Master Thesis Econometrics

Primary Biliary Cholangitis (PBC): Joint Model Selection for Alkaline Phosphatase and Serum Bilirubin

Abstract

Primary Biliary Cholangitis (PBC) is a chronic autoimmune disease of the liver. Currently, there is no cure available for PBC and most patients are treated with Ursodeoxylic Acid (UDCA) that slows down the course (Parés et al. (2006)). Retrieving measurable indicators associated with the progression of the disease is important to evaluate the prognosis and the effect of (new) medical therapies. This thesis shows the prognostic significance of two biomarkers, Alkaline Phosphatase and Serum Bilirubin, in a joint modeling framework. Two joint models are constructed, one for each marker. Relating the true levels of the biomarker to the risk of an event (death or liver transplantation) through a weighted cumulative function shows the best fit and accuracy up to 15 years of follow-up, with exception of the joint model regarding longitudinal measurements of Serum Bilirubin at 15 years of follow-up. To validate the discriminative ability of the fitted joint model and its alternative association structures for Alkaline Phosphatase and Serum Bilirubin, the Area Under the Curve (AUC) of both models is computed. The accuracy of the fitted joint models regarding the prediction of future events is quantified by an estimate of its Prediction Error (PE). The accuracy measures show that both models are performing considerably well in the discrimination between patients and the prediction of future events.

Keywords: Primary Biliary Cholangitis (PBC), Alkaline Phosphatase, Serum Bilirubin, joint models, dynamic individual predictions, linear mixed-effects (LME) model, proportional hazards model, Ursodeoxylic Acid (UDCA)

Supervisor:
Prof. dr. Richard Paap
Erasmus School of Economics
Erasmus University Rotterdam, Rotterdam

Co-reader:
Prof. dr. Dennis Fok
Erasmus School of Economics
Erasmus University Rotterdam, Rotterdam

Supervisor:
Dr. Bettina E. Hansen
Department of Gastroenterology and Hepatology
Erasmus University Medical Centre, Rotterdam

Author:
Floris B. Hendrix
Studentnumber: 328540

2016

Erasmus University Rotterdam
Rotterdam

Contents

1 Introduction

Primary Biliary Cholangitis (PBC) is a relatively rare, chronic, autoimmune disease of the liver. This slowly progressive disease will harm the liver permanently and can lead to cirrhosis of the small bile ducts within the liver, which may cause liver cancer, liver failure or premature death (Kaplan and Gershwin (2005)). PBC has only been determined in adults (Trivedi et al. (2015)), especially women above the age of 40, where approximately 0.1% developes PBC (Selmi et al. (2011)).

Currently, the only treatment according to the guidelines of the European Association for the Study of the Liver (EASL) is Ursodeoxylic Acid (UDCA) (Lindor et al. (2009); Liver et al. (2009)). Most patients are treated with this drug that is able to slow down the progression of the disease and thereby extend patients' lifes (Parés et al. (2006)). Although it is the only approved therapy, the symptoms are not yet diminished as desired and UDCA is in particular effective in an early stage of the disease. Up to 40 % of the patients have an ineffective reaction on their treatment with UDCA and thereby an increased risk on serious liver damage, which eventually can lead to liver transplantation or premature death (Lammers et al. (2014)). Therefore, it is of great importance to identify patients in an early stage of this hepatobiliary disease. Patients with an insufficient response to UDCA could take advantage of secondary treatments (Lammers et al. (2015)). To evaluate the effect of (new) medical therapies, (early) measurable indicators associated with the failure status of the patient are needed.

Recently, Lammers et al. (2014) performed a meta-analysis on the prognostic significance of two biomarkes, Alkaline Phosphatase and Serum Bilirubin, regarding the risk of an event (death or liver transplantation) for patients diagnosed with PBC. Biomarkers are repeatedly measured indicators with a strong prognostic capability on time-to-event data. Lammers et al. (2014) adopted the proportional hazards function to show whether levels of Alkaline Phophatase and Serum Bilirubin are able to predict an event. An increase in the levels of these biomarkers is associated with an increase in the hazard ratio and thereby a detoriated health status (Lammers et al. (2014)). These meaningful changes associated with the patients' failure status appear at different times in the course of the disease. Alkaline Phosphatase already exhibits changes in an early state of the hepatobiliary disease and is essential for the diagnosis of PBC (Lindor et al. (2009)). Serum Bilirubin tends to increase in a more advanced stage of the disease.

To establish an appropriate prognosis, physicians often use risk scores based on the patients' physical health condition and laboratory test results. Although, physicians measure the biomarkers of the patients frequently, the risk scores are often based on the last available observation. This approach does not capture the rate of change in the biomarker levels, which not only differs between patients, but also differs over time for the same patient (Rizopoulos et al. (2013)). With the development of joint modeling frameworks for longitudinal and time-to-event data, all available observations of the biomarker can be related to the time-to-event and the probability of transplantation-free survival can be updated after each new measurement of the biomarker (Sène et al. (2014)); in this context Alkaline Phophatase or Serum Bilirubin.

In this thesis, the repeatedly measured levels of Alkaline Phosphatase and Serum Bilirubin are examined on their ability to predict the risk of experiencing an event (death or a liver transplantation) using a joint modeling framework. The main aim of this thesis is to find an accurate joint model in order to relate the longitudinal measurements of Alkaline Phosphatase or Serum Bilirubin to the probability of survival for patients diagnosed with PBC. Due to the fact that the two biomarkers, Alkaline Phosphatase and Serum Bilirubin, tend to show meaningful changes in different stages of the disease (Lammers et al. (2015)), their predictive performance is investigated at various points in time during follow-up. The study is conducted using the database provided by the Global PBC Study Group, which represents the most comprehensive database regarding international PBC research.

In general sense, the joint model can be distinguished in two submodels, respectively a model that describes the trajectory of Alkaline Phosphatase or Serum Bilirubin and a model that analyses the survival process of the patient. The first model, to describe the complete trajectory of the two biomarkers individually, is the linear mixed-effects model. Along with the survival function and proportional hazards function to describe the analysis of time-to-event data, it represents the two building blocks that constitute the joint model (Rizopoulos (2012b)). The parameters of the joint models are estimated by Maximum likelihood (ML) estimation using numerical optimization (EM and Quasi-Newton algorithm) and numerical integration (Gauss-Kronrod rule and pseudo adaptive Gauss-Hermite rule). Different parameterizations of the biomarkers' trajectory and baseline covariates are incorporated in the joint model to improve its fit and prognostic value (Rizopoulos (2012b)). Individual dynamic predictions are derived from the fitted joint models using a Monte Carlo simulation scheme.

To develop a valid inference on the prognostic significance of Alkaline Phophatase and Serum Bilirubin in the joint modeling framework, two accuracy measures have been adopted. The Area Under the Curve (AUC) or C-statistic is computed to validate its discriminative ability (Rizopoulos (2012b)). The AUC is an accuracy measure that indicates how well the fitted joint model is able to discriminate patients between high and low risk in experiencing an event (death or liver transplantation). How well the fitted model is able to predict future events is called calibration. The Prediction Error (PE) is used to quantify the predictive ability of different parameterizations of the joint model (Henderson et al. (2002)).

