimated the joint model using the EM algorithm, we can easily make use of the observed score function to calculate the Hessian matrix and subsequently standard errors using the observed information matrix. Using similar computations as seen above in the derivation of the U_ϑ, we can rewrite the Hessian matrix in the following form:

$$\begin{aligned}
\frac{\partial U_i(\vartheta)}{\partial \vartheta} &= \frac{\partial}{\partial \vartheta} \int C(\vartheta, b_i) p(b_i | T_i, \delta_i, Y_i; \vartheta)\, db_i \\
&= \int \frac{\partial C(\vartheta, b_i)}{\partial \vartheta} p(b_i | T_i, \delta_i, Y_i; \vartheta) + C(\vartheta, b_i) \frac{\partial p(b_i | T_i, \delta_i, Y_i; \vartheta)}{\partial \vartheta}\, db_i \\
&= \int \frac{\partial C(\vartheta, b_i)}{\partial \vartheta} p(b_i | T_i, \delta_i, Y_i; \vartheta)\, db_i + \int C(\vartheta, b_i) \frac{\partial p(b_i | T_i, \delta_i, Y_i; \vartheta)}{\partial \vartheta}\, db_i
\end{aligned}$$

The asymptotic MLE for the variance of the parameter estimates is given below. It is based on the estimated observed information matrix, not on the expected, due to the drop-out caused by the occurrence of the events:

$$\begin{aligned}
\mathbb{V}(\hat{\vartheta}) &= \mathcal{I}(\hat{\vartheta})^{-1} \\
&= \left(- \sum_{i=1}^{n} \frac{\partial U_i(\vartheta)}{\partial \vartheta} \bigg|_{\vartheta = \hat{\vartheta}} \right)^{-1} \\
&= \left(- \sum_{i=1}^{n} \frac{\partial^2 \log p(T_i, \delta_i, Y_i; \vartheta)}{\partial \vartheta^T \partial \vartheta} \bigg|_{\vartheta = \hat{\vartheta}} \right)^{-1}.
\end{aligned}$$

In general, this is valid unless the baseline risk function for the survival part is unspecified. While in the Cox model the standard errors and inference for the regression coefficients enjoy nice asymptotic properties similar to those of asymptotic maximum likelihood theory, even not having specified an appropriate baseline risk function [5], under the joint modelling framework this feature can not carry over [80]. The main problem is relative to the high dimension of the parameter vector ϑ and the need for techniques as Bootstrapping [29] that require an high computational effort.
A feasible alternative is to postulate a flexible but parametric model for the baseline hazard function $h_0(t)$, with particular reference to splines or

a piecewise-constant model. This approach leads at least two advantages: first, these models can be made arbitrarily flexible by increasing the number of knots, and thus capture various shapes of baseline hazard function, and second, under such models, estimation of standard errors directly follows from asymptotic maximum likelihood theory [22].

3.3.2 Asymptotic inference

Having fitted the joint model under a maximum likelihood framework, the standard asymptotic likelihood inference tests are directly available. In general, if we are interested in testing the null hypothesis

$$\begin{cases} H_0 : \vartheta = \vartheta_0 \\ H_a : \vartheta \neq \vartheta_0 \end{cases}$$

we could use a:

- **Likelihood Ratio Test.**

$$LRT = -2ln\left[\frac{L_{ML}(\hat{\vartheta}_0)}{L_{ML}(\hat{\vartheta})}\right] \longrightarrow \chi^2_{df}$$

  where $L_{ML}(\hat{\vartheta}_0)$ is the maximum likelihood function evaluated under the null ($\hat{\vartheta}_0$) and alternative hypothesis ($\hat{\vartheta}$) respectively.

- **Score Test.**

$$U = U_\vartheta(\hat{\vartheta}_0)^T \mathcal{I}(\hat{\vartheta_0})^{-1} U_\vartheta(\hat{\vartheta}_0) \longrightarrow \chi^2_{df}$$

- **Wald Test.**

$$W = (\hat{\vartheta} - \vartheta_0)^T \mathcal{I}(\hat{\vartheta_0})^{-1} (\hat{\vartheta} - \vartheta_0) \longrightarrow \chi^2_{df}$$

that, under H_0, all tests follow asymptotically a chi-squared distribution with *df* equal to the difference in number of parameters between two models. If only one parameter is tested, the Wald test has the following form:

$$W = \frac{\hat{\vartheta}_j - \vartheta_{0j}}{se(\hat{\vartheta}_j)} \longrightarrow \mathcal{N}(0,1)$$

These test statistics are asymptotically equivalent. However, when we are dealing with finite samples, they usually differ. In this case, the likelihood

ratio test is generally considered the most reliable and the Wald test the least reliable. The score and Wald test require fitting the model only under the null and alternative hypotheses, respectively, whereas the likelihood ratio test is a bit more computationally expensive because requires to fit the joint model under both hypothesis. If there are missing data in the variable we are interested to test for, then the score test will be more efficient since it requires fitting the model only under the null and therefore, avoids a case-wise deletion of missing values (i.e., excluding subjects who have a missing value in the variable of interest).
As seen in Section 1.5.1, a problem with the Wald test for testing the fixed effects in the linear mixed sub-model is that it is based on standard errors which underestimate the true variability in $\hat{\beta}$ because they do not take into account the variability introduced by estimating the covariance matrix for the random effects [27]. For this reason, typically an approximate F distribution with appropriate degrees of freedom (see Section 1.5.1). In joint models this problem could be exacerbated because we do not only ignore the fact that we estimate the variance components, but also that we need to estimate the survival process. Asymptotically, we expect that the Wald statistic will follow the claimed chi-squared distribution, but in finite samples there has not been much work in the joint modelling literature to investigate its properties. Therefore, it is generally advisable to prefer likelihood ratio tests even though they are more computationally expensive.
All three tests are only appropriate for the comparison of two nested models.

When interest lies in comparing non-nested models, information criteria are typically used. The main idea behind these criteria is to compare two models based on their maximized log-likelihood value, but to penalize for the use of too many parameters. The two most commonly used information criteria are the Akaike's Information Criterion (AIC) [2] and the Bayesian Information Criterion (BIC) [89]. As already seen in Section 1.5.3, the two indices are the following:

$$AIC = -2l(\hat{\vartheta}) + 2p$$

$$BIC = -2l(\hat{\vartheta}) + p \log N$$

where p denotes the number of the parameters in the model. The BIC applies a greater penalty for the models with more parameters than does the AIC, because it multiplies the number of parameters fitting with the natural logarithm of the number of total observation used. In general, the choice of the model follows the rule "smaller is better", however they do not always agree. AIC tends to select more elaborate models than BIC due to the fact that the latter penalizes much more heavily for the complexity of the

model. An additional important issue arises when we are interested in testing whether an extra random effect should be included in the joint model. This in fact corresponds to increasing the dimensionality of the random effects design matrix D with extra variance components. In this case the model under the null hypothesis is obtained by setting some of the elements of D to zero in the full model. At least one of these elements is always an element in the diagonal of D (i.e., a variance parameter), meaning that under the null some parameters are set to a value on the boundary of their parameter space. The problem under this setting is that the classical maximum likelihood asymptotic arguments do not apply to boundary cases. In particular, some work on this topic in the linear mixed models framework by Stram and Lee [97] following the results of Self and Liang [90], and later by Verbeke and Molenberghs [109] and [58] has shown that all three tests statistics we have seen above do not follow the claimed χ^2_p distribution under the null. As an alternative, it has been suggested to use mixtures of chi-squared distributions with appropriately chosen degrees of freedom. However, Greven et al. [34] have demonstrated that even this choice could be rather conservative in some settings, and they have instead proposed a simulation based approach to approximate the distribution of the likelihood ratio test statistic under the null. Within the joint modelling framework there has not been much work about this issue. As a practical guideline we would suggest using a higher type I error rate, e.g., 10% to 15%, to guarantee that we do not oversimplify the random-effects structure of the posited joint model.

3.4 Estimation of the random effects

In the context of precision medicine, the interest lies in deriving patient-specific predictions for one or both longitudinal and survival outcomes. To derive such predictions, an estimate of the random effects vector b_i is required. In the Linear Mixed Model we have assumed random effects b_i follow a normal distribution $\mathcal{N}(0, D)$ and their estimation has a close-form solution. However, in joint model the distribution of random effects plays a more prominent role because the random effects explain all interdependencies, in fact its expresses the correlation among repeated measurements on the same subject and the association between the longitudinal process and the survival outcome. Random effects are estimated using the Bayesian paradigm.

Proposition 1. *In particular, assuming that $p(b_i; \vartheta)$ is the prior distribution, and that $p(T_i, \delta_i|b_i; \vartheta)p(Y_i|b_i; \vartheta)$ is the conditional likelihood part, we*

can derive the corresponding posterior distribution:

$$p(b_i|T_i, \delta_i, Y_i; \vartheta) = \frac{p(T_i, \delta_i|b_i; \vartheta)p(Y_i|b_i; \vartheta)p(b_i; \vartheta)}{p(T_i, \delta_i, Y_i; \vartheta)}$$

$$\propto p(T_i, \delta_i|b_i; \vartheta)p(Y_i|b_i; \vartheta)p(b_i; \vartheta)$$

In joint models it does not have a closed-form solution and it has to be numerically computed. However, as the number of longitudinal measurements n_i increases, this distribution converges to a normal distribution as happened with the linear mixed models. To describe this posterior distribution an empirical Bayes approach is employed and requires replacing ϑ with its estimate $\hat{\vartheta}$. Derived indexes of location are the posterior mean

$$\bar{b}_i = \int b_i p(b_i|T_i, \delta_i, Y_i; \vartheta)\, db_i$$

and the posterior mode

$$\hat{b}_i = \underset{b}{\text{argmax}}[\log p(b_i|T_i, \delta_i, Y_i; \vartheta)]$$

while its dispersion is defined in term of posterior variance

$$\mathbb{V}(b_i) = \int (b_i - \bar{b}_i)^2 p(b_i|T_i, \delta_i, Y_i; \vartheta)\, db_i$$

or inverse Hessian matrix

$$H_i = \left(-\frac{\partial^2 \log p(b|T_i, \delta_i, Y_i; \vartheta)}{\partial b^T \partial b}\bigg|_{b=\hat{b}_i} \right)^{-1}$$

3.5 Bayesian Joint Model estimation

The maximum likelihood (ML) approach has been described in the section 3.3.1. Another approach to estimate the parameters of Joint Model is using Bayesian techniques. In this case, the estimation of the joint model's parameters proceeds using Markov chain Monte Carlo (MCMC) algorithms. As in the ML approach, the expression for the posterior distribution of the model parameters is derived under the assumptions of full conditional independence of the longitudinal and survival parts given the random effects (see Section 3.3.1), both the longitudinal and event time process are assumed

independent, and the longitudinal responses of each subject are assumed independent. As previous defined, the likelihood function for the i-th subject is given by:

$$p(T_i, \delta_i, Y_i|b_i, \Theta) = p(T_i, \delta_i|b_i, \Theta) \prod_j^{n_i} p(y_{ij}|b_i, \Theta)$$

where Θ is a random variable that denotes the full parameter vector, and $p(\cdot)$ is an appropriate probability density function (see Section 3.3.1). Under this assumptions, the posterior distribution is given by:

$$p(\Theta, b) \propto \prod_i^n p(T_i, \delta_i|b_i, \Theta) \prod_j^{n_i} p(y_{ij}|b_i, \Theta) p(b_i|\Theta) p(\Theta)$$

For Θ, we usually take standard prior distributions. In particular, for the fixed effects of longitudinal submodel, for the regression parameters of the survival sub-model and for the association parameter between longitudinal and survival parts α, we use independent univariate normal priors. This point derives from the fact that subjects are assumed to be independent of each other. However, even if it is not the case of this thesis, it could happen that there are some clusters where the subjects are much more related than others. For example, in case the subjects are enrolled from different hospital, or different departments within the same hospital, the subjects within the same cluster could be more similar than subjects coming from two different groups for treatment, severity or other clinical aspects. In these cases, it could be better to assume different priors for each center or, as more commonly done, to manage the interdependencies by including the cluster variable among the fixed effects in the model.
As for the model under the maximum likelihood approach, particular attention must be paid to the specification of the models for the longitudinal and survival submodels which can produce biased estimates of the coefficients.
The bayesian approach meets less problem of convergence of ML estimation, even if it is computationally more expensive, because involves many iteration of the MCMC algorithms. However, when interest lies in considering several longitudinal trajectories simultaneously, we need a multivariate joint model that works under bayesian approach as we will briefly see in the next section 3.6.

3.6 Multivariate Joint Model

In clinical research, and in particular in precision medicine, the interest lies in developing an algorithm that adapts to the characteristics of each pa-

tient. In this way, an obviously extension of the Joint Model is given by the possibility to follow the patients by studying more than one longitudinal outcome. Moreover, it is not excluded that these can be modelled by distributions other than the Gaussian (Binomial, Poisson, ...) by using an appropriate generalized linear mixed model. Extending to a multivariate case with K longitudinal trajectories is mathematically straightforward. Formally we have:

$$\begin{cases} g_k(\mathbb{E}[Y_{ki}|b_{ki}]) = \eta_{ki}(t) = X_{ki}^T(t)\beta_k + Z_{ki}^T(t)b_{ki} \\ h_i(t) = h_0(t)\exp\left(\gamma^T W_i + \sum_{k=1}^K \alpha_k \eta_{ki}(t)\right) \end{cases}$$

where the first row gives a multivariate generalized linear model, $g_k(\cdot)$ denotes a known one-to-one monotonic link function, and the second row is a survival model with K association parameters $\alpha_1, ..., \alpha_k$ that link each longitudinal outcome with the survival process. However, with the increase of longitudinal outcomes, the number of random effects b_{ki} to estimate grows considerably and standard methods of estimation become computationally prohibitive both under Maximum Likelihood and under Bayesian context.
A two-stage approach is been proposed to overcome these problems (see [56]). Joint Models with Multiple Longitudinal Outcomes and a Time-to-Event Outcome): in the first stage we fit the longitudinal outcomes using a multivariate mixed model and the output of this model is used to fit a survival submodel in the second stage. However, as seen in the Section 3.1 some papers have shown that this approach returns in biased estimates [103] [80] [115]. In a paper of 2019 presented below, Rizopoulos proposes a bayesian adaptation of the two-stage approach using a correction factor based on importance sampling theory [72] to remove the bias and reduce the computational time. In fact, importance sampling allows the use of a sample generated from a different distribution than the distribution of interest and adjust it through weights to look like a sample from the distribution of interest.

3.6.1 Corrected two-stage approach

The two-stage approach is an intuitive and the most often use solution to overcome the too expensive computational cost of fitting the full multivariate joint model. It moves on a Bayesian framework, following Rizopoulos [56] we describe the two-stage as follows.

Stage I

We fit a multivariate mixed model for the longitudinal outcomes using either MCMC or HMC, and we obtain a sample $\{\vartheta_y^{(m)}, b^{(m)}; m = 1, ..., M\}$ of size

M from the posterior,

$$p(\vartheta_y, b|Y) \propto \prod_{i=1}^{n} \prod_{k=1}^{K} \prod_{j=1}^{n_{ki}} p(Y_{kj}|b_{ki}, \vartheta)\, p(b_{ki}|\vartheta)\, p(\vartheta_y)$$

where ϑ_y denotes the subset of the parameters that are included in the definition of the longitudinal sub-models (including the parameters in the distribution of the random effects).

Stage II

Using the sample from Stage I, we obtain a sample for the parameters of the survival sub-model $\{\vartheta_t^{(m)}; m = 1, ..., M\}$ from the corresponding posterior distribution,

$$p(\vartheta_t|T, \delta, \vartheta_y^{(m)}, b^{(m)}) \propto \prod_{i=1}^{n} p(T_i, \delta_i|\vartheta_t, b_i^{(m)}, \vartheta_y^{(m)})\, p(\vartheta_t)$$

where ϑ_t denotes the subset of the parameters that are included in the definition of the survival sub-model.

Even if this two stage procedure entails the same number of iterations as the full Bayesian estimation of the multivariate joint model, the computational benefits derive from the fact that we do not need to numerically integrate the survival sub-model density function in Stage I. However, as already mentioned, this approach results in biased estimates and requires a correction factor.

Importance sampling correction

To handle this problem, Rizopoulos [56] propose the correction of the estimates obtained from the two-stage approach using importance sampling weights [72].
The realizations $\{\vartheta_t^{(m)}, \vartheta_y^{(m)}, b^{(m)}; m = 1, ..., M\}$ obtained with the two-stage approach can be considered a weighted sample from the full posterior distribution of the multivariate joint model with weights given by:

$$w^{(m)} = \frac{p(\vartheta_t^{(m)}|T, \delta, \vartheta_y^{(m)}, b^{(m)})\, p(\vartheta_y^{(m)}, b^{(m)}|Y, T, \delta)}{p(\vartheta_t^{(m)}|T, \delta, \vartheta_y^{(m)}, b^{(m)})\, p(\vartheta_y^{(m)}, b^{(m)}|Y)}$$

where the numerator is given by the posterior distribution of the multivariate joint model, while the denominator is the product of the posterior distributions from each of the two stages.

We can observe that the difference between fitting the full joint model versus the two-stage approach comes from the second term in the numerator and denominator. By expanding these two terms we obtain:

$$\begin{aligned} w^{(m)} &= \frac{p(\vartheta_y^{(m)}, b^{(m)}|Y,T,\delta)}{p(\vartheta_y^{(m)}, b^{(m)}|Y)} \\ &\propto \frac{\prod_i p(Y_i|b_i^{(m)}, \vartheta_y^{(m)})\, p(T_i,\delta_i|b_i^{(m)}, \vartheta_y^{(m)})\, p(b_i^{(m)}|\vartheta_y^{(m)})\, p(\vartheta_y^{(m)})}{\prod_i p(Y_i|b_i^{(m)}, \vartheta_y^{(m)})\, p(b_i^{(m)}|\vartheta_y^{(m)})\, p(\vartheta_y^{(m)})} \\ &= \prod_i p(T_i,\delta_i|b_i^{(m)}, \vartheta_y^{(m)}) \\ &= \prod_i \int p(T_i,\delta_i|b_i^{(m)}, \vartheta_y^{(m)}, \vartheta_t)\, d\vartheta_t = \varpi^{(m)} \end{aligned}$$

To calculate $\varpi^{(m)}$ we need to resolve the integral by using a Laplace approximation:

$$\begin{aligned} \varpi^{(m)} &= \prod_i \int p(T_i,\delta_i|b_i^{(m)}, \vartheta_y^{(m)}, \vartheta_t)\, d\vartheta_t \\ &\approx \frac{p(T_i,\delta_i|b_i^{(m)}, \vartheta_y^{(m)}, \hat{\vartheta}_t^{(m)})}{(2\pi)^{-\frac{q}{2}}|\hat{\Sigma}^{(m)}|^{\frac{1}{2}}} \end{aligned}$$

where q denotes the dimensionality of the ϑ_t vector,

$$\hat{\vartheta}_t^{(m)} = \underset{\vartheta_t}{\text{argmax}}[\log(p(T_i,\delta_i|b_i^{(m)}, \vartheta_y^{(m)}, \vartheta_t))]$$

and

$$|\hat{\Sigma}^{(m)}| = \left| -\frac{\partial^2 \log(p(T_i,\delta_i|b_i^{(m)}, \vartheta_y^{(m)}, \vartheta_t))}{\partial \vartheta_t^T \partial \vartheta_t} \right|_{\vartheta_t = \hat{\vartheta}_t^{(m)}} \Bigg|$$

is the determinant of the Hessian matrix for the ϑ_t parameters.
Then, the self-normalized weights are given by

$$\tilde{w}^{(m)} = \frac{\varpi^{(m)}}{\sum_{m=1}^{M} \varpi^{(m)}}.$$

To evaluate the performance of this approach, Rizopoulos [56] performs a proof-of-concept simulation study and he concludes that

- despite the extra burden performing the Laplace approximation, the two-stage approach with importance sampling correction has minimal computational cost in comparison with the full multivariate joint model;

- the estimate for the parameters of the longitudinal sub-models are similar between multivariate joint model and two-stage approach with importance sampling correction;
- the estimates for the parameters of the survival sub-model are considerably biased in comparison with the full multivariate joint model, even if it was less strong that the simple two-stage approach when we use the importance sampling correction.

The corrected two-stage model produces unbiased estimate for both fixed effects and the variance components of the longitudinal sub-models. However, there is a considerable difference between this approach and the multivariate joint model with regards to the posterior of the random effects. This observation suggests that the weights obtained in the corrected two-stage model could be further improved by updating (in the second stage) not only the parameters of the survival sub-model ϑ_t but also the random effects b.