Recent studies of cancer recurrence (Proust-Lima and Taylor (2009)), oartic stenosis (Andrinopoulou et al. (2015)) and treatment duration on HIV patients (Brilleman et al. (2016)) show the meaningful contribution of a joint modeling framework in the development of a prognosis. The model is able to accomodate all available longitudinal observations of a biomarker and update its survival probability at every new observation. In the context of Primary Biliary Cholangitis (PBC), the prognostic significance of Alkaline Phosphatase and Serum Bilirubin regarding the patients' failure status (death or liver transplantation) is concluded using Cox proportional hazards function (Lammers et al. (2014)). At different time points the last available biomarker observation is used to compute its association with the hazard ratio. Using only the last available observation of the biomarker obviously discards useful information about the biomarker and its trajectory. This is the first time the joint modeling framework is adopted to estimate the risk of an event for patients diagnosed with PBC to include all available information of the two biomarkers. This study provides physicians a more flexible and dynamic model to discriminate patients using Alkaline Phosphatase and Serum bilirubin. Improving the knowledge about the course of PBC and its biomarkers is essential for the development and approval of new therapies.

Corresponding to the literature, the fitted joint model is able to show the prognostic significance of Alkaline Phosphatase and Serum Bilirubin and their ability to predict clinical outcomes (death or liver transplantation). Increased levels of both biomarkers are associated with a higher risk of an event. Including all available information in the Cox proportional hazards function by using a weighted cumulative parameterization of the biomarker measurements provides the best joint model up to 15 years of follow-up according to its goodness-of-fit statistics and accuracy measures. Except for the joint model of Serum Bilirubin at 15 years of follow-up, where the slope of the biomarker trajectory takes a more important role in the prediction of an event and its discriminative ability. Due to the general incline of the biomarker values, the height of the biomarker contains less prognostic strength regarding the patients failure status compared to the biomarkers' slope.

The accuracy of the fitted joint models of the two biomarkers, Alkaline Phosphatase and

Serum Bilirubin, are declining over time. The general incline of the biomarker levels makes it harder to discriminate between patients and to predict future events as time progresses. The model involving levels of Alkaline Phosphatase is less accurate in both discimination and calibration compared to the joint model of Serum Bilirubin. This is due to the fact that the levels of Alkaline Phosphatase show meaningful changes during the whole spectrum of the disease, while Serum Bilirubin exhibits elevated levels in a more advanced stage of the disease (Lammers et al. (2015)).

In addition to the longitudinal data of the biomarkers, different baseline covariates are included in the survival functions. In the joint model with the longitudinal measurements of Serum Bilirubin, the level of Albumine times the Upper Limit of Normal (ULN) and the platelet count per 109/L at 1 year of follow-up and the age at the beginning of the follow-up are significant baseline covariates associated with the risk of an event. The levels of Albumine are multiplied by the Upper Limit of Normal (ULN) to correct for different use of normal values in various laboratories. The platelet count is reported following the standard measure of 109/L. A decrease in Albumine and platelet count at 1 year of follow-up is associated with an increased risk of having an event. In addition, older patients at the beginning of the follow-up tend to experience an event earlier. The joint model including the trajectory of Alkaline Phosphatase extends the set of significant baseline covariates with the logarithmic scaled level of Serum Bilirubin at 1 year of follow-up. A higher logarithmic scaled level of Serum Bilirubin at 1 year of follow-up leads to an incline in risk of an event. The logarithmic scaled level of Serum Bilirubin at 1 year of follow-up is uncorrelated with the trajectory of Alkaline Phosphatase due to the later appearance of meaningful changes in Serum Bilirubin.

This thesis contributes the vast majority in literature in confirming the prognostic significance of Alkaline Phosphatase and Serum Bilirubin. The accuracy measures correspond to earlier studies supporting the fact that elevated levels of Alkaline Phosphatase are spread over the complete course of PBC, while Serum Bilirubin shows meaningful changes in an advanced stage of the disease. The significant baseline covariates (Albumine times the ULN at 1 year of follow-up, age at the beginning of the follow-up, logarithmic scaled Serum Bilirubin times the ULN at 1 year of follow-up and platelet count per 109/L at 1 year of follow-up) are already frequently used as components of risk scores like the Globe score (Lammers et al. (2015)). Levels of Alkaline Phosphatase and Serum Bilirubin could be used as surrogate end points in research on new drugs as suggested by Lammers et al. (2014) and the selected joint model provides physicicians a more dynamic and flexible model to develop a diagnosis for Primary Biliary Cholangitis (PBC).

The outline of this paper is as follows. Section 2 gives a review of the relevant literature regarding the joint modeling framework. In section 3, the methodology is presented, discussing the components to formulate the joint log likelihood function, which is essential for estimating the joint model parameters. In addition, it explains the different parameterizations and how the fitted joint model is used to derive dynamic individual predictions. Section 3 ends with the explanation of the accuracy measures. In Section 4, the data provided by the Global PBC Study Group is explained in detail, followed by Section 5, where the estimation and validation results are discussed. Section 6 concludes the thesis and provides a range of possible future studies.

2 Literature Review

Alkaline Phosphatase is an early discovered biomarker used for the prediction of clinical outcomes (death or liver transplantation) in hepatobiliary research (Warnes (1972)). Shapiro et al. (1979) state that levels of Serum Bilirubin constitute a predictive indicator of the prognosis in Primary Biliary Cholangitis (PBC). Recently, the prognostic significance of Alkaline Phosphatase and Serum Bilirubin is evaluated in a meta-analysis conducted by Lammers et al. (2014). They concluded both biomarkers to be correlated with the clinical outcomes of patients diagnosed with PBC.

To link time-dependent endogenous covariates like Alkaline Phosphatase and Serum Bilirubin to the risk of an event, a joint model is developed and extensively decribed in literature (Faucett and Thomas (1996); Wulfsohn and Tsiatis (1997); Henderson et al. (2000); Tsiatis and Davidian (2004)). As described earlier in the introduction, the joint model consists of a longitudinal process and a survival process. The correlated continuous observations of biomarkers are captured by a linear mixed-effects model, where the correlation of the longitudinal responses is assumed to be captured by their shared random effects (Harville (1977); Laird and Ware (1982); Verbeke and Molenberghs (2009)). Because Maximum Likelihood (ML) estimation and Restricted Maximum Likelihood (REML) in general lead to no closed-form solutions, Lindstrom and Bates (1988) introduce the use of numerical optimization explaining the Expectation-Maximization (EM) algorithm and the Quasi-Newton Algorithm in the context of the linear mixed-effects model. The estimation of the random effects covariance matrix is at that moment still a numerical burden. To ensure the optimization leads to a positive semi-definite matrix for the random effects covariance matrix, Pinheiro and Bates (1996) introduced five unconstrained parameterizations for the optimization of the covariance matrix, respectively Cholesky, Log-Cholesky, Spherical, Matrix-Logarithm and the Givens parameterization. The Log-Cholesky and Spherical parameterization lead to the best predictive performances. Problems occuring when selecting the best linear mixed-effects model estimated with REML is examined by Gurka (2006). Selection of linear mixed-effects models could benefit from comparing the likelihoods using both REML and ML.

To relate time-independent covariates or time-dependent endogenous covariates, most statisticians use the popular proportional hazards model, introduced by Cox (1972). The model assumes these covariates have a multiplicative effect on its hazard ratio. Tsiatis et al. (1995) proposed a framework to relate the proportional hazards function to the linear mixed-effects model through shared random effects. However, every time an event occured the linear mixed-effects model has to be estimated and incorporated in the joint model, computing the same amount of linear mixed-effects models as events occured in the sample. To overcome this computational burden numerical optimization methods are adopted to maximize a joint log likelihood function (Wulfsohn and Tsiatis (1997)). Even then, the estimation of the complete joint log likelihood often lead to no closed-form solution and it remained computational intensive. Rizopoulos (2012b) developed a numerical integration technique to speed up convergence. The pseudo adaptive Gauss-Hermite rule estimates the integrals of the joint log likelihood and improves the computational intensive EM algorithm.

In contrast to the maximum likelihood estimation, Faucett and Thomas (1996) introduced a Bayesian approach using Markov Chain Monte Carlo (MCMC). The specification of an appropriate joint log likelihood is the key component in the Bayesian approach as well. The literature is inconclusive on the use of either two approaches.

This thesis uses the joint modeling framework that is proposed by Tsiatis et al. (1995) in combination with the numerical optimization technique developed by Rizopoulos (2012b). The Monte Carlo scheme used to develop the individual dynamic predictions and to validate the two models of Alkaline Phosphatase and Serum Bilirubin is recently introduced by Proust-Lima and Taylor (2009) and extended to compute accuracy measures for the joint modeling framework (Rizopoulos (2012b)). These methods and techniques are conducted to select a joint model for both Alkaline Phosphatase and Serum Bilirubin and to create a more dynamic and flexible environment for physicicians to develop a diagnosis and to test new medicines regarding Primary

Biliary Cholangitis (PBC).

3 Methods

3.1 Joint Model

The joint model is a framework that relates the longitudinal measurements of the biomarkers, Alkaline Phosphatase and Serum Bilirubin, to the time-to-event data through shared random effects that capture the association between the two processes. A linear mixed-effects model is used to describe the longitudinal trajectory of the biomarkers, while the time-to-event process is captured by the Cox proportional hazards function and the survival function. However, the Cox proportional hazards function is not able to incorporate the observed levels of the two biomarkers directly, because they represent time-dependent endogenous covariates (Rizopoulos (2012b)). To relate the two time-dependent endogenous covariates, Alkaline Phosphatase and Serum Bilirubin, to the hazard ratio, the model should accomodate for the special features of these prognostic biochemical variables.

The joint modeling framework for longitudinal and time-to-event data accounts for three features of the biomarkers. First of all, time-dependent endogenous biomarkers are conditionally depending on the existence of the patient. Logically, no measurements of Alkaline Phosphatase and Serum Bilirubin are obtained after the patient is deceased. So, the model should account for censoring. Secondly, the proportional hazards model assumes covariates to contain no error term, while the observed responses of the biomarkers in the dataset contain measurement errors. In addition, the complete longitudinal history is not known due to the irregular visits of the patients and the absence of continuous time-series of the biomarkers. To reconstuct the complete history of the biomarker and to find the true and unobserved biomarker value without an error, a linear mixed-effects model is adopted (Laird and Ware (1982); Harville (1977); Verbeke and Molenberghs (2009)). The linear mixed-effects model constitutes the first building block of the joint model and allows one to model the patient-specific trajectory of the biomarkers Alkaline Phosphatase and Serum Bilirubin.

The survival process is captured by the survival function and the Cox proportional hazards function and constitute the second building block of the joint model. This process analyses the time until a patient experiences an event. In the context of this study, an event has occured if a patient is deceased or received a transplant of the liver; these are considered clinical endpoints. The time-to-event is unknown if the patient is still alive after follow-up or if the patient is dropped out earlier. In this case, the time-to-event is censored and no event has occured. The association between the two building blocks, the longitudinal and survival process, is described by their shared random effects. The shared random effects approach allows to estimate the parameters of a joint log likelihood function using full Maximum Likelihood (ML).

The rest of the methodology is organized as follows. In the next subsection, the linear mixed-effects model is discussed regarding the analysis of the longitudinal responses and to find its continuous trajectory, followed by an explanation of missingness in the longitudinal data of the two biomarkers. Section 3.1.3 explains the second building block of the joint model, where the true and unobserved levels of Alkaline Phosphatase and Serum Bilirubin are introduced in the survival submodel. Section 3.1.4 discusses the involvement of censoring in the survival process and how this thesis accomodates for this feature. Section 3.2 shows how the joint model parameters are estimated using Maximum Likelihood (ML) estimation and describes alternative association structures between the two biomarkers and its corresponding hazard ratio. Section 3.3 shows the ability of deriving dynamic individual predictions from the fitted joint model. Finally, two accuracy measures are put forth to validate the prognostic significance of the joint models for Alkaline Phosphatase and Serum Bilirubin.

3.1.1 Longitudinal Submodel

The trajectory of the time-dependent endogenous covariates, Alkaline Phosphatase and Serum Bilirubin, is analysed with a linear mixed-effects model. The observed longitudinal measurements of the biomarkers, denoted by y_i, are expected to be correlated for the same patient. When the longitudinal observations are correlated, standard statistical tools like a t-test or a simple linear regression are not appropriate to use (Rizopoulos (2012b)). The linear mixed-effects model accomodates for correlated data through its random effects b_i (Rizopoulos (2012b)). Given the random effects, the observed levels of the biomarkers y_i are assumed to be independent and allows one to conduct a longitudinal data analysis with standard statistical tools. This is called the conditional independence assumption and is given by

$$p(y_i|b_i) = \prod_{j=1}^{n_i} p(y_i(t_{ij})|b_i), \tag{1}$$

where j denotes the number of observations $j = 1, 2, ..., n_i$ for patient i. The time in years is denoted by t and start at the beginning of the follow-up study.

The linear mixed-effects model is not only capable of deriving the average changes in the population, but also allows one to obtain the patient-specific response trajectory in the biomarker levels. This is one of main reasons to adopt the linear mixed-effects model. The patient-specific mean response profile over time for each patient in the population is descibed by the fixed effects, denoted by β, and the random effects, denoted by b_i. The fixed effects β are the average elevation in the observed biomarker values in the population, caused when one of the corresponding covariates has increased *ceteris paribus* by one. The random effects b_i indicate the patient-specific discrepancy with a subset of the fixed effects parameters. Hence, the linear mixed-effects model is able to capture predictions for the population with β and the individual patients through $(\beta + b_i)$ and is given by

$$\begin{aligned} y_i(t) &= m_i(t) + \epsilon_i(t); \qquad &(2)\\ m_i(t) &= x_i^T(t)\beta + z_i^T(t)b_i; &(3)\\ b_i &\sim N(0, D); &(4)\\ \epsilon_i(t) &\sim N(0, \sigma^2 I_{n_i}), &(5) \end{aligned}$$

where $y_i(t)$ are the observed levels of the biomarker Alkaline Phosphatase or Serum Bilirubin for patient i at time point t, which denotes the time in years that starts at the beginning of follow-up. $x_i^T(t)$ and $z_i^T(t)$ are the independent design vectors of respectively the fixed effects and the random effects for patient $i = 1, 2, ..., N$. In this thesis different parameterizations of time are used for the vectors $x_i^T(t)$ and $z_i^T(t)$. The patients are assumed to be randomly sampled from the population, what implies that the patient-specific regression parameters b_i are randomly sampled as well. The random effects coefficient vector b_i is therefor assumed to be normally distributed with a mean zero and a covariance matrix D. This assumption becomes more robust, when the number of observations per patient increases (Rizopoulos et al. (2008)). $m_i(t)$ are the true and unobserved levels of the biomarkers without a measurement error. In addition, the error terms are denoted by ϵ_i and are assumed to be normally distributed with a mean zero and a covariance matrix $\sigma^2 I_{n_i}$, where n_i is the patient-specific total number of observations $j = 1, 2, ..., n_i$. The random effects and the error terms are assumed to be uncorrelated. The linear mixed-effects model reconstructs the continuous response trajectory of the two biomarkers, Alkaline Phosphatase and Serum Bilirubin, assuming the marker values not to be constant between visits.

In section 3.2, the maximization of the joint log likelihood function is discussed. It is shown that the maximization is conducted using a numerical optimization method; the Expectation-Maximization (EM) algorithm. To obtain initial parameters for this iterative optimization method, the linear mixed-effects model and the Cox proportional hazards model are estimated individually. So, both models are first estimated seperately, not in a joint modeling framework, to compute the initial parameter values for the EM algorithm.