Modified Stage II

A sample for the parameters of the survival sub-model $\{\vartheta_t^{(m)}, b^{(m)}; m = 1, ..., M\}$ could be obtained from the corresponding joint posterior distribution:

$$p(\vartheta_t, b|T, \delta, Y, \vartheta_y^{(m)}) \propto \prod_{i=1}^{n} \prod_{k=1}^{K} \prod_{j=1}^{n_{ki}} p(Y_{kj}|b_{ki}, \vartheta_y^{(m)})\, p(b_{ki}|\vartheta_y^{(m)})\, p(T_i, \delta_i|\vartheta_t, b_i, \vartheta_Y^{(m)})\, p(\vartheta_t).$$

Admittedly, simulating from $[\vartheta_t, b|T, \delta, Y, \vartheta_y^{(m)}]$ is more computationally expensive than simulating $[\vartheta_t|T, \delta, \vartheta_y^{(m)}, b^{(m)}]$ because we now also need to calculate the densities of the mixed-effect models for the K longitudinal trajectories. Nonetheless, the computational time remains significantly lower than fitting the full joint model. Under this modified second stage the importance sampling weights now take the form:

$$w^{(m)} = \frac{p(\vartheta_t^{(m)}, b^{(m)}|T, \delta, \vartheta_y^{(m)})\, p(\vartheta_y^{(m)}|Y, T, \delta)}{p(\vartheta_t^{(m)}, b^{(m)}|T, \delta, Y, \vartheta_y^{(m)})\, p(\vartheta_y^{(m)}, b^{(m)}|Y)}$$

where the difference lies in the second term in both numerator and denominator. By doing an expansion of these two terms similar to that used in the

previous section, we obtain:

$$\begin{aligned} w^{(m)} &= \frac{p(\vartheta_y^{(m)}|Y,T,\delta)}{p(\vartheta_y^{(m)},b^{(m)}|Y)} \\ &\propto \frac{\prod_i p(Y_i,T_i,\delta_i|\vartheta_y^{(m)})\, p(\vartheta_y^{(m)})}{\prod_i p(Y_i|b_i^{(m)},\vartheta_y^{(m)})\, p(b_i^{(m)}|\vartheta_y^{(m)})\, p(\vartheta_y^{(m)})} \\ &= \frac{\prod_i \int\int p(Y_i|b_i,\vartheta_y^{(m)})\, p(T_i,\delta_i|b_i,\vartheta_y^{(m)},\vartheta_t)\, p(b_i|\vartheta_y^{(m)})\, p(\vartheta_t)\, db_i\, d\vartheta_t}{\prod_i p(Y_i|b_i^{(m)},\vartheta_y^{(m)})\, p(b_i^{(m)}|\vartheta_y^{(m)})} \\ &= \varpi^{(m)} \end{aligned}$$

and the self normalized weights are given by:

$$\tilde{w}^{(m)} = \frac{\varpi^{(m)}}{\sum_{m=1}^{M} \varpi^{(m)}}.$$

As seen previously, the integrals in $\varpi^{(m)}$ are once again approximated using the Laplace method. Let

$$\{\hat{\vartheta}_t^T, \hat{b}_i^T\} = \underset{\vartheta_t, b_i}{\operatorname{argmax}} \left[\sum_j \log p(Y_i|b_i,\vartheta_y^{(m)}) + \log p(T_i,\delta_i|b_i,\vartheta_y^{(m)},\vartheta_t) + \log p(b_i|\vartheta_y^{(m)}) + \log p(\vartheta_t) \right]$$

and let

$$\Sigma_{b_i} = -\frac{\partial^2}{\partial b^T \partial b} \left[\log p(Y_i|b_i,\vartheta_y^{(m)}) + \log p(T_i,\delta_i|b_i,\vartheta_y^{(m)},\hat{\vartheta}_t) + \log p(b_i|\vartheta_y^{(m)}) \right]_{b=\hat{b}_i}$$

denote the Hessian matrix for the random effects, and analogously

$$\Sigma_{\vartheta_t} = -\frac{\partial^2}{\partial b^T \partial b} \left[\log p(T_i,\delta_i|\hat{b}_i,\vartheta_y^{(m)},\vartheta_t) + \log p(\vartheta_t) \right]_{\vartheta_t=\hat{\vartheta}_t}$$

denote the Hessian matrix for the ϑ_t parameters. Then, we approximate the inner integral by

$$p(Y_i,T_i,\delta_i|\vartheta_y^{(m)},\vartheta_t) \approx \frac{p(Y_i|\hat{b}_i,\vartheta_y^{(m)})\, p(T_i,\delta_i|\hat{b}_i,\vartheta_y^{(m)},\hat{\vartheta}_t)\, p(\hat{b}_i|\vartheta_y^{(m)})}{(2\pi)^{-\frac{k}{2}} |\hat{\Sigma}_{b_i}^{(m)}|^{\frac{1}{2}}}$$

where k denotes the number of random effects for each subject i. Analogously, the outer integral is approximated as

$$p(Y_i,T_i,\delta_i|\vartheta_y^{(m)}) \approx \frac{p(Y_i,T_i,\delta_i|\vartheta_y^{(m)},\hat{\vartheta}_t)}{(2\pi)^{-\frac{q}{2}} |\hat{\Sigma}_{\vartheta_t}^{(m)}|^{\frac{1}{2}}}.$$

Also the performance of this modified approach was assessed by using simulation. Rizopoulos [56] concludes:

- given the requirement for a double Laplace approximation, and the fact that the denominator does not simplify, the calculation of the $\varpi^{(m)}$ weights is more expensive than seen in the unmodified Stage II, however these approach requires a computational cost long less expensive and more faster that fitting the full joint model;
- the estimate for the parameters of the longitudinal sub-models are similar among multivariate joint model, two-stage approach with and without correction updating the random effects;
- the bias in the estimation of the parameters of the survival sub-model is eliminated.

3.7 Diagnostic

When it comes to use these models in practice, a prerequisite step is to validate the model's assumptions. The standard tools to assess these are by using the residual plots. How to obtain these plots and their characteristics have been described in the Chapters 1 and 2 when longitudinal and survival outcomes are separately modelled. The techniques used when they are jointly modelled are an adjustment of those seen in Chapters 1 and 2. A new aspect to investigate is instead related to the implications of the non-random drop-out caused by the occurrence of events.

3.7.1 Residuals for the Longitudinal part

When we analyse longitudinal data using a Linear Mixed Model, there exist the conditional (or subject-specific) residuals and the marginal (or population-averaged) residuals (see Section 1.7.1).

Conditional residuals

The conditional (subject-specific) residuals aim to validate the assumptions of the hierarchical version of the model

$$\begin{cases} Y_i = \beta^T X_i + b_i^T Z_i + \varepsilon_i \\ \varepsilon_i \sim N_{n_i}(0, \Sigma_i) \\ b_i \sim N_q(0, D) \end{cases}$$

and are defined as

$$\varepsilon_i = y_i - (\hat{\beta}^T x_i + \hat{b}_i^T z_i)$$

with corresponding standardize (studentized) version

$$\varepsilon_i^{stud} = \frac{y_i - (\hat{\beta}^T x_i + \hat{b}_i^T z_i)}{\hat{\Sigma}_i^{1/2}}$$

where $\hat{\beta}$ and $\hat{\Sigma}$ denote the MLEs while $\hat{b}_i$ are the empirical Bayes estimates. These conditional residuals can be used for checking the homoschedasticity and normality assumptions.

Marginal residuals

The marginal residuals focus on the marginal model for Y_i implied by the hierarchical representation

$$\begin{cases} Y_i = \beta^T X_i + \epsilon_i \\ \epsilon_i \sim N_{n_i}(0, V_i) \end{cases}$$

and are defined as

$$\epsilon_i = y_i - \hat{\beta}^T x_i$$

with corresponding studentized version

$$\epsilon_i^{stud} = \frac{y_i - \hat{\beta}^T x_i}{\hat{V}_i^{1/2}}$$

where $V_i = Z_i D Z_i^T + \sigma^2 I_{n_i}$ denotes the marginal covariance matrix of Y_i. These residuals can be used to investigate the misspecification of the mean structure $\beta^T X_i$ as well as to validate the assumptions for the within-subjects covariance structure V_i.

Residual and drop-out process

The onset of the event of interest or another that could prevent the continuation of the study could correspond to a discontinuation of the collection of the longitudinal information and either the follow-up measurements can no longer be collected or their distribution change significantly after the onset of the event. This type of drop-out is obviously not at random like in the case of missed visit. Thus the observed data up to the drop-out time do not constitute a random sample of the target population [110] [31]. This in turn implies that residual plots based on the observed data alone can be misleading because these residuals should not be expected to exhibit standard properties, such as zero mean and independence. In particular, it is difficult

to discern if a systematic trend could be attributable to a misspecification of the design matrix X of the fixed effects or apparent and due to the nonrandom nature of missingness.
To overcome the problems caused by the nonrandom dropout and produce residuals for the longitudinal process that can be readily used in diagnostic plots, Rizopoulos [78] proposed to augment the observed data with randomly imputed longitudinal responses under the complete data model, corresponding to the longitudinal outcomes that would have been observed had the patients not dropped out.
We should note, however, that in some clinical studies in which the terminating event is death, it may not be conceptually reasonable to consider potential values of the longitudinal outcome after the event time. Nonetheless, the multiple-imputed residuals are merely used as a mechanism to help us investigate the fit of the model, and we are not actually interested in inferences after the event time.
An analytic view on multiple imputation approach is not relevant for our purpose but it is available on Rizopoulos [80].

3.7.2 Residuals for the Survival part

A useful tool to check the assumptions for the survival part is based on the martingale residuals (see Section 2.6.2). The martingale process can be seen as the equivalent of the residual term in the generalized linear model and can be viewed as the difference between the observed number of events for the i-th subject by time t and the expected number of events by the same time based on the fitted model. Martingale residuals are mainly used for two aims. Firstly, they are used for a direct identification of excess of events in case of subjects that are poorly fit by the model, then for evaluating the appropriate functional form for covariates inserted in the model fit. In the Joint Model context, a martingale residual for the i-th subject is defined as

$$\begin{aligned} M_i &= \delta_i(t) - \int_0^t R_i(s) h_i(s|\hat{\mathcal{M}}_i(s); \hat{\vartheta})\, ds \\ &= \delta_i(t) - \int_0^t R_i(s) \hat{h}_0 \exp(\hat{\gamma}^T W_i + \hat{\alpha}\hat{m}_i(s))\, ds \end{aligned}$$

where δ_i represent the onset of event for the subject i at time t, $R_i(t)$ is the left continuous at risk process with $R_i(t) = 1$ if the subject i is still at risk at time t and $R_i(t) = 0$ otherwise, $\hat{m}_i(t) = X_i^T(t)\hat{\beta} + Z_i^T(t)\hat{b}_i$ is the estimated value of the longitudinal response at time t and $\hat{h}_0(\cdot)$ is the estimated baseline hazard function.

Another type of residuals for survival part is given by the Cox-Snell residuals based on the estimated cumulative risk function evaluated at the observed event time T_i, as presented in Chapter 2. With appropriate adjustments due to the JM, a Cox-Snell residual for the subject i is given by

$$\begin{aligned} r_i &= \int_0^{T_i} h_i(s|\hat{\mathcal{M}}_i(s);\hat{\vartheta})\, ds \\ &= \int_0^{t} \hat{h}_0 \exp(\hat{\gamma}^T W_i + \hat{\alpha}\hat{m}_i(s))\, ds \end{aligned}$$

When the chosen model fits the data well, we expect that the probability of failure after time t, i.e. $S(t) = \mathbb{P}(T_i^* > t)$ will have a standard uniform distribution, and therefore the cumulative hazard, defined as $\mathcal{H}(t) = -\log S(t)$ will have a unit exponential distribution.

3.7.3 Random-effect diagnostic

Also the random effects b have distributional assumption to verify. In general, in the context of the mixed models, the choice is to assume that the random effects are normally distributed with mean zero and covariance matrix D. However, because the random effects are latent quantities do not lend themselves to a straightforward construction of residuals. Nonetheless, it has been shown that linear mixed-effects models are relatively robust to misspecification of the random effects [107]. In the Joint Model context, there are mainly two issues for which a misspecification in random effects may influence the inference on the model's parameters. Firstly, random effects have a more prominent role in joint models than in mixed models because they are used to build the association between the longitudinal and event time processes besides capturing the correlations between the repeated measurements. Secondly, the nonrandom dropout caused by the occurrence of events complicates matters because in the missing data literature it is known that inferences in nonrandom dropout settings can be highly sensitive to modelling assumptions.
Nevertheless, simulation studies have been done to investigate these issues and the findings presented in literature suggests that parameter estimates and standard errors were rather robust to misspecification [102]. Moreover, it has been shown [41] that, as the number of repeated measurements n_i per subject increases, the misspecification of the random-effects distribution has a minimal effect in parameter estimations and their standard errors.

3.8 Summary of chapter

A review of the Joint Model for longitudinal and survival outcome, with its basic formulation, is shown in this chapter. In particular, the two stage formulation is presented to better understand the mechanism by which the longitudinal and the survival parts are joined. An introduction to the Bayesian formulation and the Multivariate Joint model is also shown. The next two chapters will describe how the predicted survival probabilities under the Joint Model are used to evaluate the improvement in prediction due to the dynamic updating of the survival curve as new longitudinal values are collected.

Chapter 4

Dynamic prediction

One of the main characteristics of the Joint model is that it allows to update the dynamic prediction of survival probabilities whenever new longitudinal data are collected on the patients. This fact give the possibility to better calibrate the prognostication and the therapy of the patients. It is obviously done in clinical practice by the experience of the physicians, but the use of results coming from a statistical model could provide a more formal way to support the clinical activity.

4.1 Survival probability

To provide the patient with an accurate diagnosis of his state of health (measured by a clinically relevant endpoint) at a certain moment in the future is of paramount importance in clinical research. Usually this is done through risk scores. Several risk scores currently used in clinical practice are based on the survival probabilities estimated at a specific time point u in the future. These scores are usually based on a multivariate survival model (e.g. Cox-PH model, Weibull Model, ...) to predict the mean survival probability in the population under study [15] [68]. This gives a comprehensive risk score where several risk factors are combined into one as a sum of the covariates considered clinically associated to the event weighted by the coefficients estimated by the model. The comprehensive risk score is applied to a baseline survival function estimated on the observed data of a patient in order to derive the survival probability (or hazard) at a specific time point for that patient.
The above is unsatisfactory for various reasons. First it is model based and models might be a rough simplification of the reality in which not all potential risk factors are known and also not all those which are known are included in the model. Second, and more significant, is that the most common approach

is based on using only the baseline information only (i.e. data collected at the first visit, diagnosis of disease, start of the treatment, ...) even when the first visit is followed by others with the relative set of measurements.
Some improvements adopted in clinical practice is given by the fact that the estimate is updated using in the algorithm the last available biomarker measurement at a time t, with $t < u$. For example, it is possible to assume of following a patient for one-month treatment with a weekly visit, and then the interest lies in estimating the probability of being free from an event of interest in the next month. Also this improvement is unsatisfactory because it involves several approximations, in particular:

- a lot of useful information collected between baseline and the last visit (i.e. at one month) would be ignored and only the first and the last data would be used;
- how data change from baseline to last visit would be ignored;
- the last value of the predictor would be held constant up to the time of event or to the last date free from the event.

Another approach, discussed by Rizopoulos in [83], is landmarking prediction. It is somehow similar to the previous one. A survival model is implemented by fitting data relative to patients at risk at a specific time point t after the baseline. In this approach, the model at time t is estimated not on the observed value of the longitudinal biomarker of interest at time t but by replacing these values with the their estimations obtained from a linear mixed model applied on the previous longitudinal history up to time t. In this context, the Joint Model provides a more precise estimate of the the survival probabilities if there is no model misspecification. The survival probabilities are continuously updated because the individual trajectories, and thus their association with the risk of the event, are updated every time the new measurements are collected on the patients.
In this sense the author talks about dynamic prediction of the survival probabilities. An improvement to this could be given by understanding what factors significantly improve the prognosis and how long it is necessary to follow a patient to provide a more precise diagnosis.
In our paper (see Chapter 6) we briefly come back to this by observing how well the Joint Model improves its prediction capability using longitudinal information coming from different time windows.
In this section, following Rizopoulos [79], we detail the theory of estimation of the survival probability at a specific time point in the context of Joint Model which will be used in Chapter 6. Indeed JM allows the simultaneous

derivation of both survival probabilities and longitudinal values for a specific time point of interest in the future.

4.1.1 Estimation of the survival probability

Let N be the number of subjects and let $\mathcal{D}_N = \{T_i, \delta_i, \mathcal{Y}_i(t); i = 1, \ldots, N\}$ denote a sample from the target population, where $T_i = min(T_i^*, C_i)$ is the observed event time for the i-th subject, with T_i^* being the random variable of the failure times and C_i a non-negative censoring variable. In addition to observing T_i we also get to see the event indicator $\delta_i = \mathbb{I}(T_i^* \leq C_i)$. We focus on the endogenous time-dependent covariate $\mathcal{Y}_i = \{\mathbf{y_i}(s), 0 \leq s \leq t\}$ where $\mathcal{Y}_i(t)$ is the vector of n_i observed repeated measurements $\mathbf{y_i}$ of a biomarker collected up time t for the i-th subject. A Joint Model to study the relationship between the longitudinal process $\mathcal{Y}_i(t)$ on the time-to-event T_i is estimated on $\mathcal{D}_n$. Let ϑ be the parameters in the estimated JM.
Let i be new subject coming from the same target population of $\mathcal{D}_n$ with a set of longitudinal measurements $\mathcal{Y}_i(t)$ collected from a baseline up to a time point t and $\delta_i = 0$ (i.e. the new patient is providing new longitudinal measurements and is therefore still alive or free of the event of interest at time t). The interest lies in estimating the survival probability for a time point $u > t$ by merging the new information of the subject i with the set of data on which we run the JM. The fact that the longitudinal biomarker is collected up to t implies that the subject i is free from the event of interest up to this time point, so the focus is to derive the conditional subject-specific survival probability at time u, given that he survived until t. In particular, for any time $u > t$ the interest is the probability that this new subject will survive (i.e. he will be free from the event of interest) at least up to u:

$$\pi_i(u|t) = \mathbb{P}(T_i^* \geq u | T_i^* > t, \mathcal{Y}_j, \mathcal{D}_n; \vartheta)$$

The time-dynamic nature of $\pi_i(u|t)$ is due to the fact the for any new information t' recorded on the subject j between t and u is possible to update of the survival probabilities $\pi_i(u|t')$ by running the model with the new longitudinal measurement.
The assumption of full conditional independence of $\{T_i, \delta_i\}$ and $\{Y_i\}$ given the random effects $\{b_i\}$ (see Section 3.3.1) on which is based the Joint Model, is necessary to estimate the subject-specific conditional survival probability. So the probability $\pi_j(u|t)$ can be rewritten integrating out the random effect as:

$$
\begin{aligned}
\mathbb{P}(T_i^* \geq u | T_i^* > t, \mathcal{Y}_i(t); \vartheta) = & \\
&= \int \mathbb{P}(T_i^* \geq u | T_i^* > t, \mathcal{Y}_i, b_i; \vartheta) \cdot p(b_i | T_i^* > t, \mathcal{Y}_i(t); \vartheta)\, db_i \\
\textit{conditional independence} \rightarrow &= \int \mathbb{P}(T_i^* \geq u | T_i^* > t, b_i; \vartheta) \cdot p(b_i | T_i^* > t, \mathcal{Y}_i(t); \vartheta)\, db_i \\
&= \int \frac{\mathcal{S}_i(u | \mathcal{M}_i(u, b_i, \vartheta); \vartheta)}{\mathcal{S}_j(t | \mathcal{M}_i(t, b_i, \vartheta); \vartheta)} \cdot p(b_i | T_i^* > t,\ \mathcal{Y}_i(t); \vartheta)\, db_i
\end{aligned}
$$

where $\mathcal{D}_n$ and any other baseline covariates are omitted, $\mathcal{S}_i(\cdot)$ is given by the formula seen in Section 3.2.1, $\mathcal{M}_i(\cdot)$ is the approximation obtained by the linear mixed effects model and it is a function of both the random effects and the parameters.
The first-order estimate of $\tilde{\pi}_j(u|t)$ can be obtained by using the empirical Bayes estimate for b_i, that is:

$$
\tilde{\pi}_i(u|t) = \frac{\mathcal{S}_i(u | \mathcal{M}_i(u, \hat{b_i}, \hat{\vartheta}); \hat{\vartheta})}{\mathcal{S}_i(t | \mathcal{M}_i(u, \hat{b_i}, \hat{\vartheta}); \hat{\vartheta})} + O(n_i^{-1})
$$

where $\hat{\vartheta}$ denotes the maximum likelihood estimates for the fixed effects in the model, $\hat{b}_i$ is the posterior mode (see Section 3.4) of the conditional distribution $p(b_i | T_i^* > t, \mathcal{Y}_i(t); \hat{\vartheta})$, and n_i is the number of longitudinal observations collected up t on the subject i. Rizopoulos [79] showed that this estimator works relatively well; however, deriving the standard error, and hence the confidence interval) for $\pi_j(u|t)$ is difficult due to the fact that it is necessary to account for the variability of both the maximum likelihood and empirical Bayes estimates.
The use of a Monte Carlo simulation schemes was proposed by Rizopoulos [79] and Proust-Lima and Taylor [73] and it is supported by the asymptotic Bayesian formulation of the joint model seen in the Section 3.5. In this context, the survival probability can be derived as follow:

$$
\begin{aligned}
\pi_i(u|t) &= \mathbb{P}(T_i^* \geq u | T_i^* > t, \mathcal{Y}_i, \mathcal{D}_n; \vartheta) \\
&= \int \mathbb{P}(T_i^* \geq u | T_i^* > t, \mathcal{Y}_i(t); \vartheta) \cdot p(\vartheta | \mathcal{D}_n)\, d\vartheta
\end{aligned}
$$

The first part of the integrand is given by the previous derivation. The second part is the posterior distribution of the parameters ϑ given the observed data $\mathcal{D}_n$. Let assume that the sample size n is sufficiently large and that it holds

$$
\{\vartheta | \mathcal{D}_n\} \sim \mathcal{N}(\hat{\vartheta}, \hat{\text{var}}(\hat{\vartheta}))
$$

then, using the previous two derivations, the following simulation scheme can be used to derive the Monte Carlo estimation of the survival probability:

$S1:$ Drawn $\vartheta^{(l)} \sim \mathcal{N}(\hat{\vartheta}, \hat{\text{var}}(\hat{\vartheta}))$

$S2:$ Drawn $b_i^{(l)} \sim \{b_i | T_i^* > t, \mathcal{Y}_i(t), \vartheta^{(l)}\} \sim t_4$

$S3:$ Compute $\pi_i^{(l)}(u|t) = \dfrac{\mathcal{S}_i(u|\mathcal{M}_i(u, b_i^{(l)}, \vartheta^{(l)}); \vartheta^{(l)})}{\mathcal{S}_i(t|\mathcal{M}_i(t, b_i^{(l)}, \vartheta^{(l)}); \vartheta^{(l)})}$

$S4:$ Repeat Steps 1-3 for each subject i, $l = 1, \ldots, L$ times, where L denotes the number of Monte Carlo samples.