For the estimation of the initial parameter values for the EM algorithm, the linear mixed-effects model is estimated using Restricted Maximum Likelihood (REML). The general Maximum Likelihood (ML) estimation leads to biased estimates of the variance components, because it does not account for the fact β is estimated (Harville (1977)). The basic idea of Restricted Maximum Likelihood (REML) estimation is that the likelihood function is adapted in a way the function is not depending on the estimated $\hat{\beta}$ and its log likelihood is based on $n-p$ degrees of freedom.

$$\begin{aligned}\ell(\theta_b, \sigma^2) = -\frac{n-p}{2} log(2\pi) + \frac{1}{2} log\left|\sum_{i=1}^{n} X_i^T X_i\right| - \frac{1}{2} log\left|\sum_{i=1}^{n} X_i^T V_i^{-1} X_i\right| \\ -\frac{1}{2}\sum_{i=1}^{n}\{log|V_i| + (y_i - X_i\hat{\beta})^T V_i^{-1}(y_i - X_i\hat{\beta})\},\end{aligned} \tag{6}$$

where $\hat{\beta}$ is replaced by the Generalized Least Squares (GLS) estimator:

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i^T V_i^{-1} X_i\right) \sum_{i=1}^{n} X_i^T V_i^{-1} y_i, \tag{7}$$

where n is the number of observed measures of Alkaline Phosphatase or Serum Bilirubin. p is the dimensionality of vector x_i, which represents a patient-specific vector of the design matrix X_i. V_i is the covariance matrix of the marginal model[1], where $V_i = Z_i D Z_i^T + \sigma^2 I_{n_i}$. $\hat{\beta}$ is the estimated value of the fixed effects coefficient.

In general, the maximization of the log likelihood function is not leading to a closed-form solution. To maximize the log likelihood function (6), one needs to use numerical optimization. Here, the Quasi-Newton method is used to obtain the optimization of equation (6) (Lindstrom and Bates (1988)). Once covariance matrix V_i is estimated, fixed effects coefficient β could be obtained by the Generalized Least Squares (GLS) estimator. The inverse covariance matrix V^{-1} is replaced by its estimate. The variances of the fixed effects parameters are derived by the computation of the variance equation of the Generalized Least Squares (GLS) estimator, given by

$$v\hat{a}r(\hat{\beta}) = \left(\sum_{i=1}^{n} X_i^T \hat{V}_i^{-1} X_i^T\right). \tag{8}$$

[1] Marginal model is written by $y_i = X_i\beta + \epsilon_i$, where $\epsilon_i \sim N(0, Z_i D Z_i^T + \sigma^2 I_{n_i})$

To ensure the covariance matrix to be positive semi-definite, the log Cholesky parameterization is applied to estimate the covariance matrix D (Pinheiro and Bates (1996)). This unconstrained estimation requires the logarithmic values of the diagonals of the Cholesky decomposition matrix[2], denoted by L, to be positive. This results in a uniquely defined upper triangle matrix L and a positive semi-definite covariance matrix D. Positive semi-definite is defined by $Z^T DZ \geq 0$.

In order to capture the trajectory of the biomarkers, eight different parameterizations of the linear mixed-effect model are proposed and described in section 5.1. The models are compared using the log likelihood statistic when they are nested, otherwise the information criteria AIC and BIC are used to determine the model with the best fit. However, comparing models with different specifications of the fixed effects is not appropriate under Resticted Maximum Likelihood (REML) estimation (Gurka (2006)). Models with different fixed effects parameters are not based on the same amount of observations due to its loss in degrees of freedom $n - p$. Likelihood functions with abberrant number of observations are not comparable. Because a likelihood ratio test is not longer valid under restricted likelihood functions, full Maximum Likelihood (ML) is adopted as well to compare the models. If both ML and REML indicate a specific linear mixed-effects model, it is assumed to be the best performing model compared to the others.

To summarize the linear mixed-effects model, the initial longitudinal parameter values for the EM algorithm of the joint model are estimated by a Restricted Maximum Likelihood (REML) estimation procedure, assuming the random effects to accomodate for the patient-specific correlation within the observed levels of Alkaline Phosphatase and Serum Bilirubin. The Quasi-Newton algorithm is adopted to maximize the log likelihood function. To ensure a positive semi-definite covariance matrix of the random effects, the covariance matrix D is decomposed by a log Cholesky parameterization. Finally, the different parameterizations of the linear mixed-effects model are compared using the log likelihood statistic when models are nested, otherwise the information criteria AIC and BIC are applied. Both ML and REML are applied to determine the fitted linear mixed-effects model.

In the next subsection, different types of missingness of the observed biomarker values, denoted by y_i, are discussed. In addition, it explains the use of a logistic regression to find a possible explanation for the missing data.

3.1.2 Missing Data

A problem involving the analysis of the longitudinal responses of the biomarkers is the presence of missing values. Missing data occur when a measurement is not reported, not performed or failed during a planned visit. In literature, three types of missing data are described, respectively Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR).

If the missing data are labeled Missing Completely At Random (MCAR), they are missing independent from the observed values y_i^o and the values the missing observations should have obtained y_i^m.

$$p(r_i|y_i^o, y_i^m; \theta_r) = p(r_i; \theta_r), \tag{9}$$

where θ_r is the vector of parameters corresponding to the indicator variable r_i, which takes a value of 1 if the observation is missing and 0 otherwise. When patients are dismissed from the follow-up study after a predetermined number of visits or a patient is missing a visit not

[2] Cholesky decomposition: $D = L^t L$, where L is an upper triangle matrix

related to the status of its disease, the missing data can be defined as Missing Completely At Random (MCAR) and the observed levels can be treated as a random sample of the population.

If the missing data mechanism is Missing At Random (MAR), the missing observations are related to the observed observations.

$$p(r_i|y_i^o, y_i^m; \theta_r) = p(r_i|y_i^o; \theta_r). \tag{10}$$

If patients are extracted from the follow-up study for secondary treatment or when they exceed a predetermined threshold, it is called Missing At Random (MAR). However, MAR is impossible to verify statistically (Little and Rubin (2014)).

The last category of missing data is called Missing Not At Random (MNAR). This means the missing value is related to the possible value it might have taken when it was not missing. If it is related to both observed and missing values it is also determined as MNAR.

$$p(r_i|y_i^o, y_i^m; \theta_r) = p(r_i|y_i^m; \theta_r) \ or \ p(r_i|y_i^o, y_i^m; \theta_r). \tag{11}$$

MNAR appears when patients are not able to visit the consultation due to their deteriorated health condition.

To verify if our missing values can be ignored, a logistic regression is used to explain the missing data regarding Alkaline Phosphatase and Serum Bilirubin by other available variables. The logistic regression is able to relate variables to a dependent categorical variable, in this case the indicator variable for missingness, and is given by

$$p(r_i = 1|X_i^T) = \frac{1}{1 + exp(-(X_i^T\beta)}, \tag{12}$$

where X_i is the design matrix of the available covariates, that the logistic regression is using to explain the categorical variable of missingness, denoted by r_i. The corresponding vector of coefficients is denoted by β.

The logistic regression is estimated with numerical optimization using the Fisher Scoring algorithm. When the logistic regression is able to clarify the missing values by its design matrix of available covariates, the missing data are not Missing Completely At Random (MCAR) and the dataset cannot be used as a random sample of the population. However, when a dataset is substantially large, a significant relationship can be found quiet easily and it is recommended to reason if the association is related to the disease. The latter statement basically means that the association has to be different for PBC patients and subjects that are not diagnosed with the disease. In addition, different cities are observed by their missingness to find discrepancies in missing percentages of the two biomarkers. For instance, if one specific city is always missing Alkaline Phosphatase due to the fact they simply do not report this biomarker for PBC patients. In Appendix A.1, the results of various logistic regressions are discussed in order to clarify the missing values of the two biomarkers. As can be observed in Appendix A.1, there are some variables able to explain missing data, but it is assumed that this significant association is mainly determined by the size of the dataset and not associated with the clinical outcomes (death or liver transplantation). Therefor, the missing longitudinal measurements of Alkaline Phosphatase and Serum Bilirubin are discarded from the dataset and treated as Missing Completely At Random (MNAR).