The second step is based on a Metropolis-Hastings algorithm with independent proposal from a multivariate t-distribution with:

- four degrees of freedom,
- centered at the empirical Bayes estimate $\hat{b_i}$,
- scale matrix $\hat{\text{var}}(\hat{b_i}) = \{-\frac{\partial^2}{\partial b^T \partial b}[\log p(T_i^* > t, \mathcal{Y}_i(t), b; \hat{\vartheta})]_{b=\hat{b_i}}\}^{-1}$.

A justification of this approach can be found in Booth and Hobert (1999) and in Rizopoulos et al. (2008). Conversely to the derivation with the first-order estimator, the maximum likelihood estimate $\hat{\vartheta}$ and the empirical Bayes estimate $\hat{b_i}$ are replaced by $\vartheta^{(l)}$ and $b_i^{(l)}$.
The realizations of $\{\pi_i^{(l)}(u|t), l = 1, \ldots, L\}$ can be used to derive estimates of the survival probabilities for example median or mean:

$$\hat{\pi}_i(u|t) = \text{median}\{\pi_i^{(l)}(u|t), l = 1, \ldots, L\}$$

or

$$\hat{\pi}_i(u|t) = \frac{1}{L} \sum_{l=1}^{L} \pi_i^{(l)}(u|t).$$

In the same manner, the standard error can be derived using the sample variance $\{\pi_i^{(l)}(u|t), l = 1, \ldots, L\}$ while the 95% confidence intervals can be replaced by the using the 2.5% and the 97.5% percentiles of the same sample (or by using the other percentiles related to other credible threshold).

4.2 Longitudinal outcome

The main advantage of using the Joint Model is given by the possibility to modelling simultaneously the survival function with respect to an event of interest and the longitudinal evolution of a marker accounting for the possible confounding effect related to the event of interest. For example, if an increase of a marker is related to the event, the researcher could underestimate the longitudinal increasing trend due to the fact that the highest values for that marker could be not collected due to the onset of the event itself (e.g. death). Due to the dual nature of the Joint Model, a dynamic prediction for the longitudinal marker is also allowed by using an approach similar to that presented for the survival probabilities in the previous sections.

4.2.1 Estimation of the longitudinal outcome

By using the approach proposed by Rizopoulos (2011), Let N be the number of subjects and let $\mathcal{D}_N = \{T_i, \delta_i, \mathcal{Y}_i(t); i = 1, \ldots, N\}$ denote a sample from the target population, where $T_i = min(T_i^*, C_i)$ is the observed event time for the i-th subject, with T_i^* being the random variable of the failure times and C_i a non-negative censoring variable. In addition to observing T_i we also get to see the event indicator $\delta_i = \mathbb{I}(T_i^* \leq C_i)$. We focus on the endogenous time-dependent covariate $\mathcal{Y}_i = \{\mathbf{y_i}(s), 0 \leq s \leq t\}$ where $\mathcal{Y}_i(t)$ is the vector of n_i observed repeated measurements $\mathbf{y_i}$ of a biomarker collected up time t for the i-th subject. A Joint Model to study the relationship between the longitudinal process $\mathcal{Y}_i(t)$ on the time-to-event T_i is estimated on $\mathcal{D}_n$. Let ϑ be the parameters in the estimated JM.
Let i be new subject coming from the same target population of $\mathcal{D}_n$ with a set of longitudinal measurements $\mathcal{Y}_i(t)$ collected from a baseline up to a time point t and $\delta_i = 0$ (i.e. the new patient is providing new longitudinal measurements and is therefore still alive or free of the event of interest at time t).
For a specific subject i who is still free from the event at follow-up time t, the interest lies in the expected value of his longitudinal trajectory from t up to a time $u > t$ accounting for the random process modelling the observed trajectory $\mathcal{Y}_i$ from baseline to t. The expected value of the marker at time $u > t$ is given by:

$$\omega_i(u|t) = \mathbb{E}[\mathbf{y_i}(u)|T_i^* > t, \mathcal{Y}_i, \mathcal{D}_n; \vartheta^*], \ u > t$$

As for the survival probability, also this prediction could be updated by collecting new values for the marker at any visits t' in the interval (t, u).

In order to account for the fact that the true parameter values ϑ^* are not know, the asymptotic Bayesian formulation of the joint model are used to calculate the expected value of $\omega_i(u|t)$ with respect to the posterior distribution of the parameters $\{\vartheta|\mathcal{D}_n\}$ as

$$\mathbb{E}[\mathbf{y_i}(u)|T_i^* > t,\ \mathcal{Y}_i(t), \mathcal{D}_n] = \int \mathbb{E}[\mathbf{y_i}(u)|T_i^* > t,\ \mathcal{Y}_i(t); \vartheta] p(\vartheta|\mathcal{D}_n)\, d\vartheta$$

The first part of the integrand can be simplified by exploiting the conditional independence assumptions as follows:

$$\begin{aligned}
&\mathbb{E}[\mathbf{y_i}(u)|T_i^* > t, \mathcal{Y}_i(t); \vartheta] = \\
&\quad = \int \mathbb{E}[\mathbf{y_i}(u)|T_i^* > t, \mathcal{Y}_i(t), b_i; \vartheta]\ p(b_i|T_i^* > t,\ \mathcal{Y}_i(t); \vartheta)\, db_i \\
&\quad = \int \mathbb{E}[\mathbf{y_i}(u)|b_i]\ p(b_i|T_i^* > t, \mathcal{Y}_i(t); \vartheta)\, db_i \\
&\quad = \int (x_i^T(u)\beta + z_i^T(u)b_i)\ p(b_i|T_i^* > t, \mathcal{Y}_i(t); \vartheta)\, db_i \\
&\quad = x_i^T(u)\beta + z_i^T(u)\overline{b}_i^{(t)}
\end{aligned}$$

where

$$\overline{b}_i^{(t)} = \int b_i \cdot p(b_i|T_i^* > t, \mathcal{Y}_i(t); \vartheta)\, db_i.$$

Under these derivations a straightforward estimator of $\omega_i(u|t)$ is obtained by replacing ϑ with its maximum likelihood estimate $\hat{\vartheta}$ and calculating the mean of the posterior distribution $p(b_i|T_i^* > t, \mathcal{Y}_i(t); \hat{\vartheta})$. A similar estimator could be derived using the mode of the posterior distribution defined as

$$\hat{b}_i(t) = \underset{b}{\text{argmax}}[\log p(b_i|T_i^* > t, \mathcal{Y}_i(t); \hat{\vartheta})]$$

and obtaining

$$\tilde{\omega}_i(u|t) = x_i^T(u)\beta + z_i^T(u)\hat{b}_i^{(t)} + O(n_i(t)^{-1})$$

using the relationship $\overline{b}_i^{(t)} = \hat{b}_i^{(t)} + O(n_i(t)^{-1})$ [101] that holds under sufficient smoothness of the distribution of $\log p(b_i|T_i^* > t, \mathcal{Y}_i(t); \hat{\vartheta})$ and in which $n_i(t)$ denotes the number of longitudinal measurements for the i-th subject by time t.
It has been proved that the mode and the mean of the posterior distribution of the random effects are typically very close to each other. In fact, the density of the longitudinal model (that is well approximated by a normal

distribution) is the leading distribution of the random effects in the joint model context (see Rizopoulos, 2012) [80]. However, the use of the mode is preferable because it is usually a better location measure than mean, especially when the posterior distribution is skewed.
As seen for the prediction of the survival probability, obtaining the standard error of the estimate $\omega_i(t)$ is very difficult because $\overline{b}_i^{(t)}$ and $\hat{b}_i^{(t)}$ are non-linear function of $\hat{\vartheta}$ that cannot be written in closed form. To overcome this problem and obtain confidence intervals for the predicted longitudinal marker, the Monte Carlo approach was proposed as for the survival probability in Section 4.1.1.
Assuming that the sample size is sufficiently large so that $\{\vartheta|\mathcal{D}_n\}$ can be well approximated by a normal distribution with mean the MLEs $\hat{\vartheta}$ and the variance-covariance matrix given by $\{\mathcal{I}(\hat{\vartheta})\}^{-1}$, the following simulation scheme can be obtained:

$$S1 : \text{Drawn}\, \vartheta^{(l)} \sim \mathcal{N}(\hat{\vartheta}, \hat{\text{var}}(\hat{\vartheta}))$$

$$S2 : \text{Drawn}\, b_i^{(l)} \sim \{b_i|T_i^* > t, \mathcal{Y}_i(t), \vartheta^{(l)}\} \sim t_4$$

$$S3 : \text{Compute}\, \omega_i^{(l)}(u|t) = x_i^T(u)\beta^{(l)} + z_i^T(u)b_i^{(l)}$$

where the first two steps are equal to those seen for the prediction of the survival probabilities in Section 4.1.1. These values are then used in the Step 3 to derive the longitudinal outcome. As done for the survival probability, also in this case the 95% confidence intervals can be obtained as credible interval by using the 2.5th and 97.5th percentiles of $\{\omega_i^{(l)}(u|t), l = 1, \ldots, L\}$. The posterior mean or median of $\{\omega_i^{(l)}(u|t)\}$ will be used as predicted value for the longitudinal outcome even if it was proved that the resulting estimates are almost indistinguishable.

4.3 Summary of chapter

This chapter introduces the concept of dynamic prediction that will be used in the next one and in the application in Chapter 6 to evaluate the goodness of the model in its clinical application. The quantities introduced in this chapter are used in the next chapter to evaluate the prediction capability of a score derived by using a Joint Model to study how a longitudinal update of a predictor could improve the prediction of the survival probability.

Chapter 5

Prediction capability

The aim of my thesis is to provide the clinician with an easy-to-interpret tool that can help him/her make decisions for patient care. Several measures were developed to assess how well a model can discriminate between patients who will experience the event in a specific time frame from patients who will experience it in a later time. In particular, some of them are more accurate or useful from a mathematical point of view whereas other are more intuitive [49]. For example, the proportion of variation explained by the covariates or other indexes based on the likelihood (e.g. AIC, BIC) are useful for a statistician, other indexes such as the area under the Receive Operating Characteristics (AUC-ROC) curves or the Net Reclassification Improvement are more accessible and understandable by most people involved in clinical research. These methods are usually applied in a diagnostic setting to evaluate the performance of a new test (given by an instrument, tool, classification algorithm, predictive score and so on...) in predicting a clinical event of interest (AUC-ROC) or to compare two different tools (NRI). These two measures were primarily designed for binary outcomes. However, with little generalisations they can be used also with time-to-event data and in this work extended to the joint model.
In the next section, an overview of these two indexes is presented and linked to Joint Model. In general, it refers about the predictive capability of a score even if the same approach can be applied to several types of measurements coming from a diagnostic test.

5.1 Discrimination based on the ROC curve

In clinical practice, several diagnostic instrument or scores are used to classified a patients in different risk categories for an event of interest. Even if

a continuous measurements could be a more appropriate index to link and model the change of a score to an associated change in the risk of an event, the categorization in two or more three levels of risk allows the researcher to define the patient's clinical picture in a more direct and intuitive way. This implies a better management. To say that a unit increase of a score involves an increase of the risk of 10% or to say that a patient with a score lower than a specific threshold have an expected survival free from an event are two methods of answering to the same question but the second is more intuitive and almost free from misunderstandings. In general, the work is done on a continuous measurements coming from statistical models, but the conclusions and the operative guidelines are then expressed in terms of levels of risk.
In a predictive setting, the capability of a score to predict a future event is firstly assessed through measures of discrimination. Among them, sensitivity, specificity, positive and negative predictive values are the most common and widely used. Starting from a continuous potential predictive score, the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) are used to assess its predictive performance. Let d_i denote the event status indicator for subject i (equal to 1 or 0 in case the patient experienced the event or not), Z_i a score measured at baseline that is believed to be strongly associated with the onset of the event and to be used to identify the patients that have greater chance of having the event. After verifying the association between the score and the event with an appropriate statistical model, a prediction rule is set to classify the i-th subject as at risk of the event when his/her observed score exceeds a specific threshold c. For each potential value of c, it is possible to define the discrimination measurements previously mentioned. In particular, the sensitivity, is defined as the true positive rate, i.e. the probability that the marker correctly classifies a subject that will experience the event:

$$\mathrm{SENS}(c) = \mathbb{P}(Z_i > c \,|\, d_i = 1)$$

whereas the specificity is defined as the probability that the marker correctly classifies an healthy subject:

$$\mathrm{SPEC}(c) = \mathbb{P}(Z_i \leq c \,|\, d_i = 0)$$

The quantity one minus specificity is called false positive rate. Sensitivity and specificity are related to the predictive score and are obtained conditioning on the observed event status, i.e. when the score is under construction and the whole clinical history of a patient is observed. In clinical practice, to measure the probability that a subject will experience the event given that

the score exceeds the chosen threshold, the positive and negative predictive values are used. The positive predictive value is defined as

$$\mathrm{PPV}(c) = \mathbb{P}(d_i = 1 \,|\, Z_i > c)$$

while the negative predictive value as

$$\mathrm{NPV}(c) = \mathbb{P}(d_i = 0 \,|\, Z_i \leq c)$$

The focus of this section is obviously related to the first two indices which are used to calculate the ROC curve.
The ROC curve displays the sensitivity against the the false positive rate, defined as one minus the specificity, for the whole range of the thresholds of the score values, i.e. for each $c \in \mathbb{R}_Z$, where $\mathbb{R}_Z$ is the space of the marker Z. Formally, the ROC curve is defined as

$$ROC(p) = \mathrm{SENS}\{(1 - \mathrm{SPEC})^{-1}(p)\}$$

where p is in $[0, 1]$ and $(1-\mathrm{SPEC})^{-1} = \inf_c\{c : (1-\mathrm{SPEC}(c) \leq p\}$. The higher the ROC curve is in the unit quadrant, the more accurate the prediction rules are. In this way, the best threshold could be defined as the value of the score that maximize the ROC curve.
The AUC, also known as c-statistic, is a summary measure of the sensitivity and specificity over the whole range of the thresholds. It is defined as

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}(p)\, dp$$

and it can be interpreted as the probability that a random diseased patient will have a greater value for a score than a random healthy patients. Given i the diseased and j the healthy subject respectively, the AUC is defined as

$$\mathrm{AUC} = \mathbb{P}(Z_i > Z_j \,|\, d_i = 1, d_j = 0)$$

The AUC is scale free and ranges between 0.5 and 1, the lowest value stays for a discrimination capability equal to flipping a coin while the highest represents an almost perfect capability of the score Z to discriminate between subjects who experience the event and those who don't. This makes the AUC an intuitive index to compare different scores, regardless of their measurement scale, in contrast to the odds ratio or other measure of efficacy.

5.1.1 Sensitivity and specificity with time-to-event endpoint

The approach described above is related to binary endpoints, some generalisations are necessary when the outcome is given by the combination of the status at a specific time-point in the future and the time to that time-point and the time-dependent ROC and AUC are used [37] [49] [10]. The event time is seen as a time-dependent binary outcome, taking the value 0 for all the time points prior to the event and the value 1 afterwards. In this way, each subject could be both healthy and diseased during the follow up and this depends only on the time-point at which the discriminative capability of the score is being evaluated. Moreover, it becomes necessary to account for the right censoring that makes the true outcome usually not observed over the whole duration of the study for every patient. At least three different couple of definitions for sensitivity and specificity proposed by Heagerty and Zheng [37] are commonly used. They depend on the manner subjects are classified as either healthy or diseased at any particular time point t.

- Cumulative Sensitivity and Dynamic Specificity: at any fixed time t each patient is classified as either healthy or diseased on the basis of his status at time t. Each subject plays the role of a healthy patient at times $t < T^*$ but of a diseased patient at later times $t \geq T^*$.

$$\text{SENS}_t^{\mathbb{C}}(c) = \mathbb{P}(Z_i > c \,|\, T_i^* \leq t)$$

$$\text{SPEC}_t^{\mathbb{D}}(c) = \mathbb{P}(Z_i \leq c \,|\, T_i^* > t)$$

- Incident Sensitivity and Static Specificity: each subject does not change the status and is classified as either a event or a non-event. Events are stratified according to the time the event occurs, non-events are defined as those subjects who are free through a fixed (static) follow-up time frame $(0, \tilde{t})$.

$$\text{SENS}_t^{\mathbb{I}}(c) = \mathbb{P}(Z_i > c \,|\, T_i^* = t)$$

$$\text{SPEC}_t^{\mathbb{S}}(c) = \mathbb{P}(Z_i \leq c \,|\, T_i^* > \tilde{t})$$

- Incident Sensitivity and Dynamic Specificity: each subject is a non-event for all $t < T_i^*$, but then plays the role of a event when $t = T_i^*$. Sensitivity measures the expected proportion of patients with a score level that exceeds the threshold c among the sub-sample of patients who have the event at time t, while specificity measures the proportion of subjects with a score level less or equal to c among those who survive beyond time t.

$$\text{SENS}_t^{\mathbb{I}}(c) = \mathbb{P}(Z_i > c \,|\, T_i^* = t)$$

$$\text{SPEC}_t^{\mathbb{D}}(c) = \mathbb{P}(Z_i \leq c \,|\, T_i^* > t)$$

After selecting a definition for the time-dependent sensitivity and specificity and choosing a relevant time-point t, ROC(t) curves and the related AUC(t) can be computed and interpreted in the same manner as in the binary context. These measures will obviously be time-varying and reflect how the accuracy of the marker evolves during follow-up. Even if the later couple of sensitivity and specificity is said to have the better properties from a mathematical point of view and it is used to derive other indexes that could be time-independent under specific assumptions, it is the less clinically relevant and intuitive while the couple cumulative sensitivity - dynamic specificity is the most commonly used in clinical practice.
Lambert and Chevret [49] remark that if on the one hand the choice of a time-point in the follow-up period could be arbitrary and not able to summarize the entire discrimination capability of the score, on the other hand the cumulative-dynamic indexes (SENS, SPEC, ROC and AUC) could provide better information when used to identify time ranges in which the score performs well or when the choice of the time point is driven by a scientific interest.
The cumulative sensitivity and dynamic specificity definition will be used along to the associated time-dependent $\text{AUC}^{\mathbb{C},\mathbb{D}}(t)$ that quantifies how well a score Z can discriminate subjects who have the event by a given time t (i.e. $T^* \leq t$) from subjects who could have the event after the given time (i.e. $T^* > t$). More formally, given two different subjects i and j, the resulting AUC is defined as

$$\text{AUC}^{\mathbb{C},\mathbb{D}}(t) = \mathbb{P}(Z_i > Z_j | T_i^* \leq t, T_j^* > t), i \neq j.$$

Several methods were proposed to assess the discriminative power in the survival context with extension to account for censoring, competing risks or marker-dependent censoring. Blanche et al. [10] did a review and comparison of the most widely used. In particular, one of the more common approaches, implemented in several statistical softwares, is based on the inverse probability of censoring weighting. It uses the probability of censoring (usually estimated by the Kaplan-Meier method) to weight the contribution of each uncensored observation by the inverse probability of remaining uncensored. In this way the uncensored observations that are very likely to have been censored are more highly weighted, and thus account for the censored values (for further details, see [10] and [104]).

5.2 The Net Reclassification Improvement

The indexes presented in the previous sections are useful to explain how well a score can classify healthy and diseased subjects. Moreover, a classification of the AUC was proposed to help the physician in the interpretation of the numeric result [39]. However, several scores, adding of element to the existing ones or new models are continuously tested to improve the prediction capability. This implies the coexistence of several methods, apparently equally good to predict the same event. Usually, the insertion of a new predictor in a fair to good model (i.e. with AUC $> 70\%$) results in a strongly statistical significant association (with a low p-value) and in a slight difference in the AUC (of about $+1\%$ or $+2\%$). This makes difficult to choose between considering and ignoring this new factor, especially when its collection is expensive in term of time, money or feasibility. To make matters worse, for a clinicians it is hard distinguish between a statistical and clinical significance, where the first could be driven by the use of a big sample of patients. Even if this issue can be addressed by a well design of the study, it is necessary to provide researchers with a useful tool to assess and quantify the improvement in risk prediction offered by a new score in addition to the AUC.
Even if their application is debated by the experts [46] [67] that suggest to be careful in the interpretation of the results, two indexes were proposed by Pencina et al. [63] to address this problem. They are the net reclassification improvement (NRI) and the integrated discrimination improvement (IDI). Due to the fact that for the NRI a practical classification was proposed [32] [63] [66], only this index will be briefly presented below and used in the analysis in Chapter 6.
Differently from the AUC which measures the discriminatory capacity of a model, the NRI index compares two different scores (divided into categories of increasing risk) in order to assess the improvement in prediction. It is based on the reclassification tables that are constructed separately for patients who experience (and for those who don't experience) the event of interest in order to quantify the correct movement in categories - upwards for events and downwards for non events - when shifting from the oldest to the newest score. An extension to a category-free NRI was also presented [64]. An example of reclassification table is shown below in Table 5.1.