3.1.3 Survival Submodel

In the second building block, the time until a specified event occurs, in this case death or liver transplantation, is analyzed using a Cox proportional hazards function and its corresponding survival function. To accomodate for the special features of the biomarkers, the true and unobserved levels of Alkaline Phosphatase and Serum Bilirubin are introduced in both the Cox proportional hazards function and the survival function. When the features of the two biomarkers are ignored and the observed biomarker levels are directly implemented in the survival functions, the parameter estimates are biased towards zero (Prentice (1982)). To overcome this problem, the true complete history of Alkaline Phosphatase and Serum Bilirubin, denoted by $\mathcal{M}_i$, is obtained by the linear mixed-effects model.

The baseline survival covariates w_i and the continuous true and unobserved levels of the biomarker $m_i(t)$ are assumed to have a multiplicative effect on the hazard ratio (Breslow (1975)). The proportional hazards function including the true levels of the biomarker is given by

$$\begin{aligned} h_i(t|\mathcal{M}_i(t), w_i) &= \lim_{dt \to 0} \mathbb{P}(t \leq T_i^* < t + dt | T_i^* \geq t, \mathcal{M}_i(t), w_i)/dt \\ &= h_0(t)\ exp(\gamma^T w_i + \alpha m_i(t)), t \geq 0 \end{aligned} \tag{13}$$

where $h_0(\cdot)$ represents the baseline hazards function; the hazard ratio when $\gamma^T w_i + \alpha m_i(t) = 0$. w_i is the vector of baseline covariates, assumed to be associated with the hazard ratio, and γ^T is its corresponding vector of coefficients. α represents the coefficient of the true levels of Alkaline Phosphatase or Serum Bilirubin m_i. The true time-to-event is written as T_i^* and $\mathcal{M}_i = \{m_i(s), 0 \leq s < t\}$ describes the complete longitudinal history of the true biomarker values up to time t. The analysis starts at the beginning of the follow-up study.

Specification of the baseline hazard function $h_0(\cdot)$ is required to avoid underestimated standard errors of the parameters estimates (Hsieh et al. (2006)). The range of the baseline hazard function is restricted if explicit distributions like the Weibull distribution or exponential distribution are adopted to describe the function. To avoid underestimated standard errors and maintain flexibility, the baseline hazard function is determined using a piecewise-constant model (Rizopoulos (2012b)). The baseline hazard function using a piecewise-constant model is given by

$$h_0(t) = \sum_{q-1}^{Q} \xi_q I(v_{q-1} < t \leq v_q) \tag{14}$$

where $0 = v_0 < v_1 < ... < v_Q$ is the partition of time in years, where v_Q is the larger than the largest observed time. ξ_q is the hazard ratio during the interval of $(v_{q-1}, v_q]$. Q denotes the number of knots and determines the flexibility of the baseline hazard function.

The hazard ratio of the proportional hazards function is only related to the *current value* of m_i at time point t. To incorporate the true complete history of Alkaline Phosphatase or Serum Bilirubin, the survival function is derived from the proportional hazards function. The survival function including $\mathcal{M}_i$ is given by

$$\begin{aligned} \mathcal{S}_i(t|\mathcal{M}_i(t), w_i) &= \mathbb{P}(T_i^* > t|\mathcal{M}_i(t), w_i) \\ &= exp\bigg(- \int_0^t h_0(s) exp\{\gamma^T w_i + \alpha m_i(s)\}\ ds \bigg). \end{aligned} \tag{15}$$

The main objective of these two functions is to develop a valid inference about the true event times T_i^* using the observed event times T_i and the event indicator variable δ_i. The observed event times are given by the minimum value of the true event time T_i^* and censoring time C_i, $T_i = min(T_i^*, C_i)$. The indicator variable δ_i takes the value of 1 if the patient is deceased or received a liver transplantation, and 0 otherwise. The indicator variable δ_i plays a prominent role in the joint log likelihood to capture censoring. The Cox proportional hazards function and the survival function explain the survival process and constitute the distribution of T_i^* and δ_i in the joint log likelihood function described in section 3.2.

The longitudinal submodel of section 3.1.1 and the survival model are jointly modeled using shared random effects b_i to capture the association between the two processes. Different parameterizations of the proportional hazards model regarding its baseline covariates w_i are used to find the model with the best fit. Baseline covariates have previously been extensively studied. Recently, a risk score (the GLOBE score) was developed by Lammers et al. (2015). Components of this GLOBE score are mainly used as the cross-sectional baseline covariates in the joint modeling framework. To obtain initial parameter estimates for the EM algorithm of the joint log likelihood, the Cox proportional hazards function is first computed with a partial log likelihood.

In section 3.2, the optimization of the joint log likelihood is discussed using Maximum likelihood (ML) estimation. When models are nested the log likelihood statistics determine the model with the best fit using a likelihood ratio test, otherwise the information criteria AIC and BIC are used to compare the models in their goodness-of-fit. In the next subsection, censoring of the time-to-event data is discussed. Not all patients experience an event during the follow-up and their true event time is thereby not available in the dataset.

3.1.4 Censoring

Censoring is an important feature regarding the analysis of a survival process. During the follow-up study not every patient experience an event time, because the event has not occured yet or the patient has dropped out of the study. This thesis concentrates on non-informative right censoring. Right censoring means the event can only be obtained after the end of the follow-up or drop out. When the patient's drop out is not related to its health condition, it is determined a random drop out or non-informative drop out. Non-informative drop out can be interpreted as Missing Completely At Random (MCAR) and benefits from the same features. Appendix A.2 shows the parameter estimates of the fitted joint models without the 184 drop outs. Excluding the patients who dropped out of follow-up shows no large differences and inferences remain the same. Hence, it is assumed that they have no considerable influence on the distribution of the event times and they are handled as if the follow-up has ended normally.

3.2 Maximum Likelihood Estimation of the Joint Model

The idea behind the joint model is that the longitudinal process of a biomarker and the survival process are estimated simultaniously. Computing the longitudinal evolutions of the biomarker (Alkaline Phosphatase or Serum Bilirubin) and the survival process simultaniously uses the available information more optimal (Wulfsohn and Tsiatis (1997)). The association between these two is captured by their shared random effects, denoted by b_i. In this section the two building blocks are linked with these shared random effects to formulate the joint log likelihood. Next to the joint log likelihood, several alternative association structures of the true and unobserved levels of Alkaline Phophatase and Serum Bilirubin are explained.

3.2.1 Joint Log Likelihood

Now both the submodels have been specified, the parameters are obtained using Maximum Likelihood (ML) estimation (Wulfsohn and Tsiatis (1997); Henderson et al. (2000); Hsieh et al. (2006)). The joint log likelihood function is derived from the joint distribution of the time-to-event T_i, the event indicator δ_i and the observed biomarker value y_i. The joint distribution is defined using the shared random effects to link the longitudinal data of the two biomarkers, Alkaline Phosphatase and Serum Bilirubin, to the survival submodel. So, besides the assumption that the random effects ensure the observations of Alkaline Phosphatase and Serum Bilirubin of each patient y_i to be independent, they also account for the relationship between the longitudinal process and the time-to-event likelihoods (Rizopoulos (2012b)). This full conditional independence assumption is denoted by the following two equations

$$p(y_i|b_i;\theta) = \prod_{j=1}^{n_i} p(y_i(t_{ij})|b_i;\theta) \tag{16}$$

$$p(y_i, T_i, \delta_i|b_i;\theta) = p(y_i|b_i;\theta)p(T_i, \delta_i|b_i;\theta), \tag{17}$$

where the conditional independence of the observed measurements of a patient y_i is described by the first equation and the conditional independence between the longitudinal process and the survival outcomes by the latter equation.