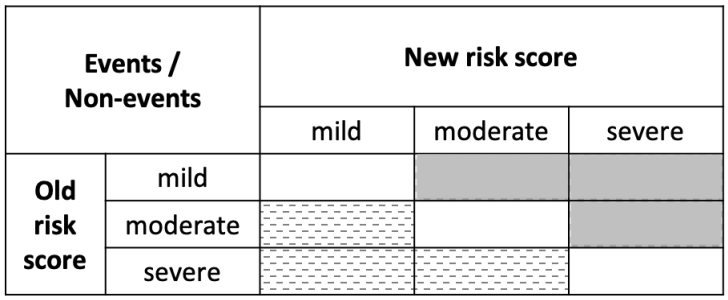

Figure 5.1: Example of reclassification table where the old and new risk scores are divided into three ordered risk categories.

In the method proposed by Pencina et al. [63], the reclassification of patients who develop and who not develop the event are considered separately. Any "upward" movement (grey cells) to an higher category for events implies an improvement in the classification while any "downward" movement (dashed cells) to a lower risk category indicates worse reclassification. The interpretation is opposite for subjects without the event. So, the improvement in reclassification can be quantified as a sum of differences in proportions of individuals moving up minus the proportion moving down for subjects with the event and the proportion of the individuals moving down minus the proportion moving up for non-events. This sum, weighted for the number of events and non-events, is called net reclassification improvement (NRI).
Consider a situation in which the predicted probabilities of the event of interest are estimated using two models that share the set of covariates, except for one new predictor. In case of two different scores, they can be standardized by using the same model (e.g. logistic model) to obtained the predicted probabilities and then these quantities can be used in the comparison. Let us categorize the predicted probabilities into a set of clinically meaningful ordinal categories of absolute risk and the cross-tabulate these two classifications. Each upward movement (up) will reflect a change into a higher category (i.e. an increase of the risk) while each downward movement ($down$) will reflect a change into a lower category (i.e. a reduction of the risk). Assuming that D is the event indicator, NRI is defined as

$$\text{NRI} = (\mathbb{P}[up|D=1] - \mathbb{P}[down|D=1]) - (\mathbb{P}[up|D=0] - \mathbb{P}[down|D=0])$$

where the four probabilities are defined as follow:

$$\mathbb{P}[up|D=1] = p_{\text{up,events}} = \frac{\#\text{events moving up}}{\#\text{events}}$$
$$\mathbb{P}[down|D=1] = p_{\text{down,events}} = \frac{\#\text{events moving down}}{\#\text{events}}$$
$$\mathbb{P}[up|D=0] = p_{\text{up,non-events}} = \frac{\#\text{non-events moving up}}{\#\text{non-events}}$$
$$\mathbb{P}[down|D=0] = p_{\text{down,non-events}} = \frac{\#\text{non-events moving down}}{\#\text{non-events}}.$$

As stated in the introduction of Section 5.2, the use of the NRI to declare a significant improvement in prediction performance is largely debated in biostatistics. We can identify at least two main drawbacks related to this index.
The first one is the problem of overfitting of the regression model used to test the association between a set of markers and an event. By using simulation it has been proved that the NRI does not provide the correct conclusion about the significance of a new marker in the model even when a large training set is used. The second issue is related to the interpretation of the index itself, both when used in categories and not. Some authors suggest to use not only the overall index that comes from the merging of events and non-events but to show both indexes. Moreover, it is commonly accepted to use the NRI, as done in this thesis, as long as the conclusions about the significant improvement in prediction performance of a new marker are also supported by the use of other methods such as the tests for the regression parameters, the change in the AUC and a clear clinical relevance of the new marker.

5.2.1 NRI for survival data

The NRI, as well as the ROC curve and the relative AUC, was born for binary outcome but it can also be applied to time-to-event outcomes. A first solution was proposed by Cook and Ridker [19] who suggested to select only subjects with a complete follow-up at a certain time point of clinical interest. The problem with this approach is that some patients will be excluded from the analysis. Another approach was developed by Steyerberg and Pencina [96]. They suggested to use the Kaplan-Meier estimator to calculate the expected number of events and non-events. Their method is proposed below. Starting from the formula used to calculate the NRI seen in the previous section and applying the Bayes theorem, it can be rewritten in an equivalent

form:

$$\text{NRI} = \frac{\mathbb{P}[\text{event}|\text{up}] \cdot \mathbb{P}[\text{up}] - \mathbb{P}[\text{event}|\text{down}] \cdot \mathbb{P}[\text{down}]}{\mathbb{P}[\text{event}]} \; + \\ + \; \frac{(1 - \mathbb{P}[\text{event}|\text{down}]) \cdot \mathbb{P}[\text{down}] - (1 - \mathbb{P}[\text{event}|\text{up}]) \cdot \mathbb{P}[\text{up}]}{1 - \mathbb{P}[\text{event}]}$$

In the survival analysis context, the quantities $\mathbb{P}[\text{event}]$, $\mathbb{P}[\text{event}|\text{up}]$ and $\mathbb{P}[\text{event}|\text{down}]$ are estimated using the Kaplan-Meier method while the proportions of people who move up and down are always available. So, the formulation of NRI is given by:

$$\text{NRI} = \frac{\mathbb{P}[\text{event}|\text{up}] \cdot n_U - \mathbb{P}[\text{event}|\text{down}] \cdot n_D}{n \cdot \mathbb{P}[\text{event}]} \; + \\ + \; \frac{(1 - \mathbb{P}[\text{event}|\text{down}]) \cdot n_D - (1 - \mathbb{P}[\text{event}|\text{up}]) \cdot n_U}{n \cdot (1 - \mathbb{P}[\text{event}])}$$

where n, n_U and n_D is the total number of subjects, the number of subject reclassified upwards and downwards respectively. The quantities in first numerator represent the expected numbers of events reclassified upwards and downwards while the expected number of non-events reclassified downwards and upwards are put in the second numerator. The two denominators represent the total expected events and non-events respectively. Moreover, the formula does not depend on the number or even the existence of predefined risk categories as it assumes probabilities of event among those reclassified upwards or downwards would be obtained pooling all subjects with the same reclassification.

This fact solved the issue related to the presence of the categories that involve an arbitrary element and a possible source of bias. Despite the use of clinically relevant categories may be useful from a clinical point of view it should be avoid and use only in presence of a categorization of the new and old risk scores that leads to the exact same interpretation. Even in this case, some authors [65] proposed and suggested the use of a category-free version of the NRI defined as NRI($>$0).

Using the definition proposed by Pencina et al. [64], denote R_{new} and R_{old} the cumulative incidences (i.e. one minus the survival probabilities) obtained from the newest and oldest survival model respectively, and assume that R_{new} and R_{old} follow a continuous distribution where any movement is considered meaningful (implying that every subject has to move either up or down, even

if slightly), for the i-th subject we obtain:

$$\mathbb{P}[R_{\text{new},i} > R_{\text{old},i} | i = event] + \mathbb{P}[R_{\text{new},i} < R_{\text{old},i} | i = event] = 1$$

More in general:

$$\mathbb{P}[\text{up}|\text{event}] + \mathbb{P}[\text{down}|\text{event}] = 1$$

that implies

$$\mathbb{P}[\text{up}|\text{event}] - \mathbb{P}[\text{down}|\text{event}] = 2 \cdot \mathbb{P}[\text{up}|\text{event}] - 1$$

Using the same approach for the non-events we obtain:

$$\mathbb{P}[\text{down}|\text{non-event}] - \mathbb{P}[\text{up}|\text{non-event}] = 1 - 2 \cdot \mathbb{P}[\text{up}|\text{non-event}]$$

From the previous derivations, the category-free NRI (NRI(>0)) can be derived as:

$$\begin{aligned}\text{NRI}(>0) &= \mathbb{P}[\text{up}|\text{event}] - \mathbb{P}[\text{down}|\text{event}] + \mathbb{P}[\text{down}|\text{non-event}] - \mathbb{P}[\text{up}|\text{non-event}] \\ &= 2 \cdot \mathbb{P}[\text{up}|\text{event}] - 1 + 1 - 2 \cdot \mathbb{P}[\text{up}|\text{non-event}] \\ &= 2 \cdot (\mathbb{P}[\text{up}|\text{event}] - \mathbb{P}[\text{up}|\text{non-event}])\end{aligned}$$

5.3 AUC and NRI in the Joint Model context

In this Section, we develop a method to assess the prediction capability of a Joint Model by using the indexes AUC and NRI described in Section 5.1.1 and 5.2.1. Both indexes are based on the evaluation on how well a score can predict an event of interest in a given time in the future. In the case of the Joint Model, this score is given by the dynamic prediction survival probabilities $\pi_i(u|t)$ with $u > t$ introduced in the Section 4.1.1. These probabilities include the contribution of following a patient from the baseline to another follow-up visit in the future. This allow the update of the model and the prediction of a patient's score whenever a new assessment of his/her marker (or markers) are available.
Rizopoulos, in Joint Model for Longitudinal and Survival Data [80] presented a derivation of the AUC in the Joint model context and an R package to calculate this quantity in practice. In particular, the function `rocJM()` [80] allows the derivation of the dynamic ROC and AUC indexes based on the incidence sensitivity and dynamic specificity seen in Section 5.1.1. This function assesses how well the Joint Model can distinguish between a patient who will experience the event of interest in a specific time window Δt having followed the longitudinal trajectory of his/her biomarker up to time t (i.e. the

subject is free from the event at time t).
However, from the point of view of this thesis, this method showed at some drawbacks which are explained briefly.

- Our interest is into the comparison of two different approaches to estimating the survival probabilities at a specific time point in the future based on two different models. We want to compare the survival probabilities estimated by using the standard approach based on the Weibull Model, with only baseline covariates, and the ones obtained from a Joint Model that used the longitudinal information collected in a specific time frame following the baseline. To assess the contribution due to the use of longitudinal data , we have to assure that the discrimination indexes are derived in the same manner under the two models except for the use of longitudinal data. Hence, the impossibility to use the method proposed by Rizopoulos that can be applied to the Joint Model but not to Weibull Model. A possible way forward could be to use a fake Joint Model where the patients contributed only with baseline data but the `rocJM()` function did not work in this way and, anyway, it was not clear if this use of the Joint Model is equivalent to the Weibull Model in practice.

- The function `rocJM()` proposed by Rizopoulos seems to be based on the incidence/dynamic definitions of sensitivity and specificity, respectively (see [80], Section 7.4.4, page 201) while we prefer to use the cumulative sensitivity and dynamic specificity as explained in Section 5.1.1.

- Even if the theory is well explained in [80], it was not clear how the function `rocJM()` works in practice.

- The application of the function `rocJM()` was not very functional in practice as the software presented various computational problems during its running and was not always able to provide results when the model changed. In particular, it has not yet been to be used with all the different parametrizations of the Joint Model (and the multivariate Joint Model) available in the R package `JM` itself.

Moreover, we need to assess the improvement due to the use of longitudinal date in deriving the survival probabilities also using the NRI. There seem to be no methods proposed in the literature.
For these reasons, we developed a new approach to compare how well the Weibull Model and the Joint Model were able to distinguish between patients who will experience an event of interest in the future and those who do not.

Following the idea underlying the approach used by Rizopoulos [80], we derived the AUC and NRI based on the predicted survival probabilities using the information collected at baseline (for the Weibull Model) and the information collected in a specific time-frame of interest following the baseline. Their derivations are presented next.

5.3.1 AUC and Joint Model

The derivation of the AUC in the Joint Model framework follows the method outlined in Section 5.1.1. It starts from the definition of a time-dependent version of cumulative sensitivity and dynamic specificity.
Let $\pi_i(u|t)$ be the dynamic survival probability for the subject $i = 1, \dots, N$ as defined in Section 4.1.1. The probability $\varphi_i(u|t) = 1 - \pi_i(u|t)$ is called dynamic cumulative incidence. For any time $u > t$, the dynamic cumulative incidence is the probability that the subject $i = 1, \dots, N$ will experience the event of interest before time u given that he/her is free from the event at time t and that we collected his/her longitudinal data $\mathcal{Y}_i = \{\mathbf{y}_i(s), 0 \leq s \leq t\}$ up to t. Let $\hat{\varphi}_i(u|t)$ the estimated cumulative incidence fot the i-th subject.

Definition 8. *The time-dependent cumulative sensitivity, having collected the longitudinal data up to $t < u$, is defined as*

$$SENS^{\mathbb{C}}_{u|t}(c) = \mathbb{P}(\hat{\varphi}_i(u|t) > c \,|\, T_i^* \leq u).$$

while the time-dependent dynamic specificity is defined as

$$SPEC^{\mathbb{D}}_{u|t}(c) = \mathbb{P}(\hat{\varphi}_i(u|t) \leq c \,|\, T_i^* > u)$$

In practice, at any fixed time point u, each patient is classified as either healthy or diseased on the basis of his/her status at time u. For our scope, we define a potential time t up to which we follow the patients regardless of whether they have all measured values in all visits up to t (i.e. missing data are possible) and whether they are actually alive at t. In this sense, even if a patient experiences the event before t the estimated dynamic cumulative incidence $\hat{\varphi}_i(u|t)$ at time $u > t$ should be higher than the one for a subject free from event up to t if the model fits well.

Definition 9. *Let i and j be two different subject, $\varphi_\cdot(u|t)$ the dynamic cumulative incidence and $T_\cdot^*$ the random variable which models the time to event where $\cdot$ stands for i or j. The resulting time-dependent AUC is defined as*

$$AUC^{\mathbb{C},\mathbb{D}}(u|t) = \mathbb{P}(\hat{\varphi}_i(u|t) > \hat{\varphi}_j(u|t) \,|\, T_i^* \leq t, T_j^* > t), i \neq j.$$

The same definition can be obtained by replacing the time-dependent cumulative sensitivity and dynamic specificity of the Definition 8 in the more general formulation seen in Section 5.1. In this way, the time-dependent ROC curve is defined as:

$$\mathrm{ROC}_{u|t}(p) = \mathrm{SENS}^{\mathbb{C}}_{u|t}\{(1 - \mathrm{SPEC}^{\mathbb{D}}_{u|t})^{-1}(p)\}$$

where p is in $[0, 1]$ and $(1 - \mathrm{SPEC}^{\mathbb{D}}_{u|t})^{-1} = \inf_c\{c : (1 - \mathrm{SPEC}^{\mathbb{D}}_{u|t}(c) \leq p\}$. With this formulation, the time-dependent AUC is given by:

$$\mathrm{AUC}^{\mathbb{C},\mathbb{D}}(u|t) = \int_0^1 \mathrm{ROC}_{u|t}(p)\, dp.$$

As seen in Section 5.1, this quantity can be weighted by using the inverse probability of censoring to weight the contribution of each uncensored observation by the inverse probability of remaining uncensored. The weighted AUC is derived by using a modified version of the definition of sensitivity. The weights are the probabilities of being uncensored when calculating the time-dependent cumulative sensitivity as follows:

$$\mathrm{SENS}^{\mathbb{C}}_{u|t}(c) = \frac{\sum_{i=1}^{N} \mathbb{I}_{(\hat{\varphi}(u|t)>c)}\, \mathbb{I}_{(T_i^* \leq u)}\, \omega_i(t)}{\sum_{i=1}^{N} \mathbb{I}_{(T_i^* \leq u)}\, \omega_i(t)}$$

where the weight ω_i is given by

$$\omega_i = \frac{1}{\mathbb{P}(C_i > t)}$$

where C is the random variable which models the censoring as defined in Section 2.2 (for further details, see [10]).

5.3.2 NRI and Joint Model

As done in Section 5.3.1, the derivation of a time-dependent NRI is based on the use of the dynamic survival probabilities $\pi_i(u|t)$ seen in Section 4.1.1. From these quantities, for each subject $i = 1, \ldots, N$ it is possible to derive the dynamic cumulative incidence $\varphi_i(u|t) = 1 - \pi_i(u|t)$ and to use it in the derivation of the category-free NRI as shown in Section 5.2.1.

Definition 10. *Let $\hat{\varphi}(u|t)_{JM}$ be the N-dimensional vectors of the estimated cumulative incidences at time $u > t$ derived from a Joint Model, where the subjects contribute with their longitudinal history in the time frame $[0, t]$. Let $\hat{\varphi}(u|0)_{WEIBULL}$ be the N-dimensional vectors of the estimated cumulative*

incidences at time $u > t$ derived from a Weibull Model, where the subjects contribute with their baseline values only. Let T_i^* be the random variable which models the time to event. The resulting category-free time-dependent NRI is defined as

$$NRI(u|t) = \frac{\sum_{i=1}^{N}(\hat{\varphi}_i(u|t)_{JM} - \hat{\varphi}_i(u|0)_{COX})\,\mathbb{I}_{(T_i*\leq u)}}{\sum_{i=1}^{N}\mathbb{I}_{(T_i*\leq u)}} - \frac{\sum_{i=1}^{N}(\hat{\varphi}_i(u|t)_{JM} - \hat{\varphi}_i(u|0)_{WEIBULL})\,\mathbb{I}_{(T_i*>u)}}{\sum_{i=1}^{N}\mathbb{I}_{(T_i*\,u)}}.$$

Note that the NRI can be used not only with the Weibull Model but with each survival model seen in Chapter 2.

5.3.3 Application in R

The applications in R in Chapter 6, are based on the use of functions coming from several statistical packages. In the application, our interest is into assess how the prediction of death at 48 months after the baseline may be improved by including the longitudinal values of several biomarkers (such as the systolic blood pressure) in a time window which spans from baseline to 6 months later. In particular, we compared a Weibull Model where only baseline values of the biomarker are inserted as covariates with a Joint Model which used all the measurements taken up to 6 months. A brief explanation of the use of each function is presented below.

The Weibull Model

For simplicity, we assume that only age, sex and systolic blood pressure (SBP) are collected. The Weibull model which explains the relationship between age, sex and SPB at baseline and the time to death can be fitted as follows

```
weibull_baseline<-survreg(Surv(time, death) ~ SBP + age + sex,
          data=base)
```

where the time-to event variable is given by `Surv(time, death)` and the dataset contains only the baseline information (`data=base`). The cumulative incidence of death for each subject at a specific time-point (e.g. 48 months) in the future is computed with the function `get.risk.survreg()` applied to the object `weibull_baseline`

```
base$risk_baseline <- get.risk.survreg(weibull_baseline, t0=48)
```

where the option `t0=48` sets the time point at which we want to calculate the cumulative incidence. The time-dependent AUC is obtained by using the function `timeROC` as follows:

```
auc_baseline <- timeROC(T=base$time,delta=base$death,
          weighting="marginal",
          marker=base$risk_baseline,cause=1,
          times=c(48), iid=T)
summary(auc_baseline)
auc_baseline$AUC
confint(auc_baseline)
```

The Joint Model

Again for sake of example, we assume that age and sex are kept constant at their baseline value while systolic blood pressure (SPB) are collected at baseline (i.e. time=0) and at 1, 3 and 6, 12, 18, 24, 36, 48 and 60 months later (SBP). The dataset `data=long` has a long shape where each subject has a row for each visit from 0 to 60 months after baseline.

```
    id SBP months
1    1 120      0
2    1 130      1
3    2 110      0
4    2 145      1
5    2 130      3
6    2 150     12
7    2 130     24
8    2 140     36
9    2  90     48
10   2 130     60
11   3 130      0
12   3 120      1
13   3 125      3
14   3 130     12
...
```

The function `jointModel()` is used to derive the Joint Model which explains the relationship between age, sex. The longitudinal trajectory of the SPB is at baseline. The longitudinal and survival submodels should be defined as two different objects.

The longitudinal submodel is defined as a linear mixed model with intercept and time both as fixed and random effects. No variance-covariance matrix for the residuals and among the random effects is specified. The function is `lme()`.

```
lmefit_sbp <- lme(SBP ~ times,
          random = ~ time | id ,
          data=long)
```

where the `id` option explains that there are repeated measurements for each patients identified by the `id` variable.
The survival submodel is fitted by using the function `coxph()` with the baseline fixed covariates (i.e. age and sex) while the longitudinal covariate must not be included. Despite the name, if no other options are specified, the function fits a Weibull Model.

```
coxfit_sbp <- coxph(Surv(time, death) ~ age + sex,
          data=base, x=TRUE)
```

Finally, the two submodels are joined in the function `jointModel()` where they are fitted again to derive the association between the longitudinal trajectory and time to death.

```
jmfit_sbp <- jointModel(lmefit_sbp, coxfit_sbp, timeVar="times")
```

Note that `time` and `times` are two different variables where the first one contains the time to event (or censoring) and the second one the time at which the longitudinal marker is collected. From the model contained in `jmfit_sbp` it is possible to derive the prediction of the dynamic survival probabilities by using the function `survfitJM()`. Before, we need to create a subset of `long` with only the data collected in the time window of interest (e.g. from baseline to month 6).

```
surv_6 <- c(rep(NA, nrow(base)))
dataset <- long[long$timess<=6,]
id<-dataset[dataset$timess==0,]$id
j <- 1
for(i in id)
{
   surv_6[j]<-unlist(survfitJM(jmfit_sbp,
          newdata = dataset[dataset$id==i,],
          survTimes = c(48), simulate=F)$summaries)[2]
   print(j)
```

```
   j <- j + 1
}
base$risk_6 <- 1-surv_6
```

In the option of the function `survfitJM()` we specified that we are interested in the estimation of the survival probabilities at 48 months after the baseline for each subject using only the data collected up to 6 months. Then we can derive the cumulative incidence as one minus the survival probabilities.

Time dependent AUC and NRI

The function `timeROC()` used to calculate the AUC under the Weibull model with the baseline covariate, is now used to derive the AUC by using the dynamic cumulative incidence calculated by using the Joint Model.

```
auc_6month <- timeROC(T=base$time,delta=base$death,
          weighting="marginal",
          marker=base$risk_6,cause=1,
          times=c(48), iid=T)
```

and the function `compare()` is used to compare the two AUC and test the difference as follows

```
compare(auc_baseline, auc_6month)
```

The function `nricens()` is used to obtain the time-dependent NRI by comparing the cumulative incidence of death at 48 months under the Weibull (`base$risk_baseline`) and the Joint model (`base$risk_6`).

```
nricens(time=base$time, event=base$death,
          p.std=base$risk_baseline, p.new=base$risk_6,
          t0=48, updown = "diff", cut=0, alpha=0.05,
          set.seed(1234))
```

This is repeated for each marker of interest.
Moreover, a multivariate Joint Model (`mvJointModelBayes()`) is also used to modelling all the markers of interest simultaneously. Some changes are required to the code, in particular in the definition of longitudinal submodel. The function `mvglmer()` allows to run a multivariate mixed model. For example, we want to model the trajectories of the systolic blood pressure (`SBP`) and the creatinine (`CREAT`). The longitudinal submodel is defined as follows:

```
lmefit_mv <- mvglmer(list
          (SBP ~ months + (months | id),
          CREAT ~ months + (months | id)),
          data = long,
          families = list(gaussian, gaussian))
```

where the option `families` allows to define which distribution each variable follows. The survival submodel is defined as in the univariate case:

```
coxfit_mv <- coxph(Surv(month, death) ~ age + sex,
          data=base, model=TRUE)
```

and the multivariate Joint Model is derived by the function `mvJointModelBayes()` that combines the previous two submodels:

```
jmfit_mv <- mvJointModelBayes(lmefit_mv, coxfit_mv,
          timeVar="months")
```

Also in this case, the AUC and the NRI are obtained using the predicted survival probabilities estimated by using the function `survfitJM()` applied to an object of the class `mvJointModelBayes()`.

5.4 Summary of chapter

This chapter presents an in-depth deep review of the methods to assess the prediction capability in a survival context. Section 5.1 and 5.2 are an introduction of the AUC and NRI indexes in the binary case, while Section 5.1.1 and 5.2.1 regard their formulation with survival data. Moreover, for the first time the use of the NRI was proposed and applied to a dynamic risk prediction score obtained by using the Joint Model in order to evaluate the gain related to the update of the longitudinal trajectory of a marker (see Section 5.3). Strengths and limitations of these two approaches are shown.

Chapter 6

A case study in prognostication of death in heart failure patients

In this chapter, an application of the Joint Model and the derivations of the dynamic risk prediction is used to assess if the use of longitudinal data is associated to an improvement in the prognostication of death in patients with reduced heart function. In particular, a comparison with the most widely used method (Weibull model for survival analysis with baseline covariates) and the Joint Model is proposed to prove how is important to include the use of longitudinal data in the calculation of the risk. From a clinical point of view, monitoring, and therefore analysing patient data in the 6 months following heart failure, can help to improve the prognosis in terms of survival probability at 48 months. Several scores have been developed over the years [15], but none of them modelled the longitudinal data directly. Usually, they are based on multivariate models with baseline covariates only. The Joint Model could be a good way to model the longitudinal trajectory of the biomarkers that are collected during the follow-up visits. The aim of this case of study is therefore twofold: on one hand, the purpose was to prove that the longitudinal data are relevant and useful in the prediction of death, on the other hand it was necessary to find an index that could directly assess how much the prediction of death could improve. To reach our objectives, we came to the decision to use the Joint Model to include the longitudinal trajectories of the biomarkers and to use two indexes common in the clinical practice, i.e. the AUC and the NRI, to assess the improvement due to this statistical model. In our study, the AUCs calculated using the Weibull and Joint model respectively are used to assess the discrimination capability of both of them while the NRI is used to quantify the gain related to the dy-

namic update of the risk prediction.
The content of this chapter is published in *Testing longitudinal data for prognostication in ambulatory heart failure patients with reduced ejection fraction. A proof of principle from the GISSI-HF database* [17].

6.1 Motivation

Several therapeutic decisions in heart failure patients with reduced ejection fraction (HFrEF) are based on life expectancy [70]. Both clinicians and patients need reliable estimates to consciously decide whether to proceed with further advanced interventions.
However, there is a significant lack of adequate prognostic tools for this specific population. Recent works have highlighted the limited accuracy of available prognostic models [15], which were generated from a wider range of heart failure patients, regardless of their left ventricular systolic function (i.e. reduced vs preserved), clinical setting (i.e. acute vs. chronic) and medical therapy (i.e. with or without implanted cardiac device) ([15], [30], [74]). In addition, most of them were created to predict all cause mortality, but the contribution of cardiovascular causes is known to be significantly greater in HFrEF than in heart failure with preserved ejection fraction (HFpEF) patients [105]. Cardiovascular parameters are therefore expected to have a greater prognostic impact in HFrEF than in HFpEF, in which the burden of comorbidities appear as the most important determinant of death [3].
Several cardiovascular parameters used to estimate prognosis in HFrEF vary considerably with time, even within few days or months of follow-up. This variation, which may be physiological, due to treatment effects and/or simply measurement error, may contribute to reduce the accuracy of prediction models. In particular, low systolic blood pressure (SBP) is an important marker of low cardiac output in HFrEF, and it has been indicated as one of the most powerful predictors of worse prognosis readily available at the patient's examination [7,8]. However, continuous titration of heart failure medications and variations in cardiac output may determine significant modifications in SBP values between visits. Thus, the use of a longitudinal collection of SBP values should better capture the "real" SBP value and potentially increase the accuracy of prognostication, particularly in those with an initially low SBP [87]-[9].
Based on the above considerations, in the attempt to design a more reliable prognostic tool for HFrEF patients who may be candidate to advanced heart failure therapeutics, we hypothesized that the use of longitudinal values of parameters with prognostic value, and particularly SBP, would determine

an increase in the accuracy of prognostic estimations in HFrEF patients. To test this hypothesis, we used longitudinal data from the *Gruppo Italiano per lo Studio della Sopravvivenza nell'Insufficienza Cardiacaâ€"Heart Failure* (GISSI-HF) study, a pragmatic trial of mainly HFrEF patients.

6.2 Methods

6.2.1 Study design, setting and participants

The GISSI-HF trial was a randomized placebo-controlled pragmatic nested trial, which was designed to investigate the effects of n-3 polyunsaturated fatty acids and rosuvastatin on mortality and morbidity in patients with clinical evidence of stable chronic heart failure [100]. Patients were enrolled between 2002 and 2005, with mandatory follow-up visits with clinical examination and blood testing at 1, 3, 6 months, and every 6 months thereafter, up to a maximum of 60 months.Median follow-up was indeed 4 years (51 months, IQR = 44-56) and last follow-up visit was completed on March 31, 2008. Patients were included irrespective of heart failure , left ventricular ejection fraction (LVEF), and age. Outcomes were adjudicated by an ad-hoc committee. All patients gave their written informed consent to the participation in the study, which was approved by the institutional review board of each participating center.

6.2.2 Sample selection

We took advantage of data being collected at 1, 3, 6-month follow-up visits, considering a 6-month time window as a clinically reasonable watchful waiting time to evaluate potential significant longitudinal changes in parameters of interest, and to allow taking decisions on whether to candidate a patient to further interventions, if needed. We started from a list of variables previously demonstrated as independent prognostic markers of all-cause mortality in the GISSI-HF population [7] (i.e. age, sex, body mass index, NYHA, diabetes mellitus, chronic obstructive pulmonary disease, SBP, HR, LVEF, creatinine, hemoglobin, uric acid). Among them, we subsequently choose five continuous candidate variables of interest, based on their widespread clinical availability, easy measurement repeatability and ample variability within few months. These parameters included SBP, HR, hemoglobin, creatinine and uric acid. From the initial 6975 patients with 66980 study visits, we therefore excluded 330 patients with 27783 visits missing at least one of these five parameters of interest. Additional 1176 patients with 6991 visits were excluded because of

baseline LVEF $\geq 40\%$, according to the aims of our study focused on HFrEF patients, leaving a final sample of 5469 patients with 32206 repeated visits (Figure 6.1).

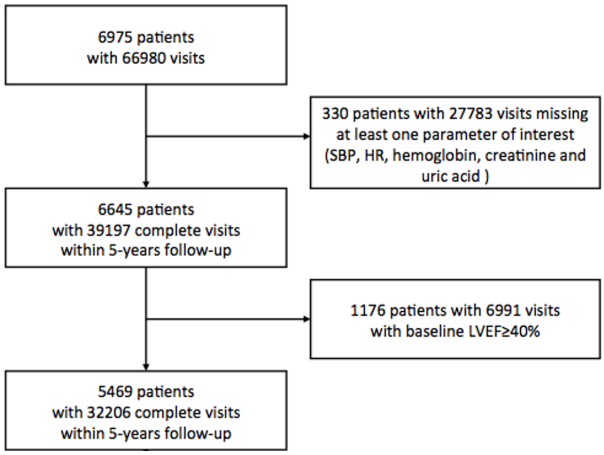

Figure 6.1: Flow chart for patient selection.

Baseline comparison between included and excluded patients revealed significant differences, mainly driven by the exclusion from our analysis of patients with a LVEF $\geq 40\%$. Excluded patients had indeed a more preserved LVEF, a lower mortality at follow-up and were more frequently older female patients, with more comorbidities, a higher SBP, a less frequent ischemic etiology and were less likely to receive beta-blockers and anti-aldosterone therapy, thus reflecting the usual characteristics of an HFpEF population.

6.3 Statistical analysis

Descriptive statistics were presented as mean and standard deviation, absolute and relative frequencies. T-test and Fisher's exact test were performed to compare the characteristics of patients alive or dead at follow-up. The coefficient of variation (i.e. $SD/mean \times 100$) was calculated for each parameter of interest (see Figure 6.2), confirming their ample variability during 6-month follow-up.

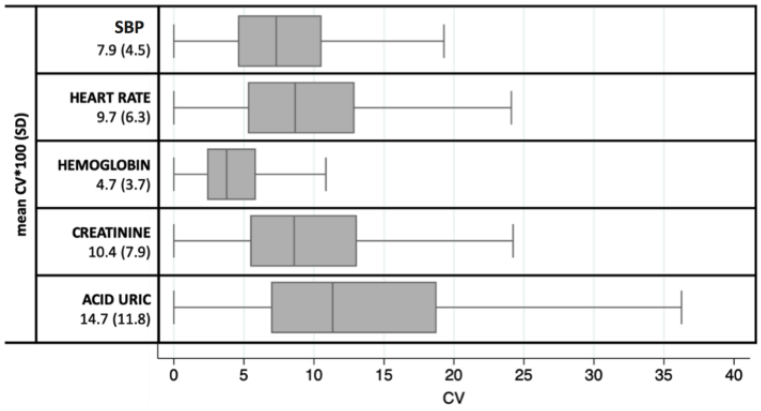

Figure 6.2: Coefficient of variations of the five parameters of interest during 6-month follow-up. For some subjects, the CV is equal to zero because the biomarker has the same value at each time point over time. This could be due to an approximation of the physician in the data entry.

6.3.1 Classical approach

After the descriptive analyses, necessary to present the composition of the sample, the statistical analysis can be divided into two steps.
In the first step, the researcher explores the variation of the longitudinal biomarkers of interest according to the survival status at the end of the follow-up. To do this, several linear mixed models (see Chapter 1) with random intercept and slope [108] were fitted for each biomarker of interest (i.e. systolic blood pressure, heart rate, haemoglobin, creatinine, uric acid). The objective was to show that their values changed significantly during the first 6 months after baseline and differently for patients who for patients who would have survived or died within 60 months. So, to ignore this change by using only the baseline value in a survival model would have involved a fairly significant loss of information because patients could have had the same biomarker value at baseline but different after 6 months. The main structure of each model for the i-th patient (with $i = 1, ..., 5469$) has the following formulation:

$$\begin{aligned}\text{BIOMARKER}_i =& \beta_0 + \beta_1 \text{TIME}_i + \beta_2 \text{DEATH}_i + \beta_3 \text{TIME}_i \times \text{DEATH}_i + \\ &+ b_{0,i} + b_{1,i} \text{TIME}_i + \psi \text{X}_i\end{aligned}$$

where

- TIME is a discrete variables that assumes values in 0,1,3 and 6 months,

- DEATH is a dichotomous variable that explains the status at the end of the follow-up (i.e. 60 months) and it is equal to 0 if patient is still alive or 1 if the patient dies during the follow-up,
- X is a matrix that contains the other covariates of clinical interest: age, sex, BMI, NYHA, diabetes, COPD, LVEF and systolic blood pressure, heart rate, haemoglobin, creatinine, uric acid at baseline when they were not the biomarker studied longitudinally,
- $\beta_0, \ldots, \beta_3, \psi$ are the fixed effects associated at each covariate,
- $b_{0,i}$ is the random intercept who explains the subject specific variation around the mean intercept β_0 for the whole sample,
- $b_{1,i}$ is the random slope who explains the subject specific variation around the mean effect of TIME estimated on the whole sample.

An example of the coefficients obtained with the statistical software STATA is shown in Figure 6.3.

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
months						
0	0	(base)				
1	.6222576	.2595137	2.40	0.016	.11362	1.130895
3	1.264383	.2745432	4.61	0.000	.7262882	1.802478
6	1.478516	.2886222	5.12	0.000	.9128273	2.044205
death						
alive	0	(base)				
dead	-2.763992	.5187404	-5.33	0.000	-3.780705	-1.74728
months#death						
1#dead	-.1286177	.48488	-0.27	0.791	-1.078965	.8217296
3#dead	-1.418392	.5208153	-2.72	0.006	-2.439171	-.3976127
6#dead	-1.308543	.5613487	-2.33	0.020	-2.408766	-.2083201
eta	.3589481	.0195049	18.40	0.000	.3207192	.397177
sex						
F	0	(base)				
M	-2.285086	.4969194	-4.60	0.000	-3.25903	-1.311142
bmi	.4735374	.044594	10.62	0.000	.3861349	.56094
nyha						
II	0	(base)				
III-IV	-1.981441	.4103211	-4.83	0.000	-2.785656	-1.177227
DIABETES						
No	0	(base)				
Yes	2.696535	.4289253	6.29	0.000	1.855857	3.537213
COPD						
No	0	(base)				
Yes	.5011661	.4733857	1.06	0.290	-.4266528	1.428985
LVEF	.3815219	.0324491	11.76	0.000	.3179228	.445121
Heart rate	.0360041	.0096594	3.73	0.000	.017072	.0549361
Hemoglobin	.7148192	.0950947	7.52	0.000	.5284369	.9012014
Creatinine	-1.577715	.3327484	-4.74	0.000	-2.22989	-.92554
Uric acid	-.4823994	.0660466	-7.30	0.000	-.6118483	-.3529504
_cons	72.01852	2.564776	28.08	0.000	66.99165	77.04538

Figure 6.3: STATA's output with the coefficients for the fixed effects of a Mixed Model to study the longitudinal evolution of the Systolic Blood Pressure (SPB). In this example, it is possible to observe that the patients who die in the first 6 months after the baseline have a mean SBP of 2.76 points less than live patients (95%CI $= -3.78, -1.74; p < 0.001$) at baseline. The alive patients have a significant increase of 0.62 points at 1 month and an increase of 1.48 point at 6 months in comparison with the baseline value. For example, the increase at 6 months in patient who die during the first 6 months can be derived as $1.47 - 1.30 = 0.17$, i.e. in the death patients the SBP does not change in mean during the first 6 months. Observing a continuous variable such as age, we can observe that 1 increase on the age involves an increase of the SBP of 0.36 points. All the values are to be considered correct (i.e. to be kept constant) for the other covariates in the model.

So defined, the model is a linear mixed model with random intercept and slope. A first-order autoregressive covariance matrix was set to model the correlation between the subsequent visits, while a matrix with only the variance components was set to model the relationship among the random effects (see Section 1.2 for more details). From these model, our interest

is in two different elements: the coefficient of the interaction term and the marginal trajectory of the marker according to the survival status. The first one is necessary to support our hypothesis that ignoring the longitudinal evolution involves a loss of information related to the change from the baseline value of the markers. The second one is clinically relevant because it allows the researcher to study the relationship between the biomarkers and survival. To facilitate the interpretation, a spaghetti plot with the marginal means (with 95% confidence intervals) adjusted to baseline values were calculated at each time point for patients alive and dead at follow-up (see Figure 6.5). The second step regards the definition of the survival models to assess how the value (or the change) of each biomarker was associated to a clinical worsening and death in a time frame of 60 months after the baseline. To do this, two different approaches were used and compared: a "traditional" approch and a "newer" one. The "traditional" approach consists of a survival model estimated using a Weibull proportional hazards model (see Section 2.2.2 and Section 2.4.1) with each biomarker of interest and covariates measured only at baseline. The choice of the parametric Weibull will be explained in the Section 6.5.4. The formulation of the Weibull Model for the i-th subject was the following:

$$\begin{aligned} h_i(t|X) =& \gamma(h_0 t)^{\gamma-1} \exp(\beta_0 + \beta_1 \text{SBP}_i + \beta_2 \text{HEART-RATE}_i + \beta_3 \text{HAEMOGLOBIN}_i + \\ &+ \beta_4 \text{CREATININE}_i + \beta_5 \text{URIC-ACID}_i + \psi \text{X} \end{aligned}$$

where

- the first part $\gamma(h_0 t)^{\gamma-1}$ defines the baseline hazard function as defined in Section 2.4.1,
- β_0 is the intercept,
- $\beta_1, \ldots, \beta_5$ are the coefficient that explain the effect of each biomarker on the hazard,
- X is a matrix that contains the other covariates of clinical interest, necessary to avoid confounding: age, sex, BMI, NYHA, diabetes, COPD, LVEF,
- ψ is the vector that contains the coefficients for the other covariates.

6.3.2 Proposed approach

The newer approach aimed at investigating the association between the longitudinal trajectory of each biomarker of interest and time to death, using

the Joint Model (JM) for longitudinal and time-to-event data [80][82], as recently performed elsewhere [116] and as define in Chapter 3. Briefly, the advanced potential of this approach is to derive an individual prediction of the survival curve from joining baseline and longitudinal information collected over time (Figure 6.4). Joint modelling combines linear mixed effect models for temporal evolution of the repeated measurements with Weibull survival models for the time-to-event data. By applying joint modelling, all biomarker candidate values were inherently corrected for different follow-up durations between patients [79][12].

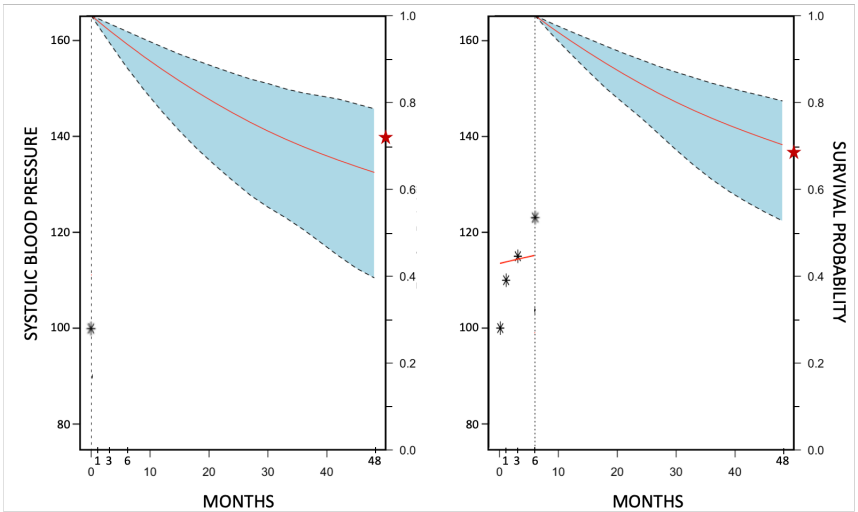

Figure 6.4: Change in prediction of 4-year survival using baseline value only vs joining baseline and longitudinal values up to 6-month follow-up. The plots depict the change in prediction of 4-year survival (right Y axis) using baseline value of systolic blood pressure only (A, left Y axis) vs joining baseline and longitudinal systolic blood pressure values up to 6-month follow-up (B, left Y axis). The increase in systolic blood pressure observed during the first 6 months of follow-up (with 3 additional values at 1, 3 and 6-month visit) determines a change in the prediction of 4-year survival (red stars), which in the example increases from 63% to 70%.