In addition, it is assumed that levels of Alkaline Phosphatase and Serum Bilirubin are not constant between visits and that the missing data and censoring is non-informative. The latter assumption means that the missing values of the two biomarkers and the censored values due to drop outs are assumed to be unrelated to the prognosis of PBC patients. Under these assumptions, the joint log likelihood function is given by

$$\begin{aligned}\ell_i(\theta) = \sum_i log\ p(T_i, \delta_i, y_i;\theta) &= \sum_i log \int p(T_i, \delta_i, y_i, b_i;\theta)\ db_i \\ &= \sum_i log \int p(T_i, \delta_i|b_i;\theta_t, \beta)\left[\prod_j p\{y_i(t_{ij})|b_i;\theta_y\}\right]p(b_i;\theta_b)\ db_i,\end{aligned} \tag{18}$$

where $\theta = (\theta_t^T, \theta_y^T, \theta_b^T)$ is the vector of unknown parameters with $\theta_t = (\gamma^T, \alpha, \theta_{h_0}^T)^T$ represents the parameters associated with the survival process and θ_{h_0} are the parameters of the piecewise constant function specifying the baseline hazards function h_0. $\theta_y = (\beta^T, \sigma^2)^T$ denotes the parameters of the observed longitudinal measurements and $\theta_b = vech(D)$ indicates the vectorization of the covariance matrix of the random effects D.

The event times density function, denoted by $p(T_i, \delta_i|b_i;\theta_t, \beta)$, and the longitudinal component of the joint log likelihood, denoted by $p(y_i|b_i;\theta)p(b_i;\theta)$, are given by

$$\begin{aligned} p(T_i, \delta_i|b_i; \theta_t, \beta) &= h_i(T_i|\mathcal{M}_i(T_i), \theta_t, \beta)^{\delta_i} \mathcal{S}_i(T_i), \theta_t, \beta) \\ &= \Big[h_0(T_i)\ exp\{\gamma^T w_i + \alpha m_i(T_i)\} \Big]^{\delta_i} \times exp\Big(- \int_0^{T_i} h_0(s)\ exp\{\gamma^T w_i + \alpha m_i(s)\} ds \Big); \end{aligned} \tag{19}$$

$$\begin{aligned} p(y_i|b_i; \theta)p(b_i; \theta) &= \prod_j p\{y_i(t_{ij})|b_i; \theta_y\} p(b_i; \theta_{b_i}) \\ &= (2\pi\sigma^2)^{n_i/2}\ exp\big\{ - \|y_i - X_i\beta - Z_i b_i\|^2/2\sigma^2 \big\} \times (2\pi)^{-q_b/2} det(D)^{-1/2} exp(-b_i^T D^{-1} b_i/2), \end{aligned} \tag{20}$$

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

Optimizing the joint log likelihood function generally leads to no closed-form solution. To maximize the joint log likelihood function with respect to its parameters, it is recommended to estimate the parameters using a numerical optimization method. In this thesis, the Expectation-Maximization (EM) algorithm is adopted to conduct a full Maximum Likelihood (ML) estimation, where the random effects are treated as missing data. The EM algorithm for the joint modeling framework is described in detail in Appendix A.3.

The gradient of the joint log likelihood function plays an essential rol in the iterative EM algorithm and thereby in the estimation of the parameters and its corresponding standard deviations. The gradient of this joint log likelihood function is called the score vector and represents the sensitivity of the likelihood function with respect to the vector of parameters θ (Rizopoulos (2012b)). The score vector is given by

$$\begin{aligned} S(\theta) &= \sum_i \frac{\partial}{\partial \theta^T} log \int p(T_i, \delta_i|b_i; \theta) p(y_i|b_i; \theta) p(b_i; \theta) db_i \\ &= \sum_i \int \Big[\frac{\partial}{\partial \theta^T} log\{p(T_i, \delta_i|b_i; \theta)\ p(y_i|b_i; \theta) p(b_i|\theta)\} \Big] \times \frac{p(T_i, \delta_i|b_i; \theta)\ p(y_i|b_i; \theta) p(b_i|\theta)}{p(T_i, \delta_i, y_i; \theta)} \\ &= \sum_i \int A(\theta, b_i) p(b_i|T_i, \delta_i, y_i; \theta) db_i, \end{aligned} \tag{21}$$

where $A(\theta, b_i) = \partial\{log\ p(T_i, \delta_i|b_i; \theta) + log\ p(y_i|b_i; \theta) + log\ p(b_i|\theta)\}/\partial\theta^T$ is called the complete data score vector. When the score vector equation (21) is computed, considering the fact that $p(b_i|T_i, \delta_i, y_i; \theta)$ has the parameter estimates θ of the previous iteration, equation (21) coincides with an Expectation-Maximization (EM) algorithm of the joint log likelihood function.

The EM algorithm iteratively alternates between the Estimation of the complete log likelihood function (E-step) and the Maximization of the estimated complete log likelihood function (M-step) untill the algorithm has converged. The E-step involves two integrals, the integral with respect to time of the survival function (15) and the one with respect to the random effects (21). In general, the integral of the survival function (15) and the integral with respect to the random effects (21) do not provide closed-form solutions. Therefor, numerical integration techniques are needed to approximate these integrals in order to obtain the maximum likelihood estimates.

The integral of the survival function (15) with resprect to time is approximated using the 15-point Gauss-Kronrod rule (Press (2007)). The integral with respect to the random effects is more challenging to approximate. Conventional numerical integration techniques like Monte Carlo sampling or Gaussian quadrature rules (Wulfsohn and Tsiatis (1997); Henderson et al.

(2000); Song et al. (2002)) are computational very intensive. That is why Rizopoulos (2012a) developed a alternative approximation technique called the pseudo adaptive Gauss-Hermite rule, a variant of the adaptive Gauss-Hermite rule, to reduce this computational burden. Using the pseudo adaptive Gauss-Hermite rule to estimate the expected score vector with respect to the parameters θ, improves the speed of convergence. The pseudo adaptive Gauss-Hermite rule approximates the score vector of the joint log likelihood by the following equation

$$E\{A(\theta, b_i)|T_i, \delta_i, y_i; \theta\} \approx 2^{q/2}|\tilde{B_i}|^{-1} \sum_{t_1 \dots t_q} \pi_t A(\theta, \tilde{r_t}) p(\tilde{r_t}|T_i, \delta_i, y_i; \theta) exp(||b_t||^2), \tag{22}$$

where $\tilde{r_t} = \tilde{b_i} + \sqrt{2}\tilde{B}^{-1}b_t$. $\tilde{b_i}$ [3] and $\tilde{B_i}$ [4] are estimated using Restricted Maximum Likelihood (REML) on the linear mixed-effects model. b_t are the abscissas, where π_t constitutes its weights. The pseudo adaptive Gauss-Hermite rule has the advantage that it needs less quadrature points than the standard Gauss-Hermite rule and it does not need to relocate the points, which makes the estimation process more adequate and less computational (Rizopoulos (2012b)).