Following the definitions seen in Section 3.2, for each biomarker, the joint model is based on two different submodels: the longitudinal and the survival submodel. The longitudinal submodel for the i-th patient is defined by the following equation:

$$\text{BIOMARKER}_i = m_i = \beta_0 + \beta_1 \text{TIME} + b_{0,i} + b_{1,i} \text{TIME}$$

where the outcome variables contains all available measurements of the marker between baseline an month 60. Not only the information related to

the first 6 months are included in this step because we have the necessity to train the model on the larger amount of data available.
The survival submodel is given by:

$$h_i(t|\mathcal{M}_i(t), W_i) = h_0(t)\exp(\gamma' W_i + \alpha m_i(t))$$

where

- W is a matrix that contains the following covariates kept fixed in their baseline value: age, sex, BMI, NYHA, diabetes, COPD, LVEF and systolic blood pressure, heart rate, haemoglobin, creatinine, uric acid when they are not the biomarker studied longitudinally,
- γ is a vector that contains the coefficient that explains the effect of each covariates on the log-hazard,
- $m_i(t)$ is the value of the longitudinal biomarker at time t derived from the longitudinal submodel,
- α is the association parameter, it is the log-hazard ratio which explains the effect of a unit increase of the marker on the hazard of death.

Both the "traditional" and the "newer' model are fitted on the overall sample to obtain the respective hazard ratios. Then, the 48-month survival probabilities were calculated for each patients by using the baseline information under the traditional model and by using information collected up 1, 3 and 6 months respectively under the JM. The predicted survival probabilities at 48 months and the respective cumulative incidences are derived using the definition seen in Section 4.1.1. These estimates were used to assess the prediction accuracy of mortality at 48 months in terms of time-dependent area under the curve (AUC) (see Definition 9) and net reclassification improvement (NRI) (see Definition 10).
All the continuous variables inserted in the models are not centered to the mean value. Even if this fact could have helped in the interpretation of the intercept and the coefficient, it was not considered necessary. No variable of power greater than 1 was inserted in any model to justify the centering in order to avoid a correlation between the variable of grade one and the sqaure or the cube of the same.
Two-tailed probabilities were reported and a P-value of 0.05 was used to define nominal statistical significance. All analyses were conducted using STATA software (version 14.2, 2015, StataCorp, College Station, TX, USA) and R (version 3.5.2, 2018, R Core Team, R Foundation for Statistical Computing, Vienna, Austria).

6.4 Results

6.4.1 Sample characteristics

The study sample included 5469 patients with baseline LVEF $< 40\%$ and 32206 repeated visits with the same amount of SBP, HR, hemoglobin, creatinine and uric acid measurements. A total of 1588 died during the study (37 by one month, 110 by 3 months, 221 by 6 months, 1452 by 48 months, with an incidence rate of 8.3% per 1 year). Tables 6.1 and 6.2 shows a comparison of patients alive and dead at follow up, confirming the univariate association between each parameter of interest and mortality [7].

Variable, mean(SD)		Overall N = 5469	Alive 3881 (71.0%)	Dead 1588 (29.0%)	p-value
Age (years)		66.8 (10.5)	65.2 (10.6)	70.8 (9.1)	< 0.001
Sex (male), n (%)		4394 (80.3)	3067 (79.0)	1327 (83.6)	< 0.001
Body mass index (kg/m^2)		26.9 (4.4)	27.2 (4.4)	26.2 (4.4)	< 0.001
NYHA III-IV, n (%)		2063 (37.7)	1224 (31.5)	839 (52.8)	< 0.001
Diabetes, n (%)		1510 (27.6)	973 (25.1)	537 (33.8)	< 0.001
COPD, n (%)		1173 (21.5)	689 (17.8)	484 (30.5)	< 0.001
Smoking, n(%)	Former (>1 year)	2194 (40.1)	1470 (37.9)	724 (45.6)	< 0.001
	Never	2470 (45.2)	1791 (46.1)	679 (42.8)	
	Current	805 (14.7)	620 (16.0)	185 (11.6)	
Hypertension, n (%)		2843 (52.0)	2014 (51.9)	829 (52.2)	0.858
Ischemic etiology, n (%)		2779 (50.8)	1828 (47.1)	951 (59.9)	< 0.001
ECG, n (%)	Sinus rhythm	4019 (73.5)	3019 (77.8)	1000 (63.0)	< 0.001
	Atrial fibrillation	825 (15.1)	513 (13.2)	312 (19.7)	
	Pacemaker	625 (11.4)	349 (9.0)	276 (17.4)	
ICD implanted, n (%)		456 (8.3)	289 (7.5)	167 (10.5)	< 0.001
Pacemaker implanted, n (%)		738 (13.5)	422 (10.9)	316 (19.9)	< 0.001

Table 6.1: Descriptive statistics at baseline according to survival status at follow-up. NYHA=NewYork Heart Association functional class; COPD=chronic obstructive pulmonary disease; ICD=internal-cardioverter defibrillator; LVEF=left ventricular ejection fraction; SBP = systolic blood pressure; ACE = angiotensin-converting enzyme inhibitor; ARB = angiotensin receptor blocker; MRA = mineralocorticoid receptor antagonist. Please, it should be noted that the p-values should be considered with caution because their high significance can be due to the big sample size on which the tests are applied. A clinical interpretation of the differences between patients alive and dead should be preferred.

Variable, mean(SD)		Overall N = 5469	Alive 3881 (71.0%)	Dead 1588 (29.0%)	p-value
LVEF (%)		30.4 (6.0)	31.0 (5.8)	29.1 (6.3)	< 0.001
SBP (mmHg)		125.1 (17.6)	126.0 (17.5)	122.9 (17.7)	< 0.001
SPB ≤110	1480 (27.1%)	104.4 (6.9)	104.6 (6.8)	104.0 (7.2)	0.120
110< SBP ≤120	1268 (23.2%)	119.1 (1.9)	119.1 (1.9)	119.2 (1.9)	0.407
120< SBP ≤140	1995 (36.5%)	133.2 (5.3)	133.3 (5.3)	132.9 (5.3)	0.148
SBP >140	726 (13.3%)	155.6 (10.4)	155.6 (10.8)	155.6 (9.2)	0.945
Heart Rate (bpm)		72.5 (13.4)	72.0 (13.5)	73.9 (13.3)	< 0.001
Hemoglobin (mg/dL)		13.75 (1.64)	13.90 (1.56)	13.38 (1.76)	< 0.001
Creatinine (mg/dL)		1.21 (0.49)	1.14 (0.42)	1.37 (0.59)	< 0.001
Uric acid (mg/dL)		6.69 (2.01)	6.53 (1.90)	7.09 (2.19)	< 0.001
Diuretics, n (%)		4915 (89.9)	3393 (87.4)	1522 (95.8)	< 0.001
ACEi/ARB, n (%)		5118 (93.6)	3667 (94.5)	1451 (91.4)	< 0.001
Beta-blocker, n (%)		3637 (66.5)	2743 (70.7)	894 (56.3)	< 0.001
MRA, n (%)		2249 (41.1)	1530 (39.4)	719 (45.3)	< 0.001
Antiplatelets, n (%)		3045 (55.7)	2199 (56.7)	846 (53.3)	0.023
Allopurinol, n (%)		1229 (22.5)	709 (18.3)	520 (32.8)	< 0.001

Table 6.2: Descriptive statistics at baseline according to survival status at follow-up. NYHA=NewYork Heart Association functional class; COPD=chronic obstructive pulmonary disease; ICD=internal-cardioverter defibrillator; LVEF=left ventricular ejection fraction; SBP = systolic blood pressure; ACE = angiotensin-converting enzyme inhibitor; ARB = angiotensin receptor blocker; MRA = mineralocorticoid receptor antagonist. Please, it should be noted that the p-values should be considered with caution because their high significance can be due to the big sample size on which the tests are applied. A clinical interpretation of the differences between patients alive and dead should be preferred.

6.4.2 Association between 6-month changes in parameters of interest and mortality

Patients alive vs dead at follow-up started off with significantly better values of each parameter of interest, but also displayed a favorable trend of each one of these parameters over the first 6 months of follow-up (Figure 6.5, i.e. SBP increased, HR and uric acid decreased to a greater extent, hemoglobin and creatinine respectively decreased and increased to a lesser extent in alive

than in dead patients).

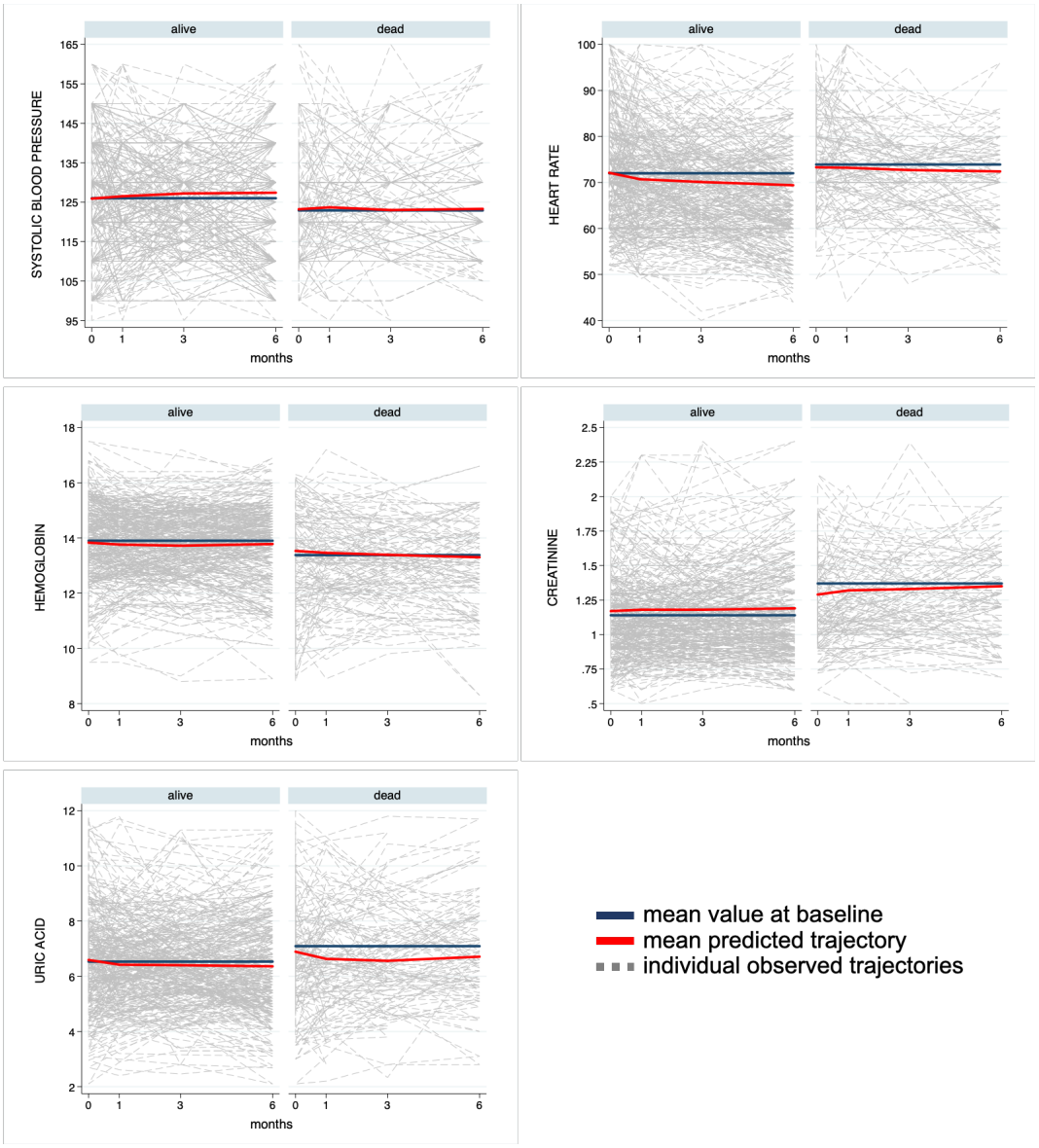

Figure 6.5: Temporal trend in five parameters of interest during the first 6 months of follow-up by survival status at last follow-up visit. Each parameter showed a favorable trend over the first 6 months of follow-up in patients alive vs. dead at 4-year follow-up. Marginal means (with 95% confidence intervals) adjusted to baseline values were calculated at each time point using linear mixed models and confirmed a significant difference in the first 6-month change of each one of these parameters between those alive and dead at 4-year follow-up, with a p-value for time by mortality interaction ≤ 0.01. See Table 6.3 for further details.

After adjustments for confounders, the 6-month time effect for each parameter was confirmed statistically significant both in the alive and dead group, with only SBP remaining substantially unchanged in the dead group (Table 6.3).More importantly, all five parameters showed a significant interaction between time and mortality, indicating that there was a significant difference in the 6-month change of each one of these parameters between those alive and dead at the end of follow-up (Figure 6.5 and Table 6.3).

		p-value for time effect	Marginal means (95%CI)				p-value for interaction patient status*time
			month 0	month 1	month 3	month 6	
SBP (mmHg)	**alive**	< 0.001	125.9 125.4 – 126.5	126.5 126.0 – 127.1	127.2 126.6 – 127.7	127.4 126.8 – 128.0	0.013
	dead	0.487	123.2 122.3 – 124.0	123.7 122.8 – 124.5	123.0 122.1 – 123.9	123.3 122.63– 124.3	
SBP ≤110 N = 1480	**alive** n = 956	<0.001	104.7 104.0 - 105.4	113.4 112.7 – 114.2	115.4 114.6 – 116.2	116.7 115.7 – 117.7	<0.001
	dead n = 524	<0.001	103.9 103.0 – 104.9	111.2 110.1 – 112.3	111.2 110.0 – 112.4	113.1 111.5 – 114.6	
110< SBP ≤120 N = 1474	**alive** n = 907	<0.001	119.1 118.5 – 119.8	122.8 122.1 – 123.5	123.8 123.0 – 124.6	125.0 123.9 – 126.1	0.004
	dead n = 361	0.020	119.2 118.2 -120.3	121.6 120.5 – 122.8	121.1 119.7 – 122.4	121.1 119.2 – 122.9	
120< SBP ≤140 N = 1995	**alive** n = 1477	<0.001	133.2 132.6 – 133.8	131.5 130.9 – 132.1	131.5 130.8 – 132.2	131.1 130.2 – 131.9	0.002
	dead n = 518	<0.001	132.8 131.8 – 133.8	129.5 128.4 – 130.6	128.3 127.1 – 129.5	127.4 125.8 – 128.9	
SBP >140 N = 667	**alive** n = 497	<0.001	155.6 154.3 – 156.8	143.0 141.7 – 144.3	142.4 141.0 – 143.8	141.0 139.3 – 142.6	0.826
	dead n = 170	<0.001	155.3 153.2 – 157.4	144.1 141.8 – 146.3	142.5 140.1 – 144.9	141.7 138.8 – 144.6	
HEART RATE (bpm)	**alive**	< 0.001	72.1 71.7 - 72.5	70.7 70.3 - 71.1	70.1 69.7 - 70.5	69.4 69.0 - 69.8	< 0.001
	dead	0.023	73.3 71.7 – 74.0	73.2 72.5 - 73.9	72.7 72.0 – 73.3	72.4 71.7 – 73.1	
HEMOGLOBIN (mg/dL)	**alive**	< 0.001	13.83 13.79 - 13.88	13.76 13.71 - 13.80	13.72 13.67 - 13.77	13.78 13.72 - 13.83	< 0.001
	dead	< 0.001	13.53 13.45 – 13.60	13.46 13.39 - 13.54	13.39 13.31 - 13.47	13.30 13.22 - 13.38	
CREATININE (mg/dL)	**alive**	0.003	1.17 1.15 - 1.18	1.18 1.16 - 1.19	1.18 1.17 - 1.19	1.19 1.17 - 1.20	< 0.001
	dead	< 0.001	1.29 1.27 - 1.31	1.32 1.30 - 1.35	1.33 1.31 - 1.35	1.35 1.33 - 1.37	
URIC ACID (mg/dL)	**alive**	< 0.001	6.59 6.54 - 6.65	6.42 6.36 – 6.48	6.40 6.34 - 6.46	6.36 6.30 - 6.43	0.005
	dead	< 0.001	6.89 6.80 - 6.99	6.63 6.54 – 6.73	6.56 6.45 – 6.66	6.71 6.59 - 6.82	

Notes: each model was adjusted age, sex, BMI, NYHA, diabetes, COPD, LVEF, heart rate, haemoglobin, creatinine, uric acid at baseline when they were not studied longitudinally

Table 6.3: Linear mixed models for longitudinal analysis of parameters of interest by patient status at follow-up. Each model was adjusted for age, sex, BMI, NYHA, diabetes, COPD, LVEF, heart rate, haemoglobin, creatinine, uric acid at baseline when they were not studied longitudinally

6.4.3 Accuracy of prognostication using baseline only vs first 6-month longitudinal data

Each parameter of interest was significantly associated with increased (or reduced) mortality, both using the traditional and the longitudinal survival model, with stronger hazard ratios for the latter one (Table 6.4).

Covariate	Traditional Model HR 95%CI p-value	Longitudinal Model HR 95%CI p-value	Prediction accuracy at 48 months expressed as AUC (95%CI) and NRI (95%CI) using information up to				
				Traditional Model	Longitudinal model		
				0 month (ref)	1 month	3 months	6 months
SBP *(x5 mmHg increase)*	0.95 0.94 – 0.97 < 0.001	0.86 0.84 – 0.88 < 0.001	AUC	75.5 (73.9 – 77.0)	75.8 (74.2 – 77.3)	76.2** (74.6 – 77.8)	77.4*** (75.9 – 78.9)
			NRI		0.15*** (0.09 – 0.22)	0.20*** (0.14 – 0.26)	0.31*** (0.26 – 0.37)
SBP ≤110			AUC	77.4 (74.6 – 80.2)	78.2* (75.5 – 81.0)	78.8** (76.1 – 81.6)	80.1*** (77.4 – 82.7)
			NRI		0.13* (0.03 – 0.24)	0.20*** (0.10 – 0.31)	0.37*** (0.26 – 0.49)
110< SBP ≤120			AUC	75.0 (71.8 – 78.3)	75.1 (71.9 – 78.4)	76.0 (72.8 – 79.3)	77.4** (74.2 – 80.5)
			NRI		-0.01 (-0.13 – 0.12)	0.13* (0.00 – 0.27)	0.30*** (0.17 – 0.42)
120< SBP ≤140			AUC	74.8 (72.1 – 77.4)	75.5 (72.8 – 78.2)	75.5 (72.8 – 78.2)	76.9*** (74.2 – 79.5)
			NRI		0.15*** (0.05 – 0.25)	0.18*** (0.09 – 0.28)	0.24*** (0.13 – 0.34)
SBP >140			AUC	69.0 (64.1 – 74.0)	68.9 (64.0 – 73.8)	68.9 (63.9 – 73.9)	70.3 (65.4 – 75.2)
			NRI		0.06 (-0.07 – 0.21)	0.05 (-0.09 – 0.19)	0.19* (0.04 – 0.35)
Heart Rate *(x5 bpm increase)*	1.03 1.01 – 1.05 0.001	1.16 1.12 – 1.19 < 0.001	AUC	75.5 (73.9 – 77.0)	75.8 (74.2 – 77.4)	76.2** (74.7 – 77.8)	77.4*** (75.9 – 79.0)
			NRI		0.08* (0.02 – 0.14)	0.13*** (0.08 – 0.19)	0.26*** (0.21 – 0.32)
Haemoglobin *(x1 mg/dL increase)*	0.92 0.89 – 0.95 < 0.001	0.83 0.80 – 0.86 < 0.001	AUC	75.5 (73.9 – 77.0)	75.7 (74.1 – 77.2)	76.0* (74.5 – 77.6)	77.2*** (75.7 – 78.7)
			NRI		0.04 (-0.02 – 0.10)	0.15*** (0.09 – 0.21)	0.22*** (0.16 – 0.28)
Creatinine *(x1 mg/dL increase)*	1.27 1.19 – 1.35 < 0.001	1.40 1.32 – 1.48 < 0.001	AUC	75.5 (73.9 – 77.0)	76.1** (74.6 – 77.7)	76.6*** (75.1 – 78.2)	77.7*** (76.2 – 79.3)
			NRI		-0.01 (-0.07 – 0.05)	0.08** (0.02 – 0.14)	0.19*** (0.13 – 0.25)
Uric Acid *(x1 mg/dL increase)*	1.06 1.03 – 1.08 < 0.001	1.18 1.13 – 1.23 < 0.001	AUC	75.5 (73.9 – 77.0)	76.0* (74.4 – 77.6)	76.5*** (74.9 – 78.0)	77.5*** (76.0 – 79.1)
			NRI		0.07* (0.02 – 0.13)	0.19*** (0.13 – 0.25)	0.27*** (0.22 – 0.34)

Table 6.4: Multivariate survival analysis. Comparison between the Traditional Model (i.e. Weibull survival model) with each biomarker of interest and covariates measured only at baseline and the Longitudinal Model (i.e. Joint Models) with each biomarker of interest collected over time. Joint models were adjusted for age, sex, BMI, NYHA, diabetes, COPD, LVEF, SBP, heart rate, hemoglobin, creatinine, uric acid at baseline when they were not studied longitudinally. Ref= reference. See Table 6.1 for abbreviations. (*) $p < 0.05$; (**) $p < 0.01$; (***) $p < 0.001$.

The comparison of the predicted survival probabilities at 48 months revealed a significant moderate increase of prediction accuracy both in term of AUC (from 75.5% with the traditional model up to 77.1%, $p < 0.001$) and NRI (0.35, $p < 0.001$) when using longitudinal values of all parameters of interest up to 6 months (Table 6.5).