In the M-step the parameters are updated by maximizing the expected complete data log likelihood of the E-step. Once the parameters have been updated the process starts over again with the estimation of the complete log likelihood with the updated parameters and with the new expected log likelihood function the parameters are again updated. Some criteria have been established to know when the EM algorithm has converged. These criteria check whether the parameter values are converging to a stable point and are given by

$$max\{|\theta^{(it)} - \theta^{(it-1)}|/(|\theta^{(it-1)} + 10^{-3})\} < 10^{-4}; \tag{23}$$

$$\ell(\theta^{(it)}) - \ell(\theta^{(it-1)}) < 10^{-8}\{|\ell(\theta^{(it-1)})| + 10^{-8}\}, \tag{24}$$

where $\theta^{(it)}$ is the iterative parameter value with it indicating the number of times the algorithm has iterated.

Besides the fact that the score vector contributes in the realization of the EM algorithm, it is convenient in finding the standard errors of the desired parameters θ. The derivative of the score vector with respect to θ is equal to the Hessian matrix of the joint log likelihood function. Subsequently, the inverse of the negative Hessian matrix is taken to obtain the standard errors of its parameters

$$\begin{aligned} v\hat{a}r(\hat{\theta}) &= \Big\{ - \mathcal{H}(\hat{\theta}) \Big\}^{-1} \\ &= \Big\{ - \sum_{i=1}^{n} \frac{\partial \mathcal{S}_i(\theta)}{\partial \theta} \Big|_{\theta = \hat{\theta}} \Big\}^{-1}, \end{aligned} \tag{25}$$

where $\mathcal{H}$ is the Hessian matrix and $\hat{\theta}$ are the parameter estimates of the joint model. The Hessian matrix is the partial derivative of the score vector with respect to its parameters.
To summarize the estimation process of the joint model parameters and the maximization of its joint log likelihood, an enumeration of the steps has been written below

[3] $\tilde{b_i} = arg\ max_b\{logp(y_i, b; \tilde{\theta_y})\}$, where $\tilde{\theta_y}$ is the parameter vector of the linear mixed-effects model

[4] Choleski factor of $\tilde{H_i} = -\partial^2 logp(y_i, b; \tilde{\theta_y})/\partial b \partial b^T$

Step 1: Estimate the linear mixed-effects model with REML

Step 2: Estimate the proportional hazards model with partial log likelihood

Step 3: Use the parameter estimates of step 1 and step 2 as initial values of the EM algorithm

> **E-step:**
> Compute expected complete data log likelihood using the 15-point Gauss-Kronrod rule and the pseudo adaptive Gauss-Hermite rule to approximate the two integrals
>
> $$\begin{aligned} Q(\theta|\theta^{(it)}) &= E\big\{\sum_i logp(T_i, \delta_i, y_i; \theta)|T_i, \delta_i, y_i; \theta^{(it)}\big\} \\ &= \sum_i \int logp(T_i, \delta_i, y_i, b_i; \theta)p(b_i|T_i, \delta_i, y_i; \theta^{(it)}) \end{aligned} \tag{26}$$
>
> **M-step:**
> Maximize the $Q(\theta|\theta^{(it)})$ in order to update the parameter vector θ

Step 4: Iterate till the algorithm has convergenced (satisfying either of two criteria (23 or 24))

Relating the hazard ratio to the *current value* of the true levels of the two biomarkers, Alkaline Phosphatase and Serum Bilirubin, provides not always the most appropriate association. In the next subsection, four alternative association stuctures are explained to improve the prognostic significance of the two biomarkers in the context of the joint model.

3.2.2 Alternative Association Structures of the Biomarkers

The proportional hazards function introduced in section 3.1.3 shows the relation between the true and unobserved biomarker value at time t and its hazard ratio at the same point in time. α represents the strength of this relationship. Time-dependent covariates like Alkaline Phosphatase and Serum Bilirubin are often handled in a different manor than the baseline covariates. In this thesis, four alternative association stuctures of the time-dependent covariates are presented in order to improve the association between the true biomarker levels and the patient's clinical outcome.

The first extention is the substitution of m_i with a lagged parameter of the true longitudinal response. Sometimes the current measurement of the biomarker is not directly related to the hazard ratio of that point in time. In that case, the association between the true biomarker value and the risk of an event could lead to incorrect conclusions (Vacek (1997)). The parameterization of the hazards function concerning a lagged parameter is given by

$$h_i(t) = h_0(t)exp\big[\gamma^T w_i + \alpha^T m_i\{max(t-c, 0)\}\big]. \tag{27}$$

where t denotes the time in years and c is the lag in years. max is the maximum operator that ensures the time point to be bigger or equal to 0, which represents the beginning of the follow-up study.

Because the biomarkers are time-dependent, the association between the biomarker and the risk of an event can be based on different characteristics of their movements over time. The second parameterization is thereby the addition of the slope or derivative of the true biomarker level, denoted by $m_i^{'}$, and is given by

$$h_i(t) = h_0(t)exp\{\gamma^T w_i + \alpha_1^T m_i(t) + \alpha_2^T m_i^{'}(t)\}, \tag{28}$$

where

$$m_i^{'}(t) = \frac{d}{dt} m_i(t) = \frac{d}{dt}\{x_i^T(t)\beta + z_i^T(t)b_i\}. \tag{29}$$

By adding the slope, the hazard ratio is still only depending on the features of the longitudinal response at time t. To directly relate the hazard ratio to the complete history of the biomarker $\mathcal{M}_i$, one can replace the true observed response by its definite integral over time

$$h_i(t) = h_0(t)exp\bigg\{\gamma^T w_i + \alpha^T \int_0^t m_i(s)ds\bigg\}. \tag{30}$$

The downside of the cumulative parameterization above, is that it gives the same weight to all available responses of the two biomarkers, Alkaline Phosphatase and Serum Bilirubin. Ideally, there is placed more weight on the recent observation than on the observations further in the past. To capture this discrepancy in weights over time, a weighted cumulative parameterization in the functional form of a standard normal density is introduced, given by

$$h_i(t) = h_0(t)exp\bigg\{\gamma^T w_i + \alpha^T \int_0^t \bar{\omega}(t-s) m_i(s)ds\bigg\}. \tag{31}$$

where $\bar{\omega}(t-s)$ is the weight function. The standard normal density as a weight function is specified as $exp\{-(t-s)^2/2\}$. This integral is approximated using a Gauss-Kronrod rule.

Structuring the thesis, first the linear mixed-effects model and the Cox proportional hazards model are estimated individually to obtain the initial parameter values for the EM algorithm to receive the maximum likelihood estimates. Thereafter, different parameterizations of the baseline covariates are compared using goodness-of-fit statistics of its corresponding joint models. The model with the best fit in combination with significant parameters is extended with the four alternative association structures of true biomarker values. The standard fitted joint model and its parameterization in the form of the suggested alternative association structures are compared with goodness-of-fit statistics and with two accuracy measures, respectively the Area Under the Curve (AUC) and its Prediction Error (PE). The log likelihood ratio test is conducted when the models are nested, otherwise the information criteria AIC and BIC are used to conclude. Concerning its accuracy measures, the model with the highest AUC and the smallest PE is characterized as the most accurate.

In section 3.4, the two accuracy measures are discussed to compare the different parameterizations based on their ability of discrimination and calibration. But first, the derivation of the dynamic individual predictions is explained in the next subsection.