Covariate	Traditional Model HR 95%CI p-value	Multivariate Joint Model HR 95%CI p-value		Prediction accuracy at 48 months expressed as AUC (95%CI) and NRI (95%CI) using information up to			
				Traditional Model	Longitudinal model		
				0 month	1 month	3 months	6 months
SBP *(x5 mmHG increase)*	0.95 0.94 – 0.97 < 0.001	0.86 0.84 – 0.88 < 0.001	AUC	75.5 (73.9 – 77.0) ref.	75.3 (73.8 – 76.9)	76.1* (74.5 – 77.7)	77.1*** (75.6 – 78.7)
Heart Rate *(x5 bpm increase)*	1.03 1.01 – 1.05 0.001	1.15 1.11 – 1.18 < 0.001					
Haemoglobin	0.92 0.89 – 0.95 < 0.001	0.84 0.81 – 0.87 < 0.001					
Creatinine	1.27 1.19 – 1.35 < 0.001	1.24 1.15 – 1.33 < 0.001	NRI	ref.	0.18*** (0.12 – 0.24)	0.28*** (0.22 – 0.34)	0.35*** (0.29 – 0.41)
Uric Acid	1.06 1.03 – 1.08 < 0.001	1.12 1.08 – 1.16 < 0.001					

Table 6.5: Multivariate survival analysis. Comparison between the Traditional Model (i.e. survival model) with each biomarker of interest and covariates measured only at baseline and the Multivariate Joint Models with each biomarker of interest collected over time and simultaneously modelled. Joint models were adjusted for age, sex, BMI, NYHA, diabetes, COPD, LVEF. Ref= reference. See Table 6.1 for abbreviations. (*) $p < 0.05$; (**) $p < 0.01$; (***) $p < 0.001$.

When 6-month longitudinal values of each parameter of interest were studied in separate JMs, holding constant values of the remaining biomarkers at baseline, a significant increase in both AUC and NRI was noticed compared

to traditional model including baseline values only (Table 6.4), particularly for SBP.

6.4.4 The case of systolic blood pressure

In order to compare the outcome of patients with similar baseline SBP value but different SBP trajectories during the first 6 months of follow-up, we performed a sub-analysis by four SBP groups (i.e. SBP $\leq$ 110, 110 $<$ SBP $\leq$ 120, 120 $<$ SBP $\leq$ 140 and SBP $>$ 140 mmHg). Indeed, the univariate association between baseline SBP and mortality were not significant within these SBP groups (Table 6.2, grey background). Nonetheless, joining the baseline and longitudinal SBP values recorded during the first 6 months of follow-up showed different trajectories of SBP in patients alive and dead at follow-up, with a greater SBP increase in patients with baseline SBP $\leq$ 110 and 110 $<$ SBP $\leq$ 120 who survived at follow-up, and a greater SBP decrease in patients with baseline 120 $<$ SBP $\leq$ 140 who died at follow-up (Figure 6.6), which were confirmed after multivariate adjustments (Table 6.3, grey background). There was no significant interaction between time and mortality in patients with baseline SBP $>$ 140 (Figure 6.6 and Table 6.3, grey background).

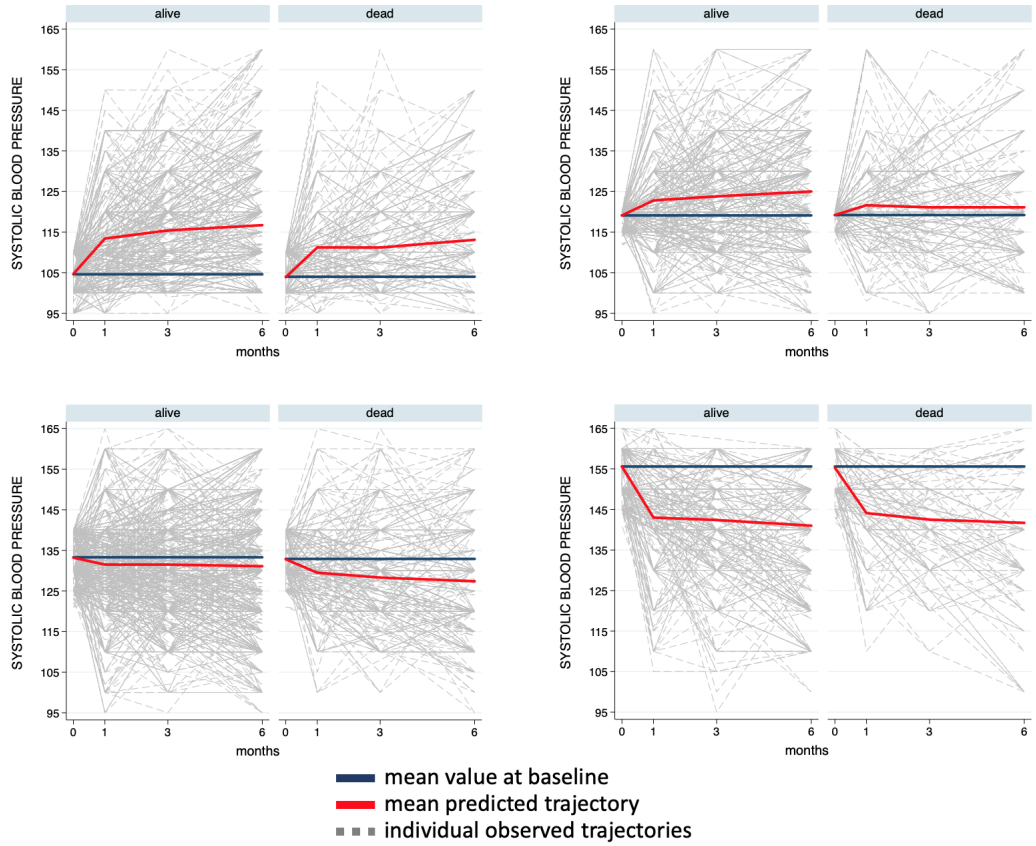

Figure 6.6: Temporal trend in systolic blood pressure during the first 6 months of follow-up by survival status at last follow-up visit. Patients were divided into 4 groups according to their SBP at baseline: SBP $\leq$ 110, 110 $<$ SBP $\leq$ 120, 120 $<$ SBP $\leq$ 140 and SBP $>$ 140 mmHg. Each parameter showed a favorable trend over the first 6 months of follow-up in patients alive vs. dead at 4-year follow-up. Marginal means (with 95% confidence intervals) adjusted to baseline values were calculated at each time point using linear mixed models and confirmed a significant difference in the first 6-month change of each one of these parameters between those alive and dead at 4-year follow-up, with a p-value for time by mortality interaction $\leq$ 0.01. See Table 6.3 for further details.

Prediction accuracy of baseline SBP decreased from its greatest (i.e. 75.8%) in the group of patients with SBP $\leq$ 110 to its least (i.e. 66.2%) in the group of patients with SBP $>$ 140 (Table 6.3, grey background). Nonetheless, even in the former group, the joining of 6-month SBP values determined a significant 2.9% increase in prediction accuracy; smaller but significant increases in AUC were found also in 110 $<$ SBP $\leq$ 120 and 120 $<$ SBP $\leq$ 140 patients, whereas no increase was noticed in SBP $>$ 140 patients. Similarly, the NRI showed a significant moderate improvement in predictive accuracy in all groups except SBP $>$ 140, confirming the greater discrimination obtained

when using a longitudinal prognostic model (Table 6.3, grey background).

6.5 Discussion

With the availability of more advanced therapeutic interventions for HFrEF patients, prognostication is becoming increasingly relevant. Concomitantly, limitations of available prognostic tools have been brought to light [15][16], and a "reboot" of prognostic modelling in heart failure advocated [26]. It has also been suggested that dynamic models using repeated measures over time would give better prediction than "one-off" measurements would give [92]. We herein took advantage of longitudinal data collected in the GISSI-HF trial to demonstrate that the use of 6-month repeated measurements of five parameters of interest (i.e. SBP, HR, hemoglobin, creatinine and uric acid) was able to improve the accuracy of 4-year all-cause mortality prediction over the use of a single measurement obtained at study enrolment.

6.5.1 Longitudinal trajectories of parameters of interest

These five parameters were chosen because of their validated independent prognostic impact, wide availability in daily clinical practice and high variability within a short period of time. Their 6-month coefficients of variation approached 10% or higher, and for SBP and HR were similar to those found in the HFrEF population of the Systolic Heart Failure Treatment with the If inhibitor ivabradine Trial (SHIFT) [11]. Acknowledging these important variations in the measurement of parameters with prognostic capability in HFrEF is of great importance. In addition, in a transversal enrolment, such as that usually undertaken in current heart failure registries and trials, the nature of the disease casually finds the patients in a fluctuating state of stability or instability within their chronic progressing condition. These aspects may contribute to the inaccuracy of available transversal prognostic models [15], and explain the existence of several "prognostic outliers" [16], who could have been erroneously judged based on a single transitory value of these parameters. Importantly, within this variability we found that 6-month trajectories of these parameters were significantly different in patients alive vs dead at final follow-up (i.e. each parameter demonstrated a significant interaction between time and mortality), supporting the rationale for the use of longitudinal data to improve accuracy of survival prediction.

6.5.2 The case of systolic blood pressure

In contrast to hypertensive patients, the concept of "reverse epidemiology" suggests that symptomatic patients with HFrEF and elevated blood pressure levels have an improved survival [43]. Thus, whereas variability of blood pressure in hypertensive patients is frequently interpreted as random fluctuation around a patient's true underlying blood pressure [60], in patients with HFrEF an increased SBP variability has been usually related to modifications in cardiac output and medications, and associated with both poorer [87] and improved outcomes [11]. Nonetheless, when HFrEF patients enrolled in the SHIFT were stratified by baseline SBP, those with low SBP at baseline and low SBP variations at follow-up had an additive deleterious effect on risk [11]. In our study, we found a significant interaction between time and mortality in the groups of patients with baseline SBP ≤ 110 and $110 < \text{SBP} \leq 120$, indicating that there was a greater increase in 6-month values of SBP change in those alive vs those dead at the end of follow-up. Accordingly, the gaining in prognostic accuracy was maximal when using longitudinal SBP values in the group of patients with baseline SBP ≤ 110. We speculate that the increase in SBP observed in these patients may be related to some recovery in LVEF. A trend toward higher risk reduction was previously found in HFrEF patients who increased their SBP to those who failed to increase their SBP 6 months after the implantation of a cardiac resynchronization device [9], further highlighting the importance of SBP augmentation as a marker of myocardial function improvement with an impact on prognosis. SBP should certainly not be considered as a substitute for measurements of LVEF, but its value lies in its comprehensiveness and simplicity, as it can be routinely measured at all outpatient visits [106], and multiple measurements in time can considerably orient physicians toward the most appropriate management of the patient.

6.5.3 Statistical modelling of longitudinal data

The change of a continuous parameter in a certain period of time(i.e. delta, usually expressed as a percent change) or the slope of the regression line derived from multiple values in time [87][12][55] have been traditionally used for easy handling longitudinal data and to be tested as predictor of outcomes. Alternatively, a continuous variable has been categorized (e.g. LVEF into LVEF $< 40\%$, 40 to 49% and $\geq 50\%$) and the subjects assigned to a "longitudinal" group based on their transition from an initial group to another during follow-up [86][8]. Joint modelling brings at least three advantages in comparison with these previous approaches. Firstly, JM handles the con-

tinuous evolutions over time of the biomarker while a time-varying Weibull or Cox model assumes unrealistically that the value remains constant between visits. Secondly, JM is more flexible, i.e. can work with biomarkers collected at different time points from a patient to another and missing values do not affect the estimation because they are automatically handled by the longitudinal sub-model [81]. Thirdly, JM takes advantage of the whole medical history of each patient, as the model is created by using all the available information, including those closest to the event (i.e. death) that are particularly valuable to draw more precise trajectories and to obtain better prognostic estimates. After creating our JM from the whole database,we arbitrarily tested its prediction accuracy by using data collected within the first 6 months of follow-up only, considering this as a clinically acceptable window of time for permitting a better estimation of 4-year survival probability. We acknowledge that about 4% of patients in the study died during these first 6 months.
To the best of our knowledge, a similar approach was previously used only by Zhang and colleagues, who demonstrated the improvement in prediction of 3-year mortality obtained by using longitudinal vs cross-sectional values of N-terminal pro B-type natriuretic peptide [116]. Nonetheless, this study and ours remain "proof of principle" studies, and further analysiswill have to be performed to validate our model in independent HFrEF cohorts, and to potentially develop a new prognostic tool to be used in the clinic, consenting the incorporation of longitudinal data. Prognostic models for specific subgroup of HFrEF patients would have to be developed [30] (e.g. ischemic vs non-ischemic HFrEF [47]), accounting also for the competitive modes of death contributing to total mortality in HFrEF, particularly pump failure vs sudden cardiac death vs. non-cardiovascular death (e.g. cancer) [16][4].

6.5.4 Strengths and limitations

Among the strengths of this analysis is the large cohort of well phenotyped heart failure patients enrolled in a pragmatic trial of substantially neutral medications (i.e. n-3 polyunsaturated fatty acids and rosuvastatin), and with several longitudinal parameters collected during a long-term follow-up. In addition, longitudinal real-world data are usually collected at the clinician's discretion, often following episodes of deterioration, whereas our trial data were recorded at prespecified time points, allowing the examination of unbiased observations. Importantly, patients with missing longitudinal variables of interest were excluded from the analysis, as well as those with preserved LVEF, according to the aims of our analysis focused on the HFrEF population only. The choice of removing patients with missing markers can

also be seen as a limitations. It can seem tendentious but it is motivated by two different aims: firstly, we needed to put ourselves in the best scenario where all markers of interest are collected, secondly, we wanted to be consistent through the entire study, using the same set of patients: the missingness of some marker would lead to regression models based on different subsets of patients. Finally, the application of the Joint Model is often affected by problem of convergence, to avoid this fact we needed to have the best and clean dataset.

Among the limitations is the retrospective nature of the analysis, performed on a database collected > 10 years ago with a utilization of disease modifying therapy (particularly beta-blocker, mineral receptor antagonists and implantable cardioverter defibrillator) lower than the one reported in contemporary real-world HFrEF European outpatients [54]. We cannot determine whether longitudinal changes in SBP were due to changes in medications or LVEF. Nonetheless, the use of HFrEF recommended medications was extremely high at enrolment, stable at 6-month follow-up (diuretics 89.6%, ACE inhibitors/angiotensin receptor blockers 93.2%, beta-blockers 71.3%, mineral receptor antagonists 40.8%), and most of these treatments could have lower and not increase SBP. As in previous analysis, variables containing information on medications were excluded from multiple analysis because of the impossibility to distinguish causality of treatment effects from confounding by indication and reverse causation [7]. Similarly, we could only speculate on the relationship between changes in SBP and LVEF; longitudinal LVEF data were lacking, thus we could not investigate this aspect any further. We finally acknowledge that the 2-3% increase observed in AUC, although statistically significant, could have little impact at the single patient level. However, this is a proof of principle study, and recent large-scale studies have reported similar AUC improvements when comparing older vs newer models [1] [15] or adding biomarkers [6].

Regarding the statistical methods, at least two drawbacks should be mentioned.

The first one regards the very small number of functions available to fit the Joint Model. We used the method and the function proposed by Rizopoulos [80] and another package was developed in STATA by Crowther [24] [25]. Some other functions are available on the web but how they work is not very clear. The package proposed by Crowther is useful in practice but the environment of STATA is less flexible and, for the purposes of our study, it does not allow for comparison between Weibull and Joint Model.

The second drawback is related to the computational issues related to the Joint Model. In general, we noticed that it requires a relative big number of repeated measurements for each subject and a good sample size. No formal

considerations are done, but in the development of this analysis we noticed than more than 4 longitudinal measurements for each subject are required and, if possible, a study design that sets the repeated measurements at fixed time points, even if the Joint Model should be implemented to manage patients with repeated measurements collected scattered over the follow-up. Another limitation is related to the computational times: if the Joint Model requires several longitudinal measurements, the increase of the sample size and the number of visits for each subject involve an increase of the computational time, specially in the calculation of the predicted survival probabilities that requires the use of simulation. If the univariate Joint Model requires from 5 to 10 minutes to run, the multivariate Joint Model takes a few hours. This fact is also related to the number of covariates inserted in the survival submodel and their distribution. Another computational limitation is related to the definition of the two submodels and the Joint Model itself. The Joint Model should allow to use several functional forms for the modelling of the time or different structure for variance/covariance matrices as well as the use of different baseline hazard function (Weibull, Cox, ...). In practice, the number of computation issues increases as the complexity of the model increase as definition of the model becomes more complex. Finally, some functions necessary to the derivation of the predicted survival probabilities are not available with all the possible combinations of the submodels for the longitudinal and survival part.

6.6 Conclusions

The findings, even if collected on a retrospective sample, proves that the use of longitudinal data over cross-sectional data can significantly improve the accuracy of survival prediction in HFrEF patients. The use of 6-month values of five continuous biomarkers, including SBP, HR, hemoglobin, creatinine and uric acid, allowed obtaining greater accuracy in prognostication than the use of baseline values only. In particular, during this window of time an increase in SBP in patients with low baseline SBP values was associated with a greater survival, potentially linked to some recovery in left ventricular function.

Limitations of available prediction scores are being increasingly acknowledged [30], and clinicians are reluctant to incorporate them in their routine clinical practice[15]. Although prognostication remains a "once and for all" moment (that is, when the patient and the physician have to take a particular decision), during this crucial moment the physician should account for the whole trajectory of the patient. In light of this principle and of our results, prog-

nostication in HFrEF becomes a dynamic process, which may be worth the waiting of few months before electing the patient to advanced therapy such as the implantation of internal cardioverter defibrillator and/or left ventricular support devices [91]. Further investigation on this matter is warranted, due to the clinical relevance of the issue and the increasing availability of longitudinal real-world data and statistical models able to comprehensively handle them for the production of more advanced prognostic tools. In general, an opportune perspective clinical trial to compare the common and the Joint Model approach is recommended to better address the problem and provide an incentive for new patient monitoring and care strategies.

Appendices

Appendix A

A proof for REML Theorem

Theorem 1. *Let $Y = X\beta + Zb + \varepsilon$ the linear mixed model obtaining by the combination of N subject-specific models $Y_i \sim N_{n_i}(X_i\beta, V_i(\alpha))$, then the REML* log*-likelihood function for $\vartheta = (\beta, \alpha)$ can be written as*

$$l_{REML}(\beta, \alpha; y) = C - \frac{1}{2} log \left| \sum_{i=1}^{N} X_i' {V_i}^{-1} X_i \right| + l_{ML}(\hat{\beta}(\alpha), \alpha; y)$$

with C a constant, and because $\left|\sum_{i=1}^{N} X_i' {V_i}^{-1} X_i\right|$ does not depend on β, it follows that the REML estimators fo α and β can also be found maximizing the REML function with respect to all parameters simultaneously.

Lemma 1. *Let A be a matrix defined as previous then*

$$U = A'Y \sim N_{n-p}(0, A'VA)$$

with $V = V(\alpha)$. Thus, the distribution of U depends on α but not β.

Proof. Because $Y \sim N_n(X\beta, V)$ we get that:

$$\mathbb{E}[U] = \mathbb{E}[A'Y] = A'\mathbb{E}[Y] = 0$$
$$\mathbb{V}[U] = \mathbb{V}[A'Y] = A'\mathbb{V}[Y]A = A'VA$$

hence:

$$U = A'Y \sim N_{n-p}(0, A'VA).$$

□

Lemma 2. *Let A a matrix defined as previous, and $G = V^{-1}X(X'V^{-1}X)^{-1}$, then*

(a) $|[A, G]|^2 = |X'X|^{-1}$

(b) $|A'VA| = |X'V^{-1}X|\,|V|\,|X'X|^{-1}$

(c) $A(A'VA)^{-1}A' = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$

Proof. Let A a matrix $n \times (n-p)$ defined as previous that satisfying

$$A'A = I_{n-p} \text{ and } AA' = I - P_X$$

and $G = V^{-1}X(X'V^{-1}X)^{-1}$.

(a)

$$\begin{aligned}
|[A,G]|^2 = |[A,G]'[A,G]| &= \left|\left[\frac{A'}{G'}\right][A,G]\right| \\
&= \left|\begin{bmatrix} A'A & A'G \\ G'A & G'G \end{bmatrix}\right| = \left|\begin{bmatrix} I & A'G \\ G'A & G'G \end{bmatrix}\right| \quad [1] \\
&= |I||G'G - G'AI^{-1}A'G| = |G'G - G'AA'G| \\
&= |G'G - G'[I - P_X]G| = |G'G - G'G + G'P_XG| \\
&= |G'X(X'X)^{-1}X'G| = |(X'X)|^{-1}
\end{aligned}$$

(b)

$$\begin{aligned}
A'VG &= A'VV^{-1}X(X'V^{-1}X)^{-1} = 0 \\
G'VG &= (X'V^{-1}X)^{-1}X'V^{-1}VV^{-1}X(X'V^{-1}X)^{-1} \\
&= (X'V^{-1}X)^{-1}(X'V^{-1}X)(X'V^{-1}X)^{-1} = (X'V^{-1}X)^{-1}
\end{aligned}$$

Hence:

$$\begin{aligned}
|[A,G]|^2|V| &= |[A,G]'||V||[A,G]| \\
&= |[A,G]'V[A,G]| \\
&= \left|\left[\frac{A'}{G'}\right]V[A,G]\right| \\
&= \left|\begin{bmatrix} A'VA & A'VG \\ G'VA & G'VG \end{bmatrix}\right| \\
&= \left|\begin{bmatrix} A'VA & 0 \\ 0 & (X'V^{-1}X)^{-1} \end{bmatrix}\right| \\
&= |A'VA||(X'V^{-1}X)|^{-1}
\end{aligned}$$

[1] The result on the determinant of a partitioned matrix gives

$$|M| = \left|\begin{bmatrix} A & B \\ B' & D \end{bmatrix}\right| = |D||A - B'D^{-1}B'| = |A||D - B'A^{-1}B'|$$

Using Lemma A.3.a we obtain

$$\begin{aligned} |[A,G]|^2|V| &= |A'VA||(X'V^{-1}X)|^{-1} \\ |(X'X)|^{-1}|V| &= |A'VA||(X'V^{-1}X)|^{-1} \\ \Longrightarrow\ |A'VA| &= |(X'X)|^{-1}|V||(X'V^{-1}X)|^{-1} \end{aligned}$$

(c) In analogy at proof of Lemma A.3.b we have:

$$([A,G]^2V)^{-1} = \begin{bmatrix} A'VA & 0 \\ 0 & (X'V^{-1}X)^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} (A'VA)^{-1} & 0 \\ 0 & X'V^{-1}X \end{bmatrix}$$

Multiplying for $[A,G]^2$ we obtain:

$$\begin{aligned} [A,G]([A,G]^2V)^{-1}[A,G]' &= [A,G]\begin{bmatrix} (A'VA)^{-1} & 0 \\ 0 & X'V^{-1}X \end{bmatrix}[A,G]' \\ V^{-1} &= A(A'VA)^{-1}A' + GX'V^{-1}XG' \\ A(A'VA)^{-1}A' &= V^{-1} - GX'V^{-1}XG' \\ &= V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1} \\ &= V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1} \end{aligned}$$

□

Now, I can use the three previous lemmas to prove the REML theorem.

Proof. The log-likelihood function for α vector of parameters of the covari-

ance matrix $V = V(\alpha)$ with respect to the sample $U = A'Y$ is given by:

$$\begin{aligned}
l(\alpha; u) &= -\frac{n-p}{2}\log 2\pi - \frac{1}{2}\log|A'VA| - \frac{1}{2}U'(A'VA)U \\
&= -\frac{n-p}{2}\log 2\pi - \frac{1}{2}\log|X'V^{-1}X| - \frac{1}{2}\log|V| - \frac{1}{2}\log|X'X|^{-1} - \frac{1}{2}y'A(A'VA)A'y \\
&= C - \frac{1}{2}\log|X'V^{-1}X| - \frac{1}{2}\log|V| - \frac{1}{2}y'A(A'VA)A'y \\
&= C - [...] - \frac{1}{2}y'A(A'VA)A'y \\
&= C - [...] - \frac{1}{2}y'(V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1})'y \\
&= C - [...] - \frac{1}{2}y'V^{-\frac{1}{2}}(I - V^{-\frac{1}{2}}X(X'V^{-\frac{1}{2}}V^{-\frac{1}{2}}X)^{-1}X'V^{-\frac{1}{2}})V^{-\frac{1}{2}}y \\
&= C - [...] - \frac{1}{2}y'V^{-\frac{1}{2}}(I - P_{V^{-\frac{1}{2}}X})V^{-\frac{1}{2}}y \\
&= C - [...] - \frac{1}{2}y'V^{-\frac{1}{2}}(I - P_{V^{-\frac{1}{2}}X})'(I - P_{V^{-\frac{1}{2}}X})V^{-\frac{1}{2}}y \\
&= C - [...] - \frac{1}{2}y'V^{-\frac{1}{2}}(I - P_{V^{-\frac{1}{2}}X})'(I - V^{-\frac{1}{2}}X(X'V^{-\frac{1}{2}}V^{-\frac{1}{2}}X)^{-1}X'V^{-\frac{1}{2}})V^{-\frac{1}{2}}y \\
&= C - [...] - \frac{1}{2}y'V^{-\frac{1}{2}}(I - P_{V^{-\frac{1}{2}}X})'(V^{-\frac{1}{2}}y - V^{-\frac{1}{2}}X(X'V^{-1}X)^{-1}X'V^{-\frac{1}{2}}V^{-\frac{1}{2}}y) \\
&= C - [...] - \frac{1}{2}y'V^{-\frac{1}{2}}(I - P_{V^{-\frac{1}{2}}X})'(V^{-\frac{1}{2}}y - V^{-\frac{1}{2}}X\hat{\beta}) \\
&= C - [...] - \frac{1}{2}(y - X\hat{\beta})'V^{-\frac{1}{2}}V^{-\frac{1}{2}}(y - X\hat{\beta}) \\
&= C - [...] - \frac{1}{2}(y - X\hat{\beta})'V^{-1}(y - X\hat{\beta}) \\
&= C - \frac{1}{2}\log|X'V^{-1}X| - \frac{1}{2}\log|V| - \frac{1}{2}(y - X\hat{\beta})'V^{-1}(y - X\hat{\beta})
\end{aligned}$$

Now we can observe that: $-\frac{1}{2}\log|V| - \frac{1}{2}(y - X\hat{\beta})'V^{-1}(y - X\hat{\beta}) \propto l_{ML}(\hat{\beta}(\alpha), \alpha; y)$ more less than a constant. Hence:

$$l_{REML}(\beta, \alpha; y) = C - \frac{1}{2}log\left|\sum_{i=1}^{N} X_i' {V_i}^{-1} X_i\right| + l_{ML}(\hat{\beta}(\alpha), \alpha; y)$$

□

Bibliography

[1] Agostoni P. et al. (2018) Multiparametric prognostic scores in chronic heart failure with reduced ejection fraction: a long-term comparison, Eur. J. Heart Fail. 20: 700-710.

[2] Akaike H. (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716-723.

[3] Alba A.C. et al. (2013) Risk prediction models for mortality in ambulatory patients with heart failure: a systematic review, Circ. Heart Fail. 6: 881-889.

[4] Ameri P. et al. (2018) Cancer diagnosis in patients with heart failure: epidemiology, clinical implications and gaps in knowledge. Eur. J. Heart Fail. 20: 879-887.

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

[6] Arzilli C. et al. (2018) N-terminal fraction of pro-B-type natriuretic peptide versus clinical risk scores for prognostic stratification in chronic systolic heart failure, Eur. J. Prev. Cardiol. 25: 889-895.

[7] Barlera S. et al. (2013) Predictors of mortality in 6975 patients with chronic heart failure in the Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico-Heart Failure trial: proposal for a nomogram. Circ. Heart Fail. 6: 31-39.

[8] Basuray A. et al. (2014) Heart failure with recovered ejection fraction: clinical description, biomarkers, and outcomes, Circulation. 129: 2380-2387.

[9] Biton Y. Et al. (2015) Inverse relationship of blood pressure to longterm outcomes and benefit of cardiac resynchronization therapy in patients

with mild heart failure: a multicenter automatic defibrillator implantation trial with cardiac resynchronization therapy long-term follow-up substudy, Circ. Heart Fail. 8: 921-926.

[10] Blanche P. et al. (2013) Estimating and comparing timedependent areas under receiver operating characteristic curves for censored event times with competing risks, Stat. Med. 32: 5381-5397.

[11] Bohm M. et al. (2016) Effect of visit-to-visit variation of heart rate and systolic blood pressure on outcomes in chronic systolic heart failure: results from the systolic heart failure treatment with the if inhibitor ivabradine trial (SHIFT) trial, J. Am. Heart Assoc. 5.

[12] Bouwens E. et al. (2019) Temporal patterns of 14 blood biomarker candidates of cardiac remodeling in relation to prognosis of patients with chronic heart failure - the bio-SHiFT study, J. Am. Heart Assoc. 8: e009555.

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

[14] Brown E. et al. (2003) A flexible B-spline model for multiple longitudinal biomarkers and survival. Biometrics 61: 64-73.

[15] Canepa M. et al. (2018) Performance of prognostic risk scores in chronic heart failure patients enrolled in the European Society of Cardiology Heart Failure long-term registry, JACC Heart Fail. 6: 452-462.

[16] Canepa M. et al. (2019) Modes of death and prognostic outliers in chronic heart failure, Am. Heart J. 208: 100-109.

[17] Canepa M., Siri G. et al. (2020) Testing longitudinal data for prognostication in ambulatory heart failure patients with reduced ejection fraction. A proof of principle from the GISSI-HF database. International Journal of Cardiology. 313: 89-96

[18] Chi Y.-Y. and Ibrahim J. (2006) Joint models for multivariate longitudinal and multivariate data. Biometrics 62: 432-445.

[19] Cook N.R. and Ridker P.M. (2009) Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. Annals of Internal Medicine 150: 795-802.

[20] Cox D. and Snell E. (1968) A general definition of residuals. Journal of the Royal Statistical Society, Series B 30: 248-275.

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

[22] Cox D. and Hinkley D. (1974) Theoretical Statistics. Chapman and Hall, London.

[23] Cox D. and Oakes D. (1984) Analysis of Survival Data. Chapman and Hall, London.

[24] Crowther M.J. et al. (2013) Adjusting for measurement error in baseline prognostic biomarkers included in a time-to-event analysis: A joint modelling approach. BMC Med Res Methodol 13.

[25] Crowther M.J. et al. (2013) Joint modeling of longitudinal and survival data. The Stata Journal 13, 165-184.

[26] Cuer B. et al. (2020) Handling informative dropout in longitudinal analysis of health-related quality of life: application of three approaches to data from the esophageal cancer clinical trial PRODIGE 5/ACCORD 17. BMC Med. Res. Methodol. 20: 223.

[27] Dempster A. et al. (1981) Estimation in covariance components models. Journal of the American Statistical Association 76: 341-353.

[28] Diggle P.J. et al. (1994) Analysis of Longitudinal Data. Oxford University Press, New York.

[29] Efron B. and Tibshirani R. (1994) An introduction to the Bootstrap. Chapman and Hall CRC Press, Boca Raton.

[30] Ferrero P. et al. (2015) Prognostic scores in heart failure - critical appraisal and practical use, Int. J. Cardiol. 188: 1-9.

[31] Fitzmaurice G. et al. (2004) Applied Longitudinal Analysis. Wiley, Hoboken.

[32] Goldman N. et al. (2016) What matters most for predicting survival? A multinational population-based cohort study, PLoS One 11: e0159273.

[33] Greenwood M. (1926) The natural duration of cancer. Reports on Public Health and Medical Subjects 33: 1-26.

[34] Greven S. et al. (2008) Restricted likelihood ratio testing for zero variance components in linear mixed models. Journal of Computational and Graphical Statistics 17: 870-891.

[35] Harville D.A. (1974) Bayesian Inference for variance components using only error contrasts, Biometrika 61: 383-385.

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

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

[38] Henderson R. et al. (2000) Joint modelling of longitudinal measurements and event time data. Biostatistics 3: 465-480.

[39] Hosmer D.W. et al. (2013) Applied Logistic Regression, Third Edition. John Wiley & Sons, New York.

[40] Hsieh F. et al. (2006) Joint modeling of survival and longitudinal data: likelihood approach revisited. Biometrics 62: 1037-1043.

[41] Huang X. et al. (2009) Latent-model robustness in joint models for a primary endpoint and a longitudinal process. Biometrics 64: 719-727

[42] Inoue E. (2018) Nricens: NRI for Risk Prediction Models With Time to Event and Binary Response Data. R Package Version 1.6.

[43] Kalantar-Zadeh K. (2004) Reverse epidemiology of conventional cardiovascular risk factors in patients with chronic heart failure, J. Am. Coll. Cardiol. 43: 1439-1444.

[44] Kalbfleisch J. and Prentice R. (2002) The Statistical Analysis of Failure Time Data. Wiley, New York.

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

[46] Kerr K.F. et al. (2014) Net Reclassification Indices for Evaluating Risk-Prediction Instruments: A Critical Review. Epidemiology 25: 114-121.

[47] Kristensen S.L. et al. (2019) Risk models for prediction of implantable cardioverter-defibrillator benefit: insights from the DANISH trial, JACC Heart Fail. 7: 717-724.

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

[49] Lambert J. and Chevret S. (2016) Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves. Statistical Methods in Medical Research 25: 2088-2102.

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

[51] Liang K.J. and Zeger S.L. (1986) Longitudinal data analysis using generalized linear models. Biometrika 73: 13-22.

[52] Lindley D.V. and Smith A.F.M. (1972) Bayes estimates for the linear model. Journal of the Royal Statistical Society, Series B 34: 1-41.

[53] Lindstrom M.J. and Bates D.M. (1988) Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures data. Journal of the American Statistical Association 83: 1014-1022.

[54] Maggioni A.P. et al. Are hospitalized or ambulatory patients with heart failure treated in accordance with European Society of Cardiology guidelines? Evidence from 12,440 patients of the ESC heart failure long-term registry. Eur. J. Heart Fail. 15: 1173-1184.

[55] Masson S. et al. (2008) Prognostic value of changes in N-terminal pro-brain natriuretic peptide in Val-HeFT (valsartan heart failure trial), J. Am. Coll. Cardiol. 52: 997-1003.

[56] Mauff K. et al. (2020) Joint models with multiple longitudinal outcomes and a time-to-event outcome: a corrected two-stage approach. Statistics and Computing volume 30: 999-1014.

[57] McCullagh P. and Nelder J. (1989) Generalize Linear Models, 2nd edition. Chapman and Hall, London.

[58] Molenberghs G. and Verbeke G. (2007) Likelihood ratio, score, and Wald tests in a constrained parameter space. The American Statistician 61: 22-27.

[59] Morrell C.H. (1998) Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. Biometrics 54: 1560-1568.

[60] Muntner P. et al. (2015) Visit-to-visit variability of blood pressure and coronary heart disease, stroke, heart failure, and mortality: a cohort study, Ann. Intern. Med. 163: 329-338.

[61] Nardi A. and Schemper M. (2003) Comparing Cox and parametric models in clinical studies, Stat. Med. 22: 3597-3610.

[62] Patterson H.D. and Thompson R. (1971) Recovery of Inter-Block Information when Block Sizes are Unequal. Biometrika 58: 545-554.

[63] Pencina M.J. et al. (2008) Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Statistics in Medicine 27: 157-172.

[64] Pencina M.J. et al. (2011) Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers, Stat. Med. 30: 11-21.

[65] Pencina M.J. et al. (2012) Interpreting Incremental Value of Markers Added to Risk Prediction Models. American Journal of Epidemiology 176: 473-481.

[66] Pencina K.M. et al. (2014) What to expect from net reclassification improvement with three categories. Statistics in Medicine 10: 4975-4987.

[67] Pepe M.S. et al (2015) The Net Reclassification Index (NRI): a Misleading Measure of Prediction Improvement Even with Independent Test Data Sets. Statistics in Biosciences 7: 282-295.

[68] Pilotto A. et al. (2008) Development and Validation of a Multidimensional Prognostic Index for One-Year Mortality from Comprehensive Geriatric Assessment in Hospitalized Older Patients. Rejuvenation Research 11: 151-161.

[69] Pinheiro J. and Bates D. (2000) Mixed-Effects Modelsin S and S-PLUS. Springer-Verlag, New York.

[70] Ponikowski P. et al. (2016) ESC guidelines for the diagnosis and treatment of acute and chronic heart failure: the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) developed with the special contribution of the Heart Failure Association (HFA) of the ESC, Eur. Heart J. 37: 2129-2200.

[71] Prentice R. (1982) Covariate measurement errors and parameter estimates in a failure time regression model. Biometrika 69: 331-342.

[72] Press W. et al. (2007) Numerical Recipes: The Art of Scientific Computing, 3rd edition. Cambridge University Press, New York.

[73] Proust-Lima C. and Taylor J. (2009) Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of post-treatment PSA: A joint modeling approach. Biostatistics 10: 535-549.

[74] Rahimi K. et al. (2014) Risk prediction in patients with heart failure: a systematic review and analysis, JACC Heart Fail. 2: 440-446.

[75] Raphael C.E. et al. (2009) Quantifying the paradoxical effect of higher systolic blood pressure on mortality in chronic heart failure, Heart. 95: 56-62.

[76] Raudenbush S.W. and Bryk A.S. (2002) Hierarchical Linear Models: Applications and data analysis methods. Thousand Oaks, Sage Publications.

[77] Rizopoulos D. et al. (2009) Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. Journal of the Royal Statistical Society, Series B 71: 637-654.

[78] Rizopoulos D. et al. (2010) Multipleimputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. Biometrics 66: 20-29.

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

[80] Rizopoulos D. (2012) Joint Models for Longitudinal and Time-to-event Data: With Applications in R, Chapman and Hall/CRC.

[81] Rizopoulos D. and Takkenberg J.J. (2014) Tools and techniques-statistics: dealing with time-varying covariates in survival analysis-joint models versus Cox models. EuroIntervention 10: 285-288.

[82] Rizopoulos D. (2016) The R, Package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC, J. Stat. Softw. 72: 1-45.

[83] Rizopoulos D. et al. (2017) Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. Biometrical Journal 59: 1261-1276.

[84] Rubin D. (1976) Inference and missing data. Biometrika 63: 581-592.

[85] Satterthwaite F.E. (1941) Synthesis of variance. Psychometrika 6: 309-316.

[86] Savarese G. et al. (2019) Prevalence and prognostic implications of longitudinal ejection fraction change in heart failure. JACC Heart Fail. 7: 306-317.

[87] Schmid F.A. et al. (2017) Prognostic value of long-term blood pressure changes in patients with chronic heart failure, Eur. J. Heart Fail. 19: 837-842.

[88] Schoenfeld D. (1982) Partial residuals for the proportional hazards regression model. Biometrika 69: 239-241.

[89] Schwarz G. (1978) Estimating the dimension of a model. Annals of Statistics 6: 461-464.

[90] Self S.G. and Liang K.Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association 82: 605-610.

[91] Shen L. et al. (2017) Declining risk of sudden death in heart failure, N. Engl. J. Med. 377: 41-51.

[92] Simpson J. and McMurray J.J.V. (2018) Prognostic modeling in heart failure: time for a reboot, JACC Heart Fail. 6: 463-464.

[93] Smith A.F.M. (1973) A general Bayesian linear model Journal of the Royal Statistical Society, Series B 35: 67-75.

[94] Snijders T.A.B. and Bosker R.J. (1999) Multilevel analysis: An introduction to basic and advanced multilevel modeling. Thousand Oaks: Sage Publications.

[95] Song X. et al. (2002) A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. Biometrics 58: 742-753.

[96] Steyerberg E.W. et al. (2010) Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 21: 128-38.

[97] Stram D.O. and Lee J.W. (1994) Variance components testing in the longitudinal mixed effects model. Biometrics 50:1171-1177.

[98] Stram D.O. and Lee J.W. (1994) Correction to: Variance components testing in the longitudinal mixed effects model. Biometrics 51: 1196.

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

[100] Tavazzi L. et al. (2008) Effect of n-3 polyunsaturated fatty acids in patients with chronic heart failure (the GISSI-HF trial): a randomised, doubleblind, placebo-controlled trial. Lancet. 372: 1223-1230.

[101] Tierney L. and Kadane J. (1986) Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association 81, 82-86.

[102] Tsiatis A. and Davidian M. (2001) A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. Biometrika 88: 447-458.

[103] Tsiatis A. and Davidian M. (2004) Joint modeling of longitudinal and time-to-event data: An overview. Statistica Sinica 14: 809-834.

[104] Uno H. et al. (2007) Evaluating prediction rules for t-year survivors with censored regression models. Journal of the American Statistical Association 102: 527-537.

[105] Vaduganathan M. et al. (2017) Mode of death in heart failure with preserved ejection fraction, J. Am. Coll. Cardiol. 69: 556-569.

[106] Ventura H.O. et al. (2017) Observations on the blood pressure paradox in heart failure, Eur. J. Heart Fail. 19: 843-845.

[107] Verbeke G. and Lesaffre E. (1997) The effect of misspecifying the random effects distribution in linear mixed models for longitudinal data. Computational Statistics and Data Analysis 23: 541-556.

[108] Verbeke G. and Molenberghs G. (2000) Linear Mixed Models for Longitudinal Data: Springer Series in Statistics.

[109] Verbeke G. and Molenberghs G. (2003) The use of score tests for inference on variance components. Biometrics 59: 254-262.

[110] Verbeke G. et al. (2008) Formal and informal model selection with incomplete data. Statistical Science 23: 201-218.

[111] West B.T. et al. (2015) Linear Mixed Models. A practical guide using statistical software. Chapman and Hall / CRC, London.

[112] Wang Y. and Taylor J. (2001) Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. Journal of the American Statistical Association 96: 895-905.

[113] Wulfsohn M. and Tsiatis A. (1997) A joint model for survival and longitudinal data measured with error. Biometrics 53: 330-339.

[114] Xu J. and Zeger S. (2001) Joint analysis of longitudinal data comprising repeated measures and time to events. Applied Statistics 50: 375-387.

[115] Ye W. et al. (2008) Semiparametric modeling of longitudinal measurements and time-to-event data - a two stage regression calibration approach. Biometrics 64: 1238-1246.

[116] Zhang J. et al. (2018) Dynamic risk stratification using serialmeasurements of plasma concentrations of natriuretic peptides in patients with heart failure, Int. J. Cardiol. 269: 196-200.