3.3 Dynamic Individual Predictions

The patient-specific predictions of survival probabilities have a dynamic feature as they can be updated at each new biomarker observation. When the joint model parameters are estimated based on a sample of the PBC population, denoted by $\mathcal{D}_n$, interest lies in the ability of the fitted joint model to obtain predictions of a new patient with a considerable number of observations. Although it is already possible with one observation of the biomarker (Rizopoulos (2012b)). The patient-specific new set of observations is denoted by y_i^*. Due to the fact that biomarker values are only available when the patient is still alive, focus lies on the computation of a patient-specific conditional survival probability. This conditional survival probability for some future point in time, denoted by u, is computed given the available information up to time t (the patient is still alive at time t, otherwise the probability is equal to zero) and is given by

$$\pi_i(u \mid t) = p(T_i^* \geq u \mid T_i^* > t, y_i^*, \mathcal{D}_n), \tag{32}$$

where $\mathcal{D}_n$ is the sample used to fit the joint model and $u > t$. T_i^* is the patient-specific true time-to-event and y_i^* denotes the set of observations for whom the conditional survival probability is estimated.

To derive the conditional survival probabilities $\pi_i(u \mid t)$, a Monte Carlo simulation is conducted developed by Proust-Lima and Taylor (2009) and Rizopoulos and Ghosh (2011). Therefor, an asymptotic Bayesian approach has to be adopted in order to formulate a posterior expectation of the conditional survival probability (32). This posterior expectation is given by

$$p(T_i^* \geq u | T_i^* > t, y_i^*, \mathcal{D}_n) = \int p(T_i^* \geq u | T_i^* > t, y_i^*; \theta) p(\theta | \mathcal{D}_n) \; d\theta. \tag{33}$$

where θ denotes the vector of parameters.

Based on the conditional independence assumption given in section 3.2.1, the first component of the integral in equation (33) can be written as

$$p(T_i^* \geq u | T_i^* > t, y_i^*; \theta) = \int \frac{\mathcal{S}_i\{u | \mathcal{M}_i(u, b_i, \theta); \theta\}}{\mathcal{S}_i\{t | \mathcal{M}_i(t, b_i, \theta); \theta\}} \; p(b_i | T_i^* > t, y_i^*; \theta) \; db_i, \tag{34}$$

where

$$\mathcal{S}_i\{t \mid \mathcal{M}_i(t, b_i), \theta\} = exp\left\{ \int_0^t h_0(s) exp\{\gamma^T w_i + \alpha m_i(s)\} \right\} ds, \tag{35}$$

denotes the patient-specific survival function. $p(b_i | T_i^* > t, \mathcal{Y}_i(t); \theta)$ is the posterior distribution of the random effects, which is essential in the Monte Carlo simulation to obtain the conditional survival probability. Dividing the product of the conditional likelihood, denoted by $p(T_i, \delta_i \mid b_i; \theta) p(y_i \mid b_i; \theta)$, and the prior distribution of the random effects, denoted by $p(b_i; \theta)$, by the joint distribution of the time-to-event T_i, the event indicator variable δ_i and the longitudinal marker observations y_i provides the posterior distribution of the random effects

$$\begin{aligned} p(b_i \mid T_i, \delta_i, y_i; \theta) &= \frac{p(T_i, \delta_i \mid b_i; \theta) p(y_i \mid b_i; \theta) p(b_i; \theta)}{p(T_i, \delta_i, y_i; \theta)} \\ &\propto p(T_i, \delta_i \mid b_i; \theta) p(y_i \mid b_i; \theta) p(b_i; \theta). \end{aligned} \tag{36}$$

where in this context y_i and T_i are respectively substituted by the new set of observations y_i^* and $(T_i^* > t)$.

The second component $p(\theta | \mathcal{D}_n)$ is the posterior distribution of the parameters conditional on the used sample to fit the joint model. Considered the sample of patients in the dataset is sufficiently large, it is assumed that the posterior distribution of θ approaches an asymptotic normal distribution. This taken into account, a Monte Carlo simulation scheme is proposed to estimate the conditional survival probability $\pi_i(u \mid t)$ (Proust-Lima and Taylor (2009); Rizopoulos and Ghosh (2011)).

1. Simulate the parameter values θ:

   $$\textit{Draw } \theta^{\ell} \sim \mathcal{N}\{\hat{\theta}, \hat{var}(\hat{\theta})\}.$$

2. Simulate the random effect parameters from the posterior distribution of the random effects conditional on the available longitudinal observations y_i^* and the simulated parameter values:

   $$\textit{Draw } b_i^{\ell} \sim p(b_i|T_i^* > t, y_i^*(t), \theta^{\ell} \propto \left\{ \prod_j^{n_i} p(y_{ij} \mid b_i, \theta^{\ell}) \right\} \mathcal{S}_i\{t \mid \mathcal{M}_i(t, b_i), \theta^{\ell}\} p(b_i, \theta^{\ell}).$$

3. Compute the predictive survival probability:

   $$\textit{Calculate } \pi_i^{\ell}(u \mid t) = \mathcal{S}_i\{u|\mathcal{M}_i(u, b_i^{\ell}, \theta^{\ell}; \theta); \theta^{\ell}\} \Big/ \mathcal{S}_i\{t|\mathcal{M}_i(t, b_i^{\ell}, \theta^{\ell}); \theta^{\ell}\}.$$

b_i^{ℓ} is simulated from its posterior distribution using a Metropolis-Hasting algorithm with multivariate t distributed proposals. The Monte Carlo simulation scheme is repeated L times ($\ell = 1, ..., L$) wherefrom the median and mean is taken to provide a value for the conditional survival probability $\pi_i(u \mid t)$. This Monte Carlo simulation scheme computes the first-order estimate of $\pi_i(u \mid t)$. Uncertainty in the Maximum Likelihood (ML) estimates of θ and Bayesian estimates of b_i is spread by using the simulated values (θ^{ℓ} and b_i^{ℓ}) instead of their expected values ($\hat{\theta}$ and $\hat{b}_i$).

In this thesis, the Monte Carlo simulation iterates 200 times and thereafter takes both the mean and median of the computed conditional survival probabilities. In section 5.3.3, individual dynamic predictions are discussed of three patients, one who is deceased before the end of follow-up, one who received a liver transplantation before the end of follow-up and one who is still alive at the end of follow-up. Predictions are made at four different points in time. Conditional survival probabilities constitute an essential component in both accuracy measures as becomes clear in the next section.

3.4 Accuracy

Besides the goodness-of-fit statistics used to compare the joint model and its parameterizations, two accuracy measures have been adopted to determine the models ability to discriminate between patients having a high or low risk of having an event and how well the the fitted model and its alternative association structures of the biomarkers are able to predict a future event.

3.4.1 Discrimination in a Joint Modeling Framework

The Receiver Operating Characteristic (ROC) or ROC curve is a statistical tool to find out how well the biomarker can discriminate between patients in having an event or not. The curve plots the true positive rate (TPR) against the false positive rate (FPR) for different biomarker threshold values. True positive is called the sensitivity of the model and appears when a biomarker threshold c correctly predicts that a patient is having an event. Specificity is the rate which the biomarker threshold correctly specifies the absence of an event. Hence the false positive rate is obtained by substracting the specificity from one. In context of the joint modeling framework the sensitivity and specificity are given by

$$p\{\pi_i(t+\Delta t|t) \leq c \mid T_i^* \in (t, t+\Delta t]\}, \tag{37}$$

and

$$p\{\pi_i(t+\Delta t|t) > c \mid T_i^* > t+\Delta t\}, \tag{38}$$

where c is the biomarkers' threshold.

In order to measure the accuracy in discrimination, the Area Under the Curve (AUC) or C-statistic of the fitted model is calculated. This is the area under the ROC-curve. The prognostic survival probabilities $\pi_i(t+\Delta t|t)$ are computed for two randomly selected patients i and f, based on the methods of section 3.3. Patient i has experienced an event (death or liver transplantation) during $(t+\Delta t)$ and the time-to-event of patient f is unknown at time point $(t+\Delta t)$ due to the absence of an event. The discriminative ability of the fitted joint model for these randomly selected patients $\{i, f\}$ is specified by the following equation

$$AUC(t, \Delta t) = p\big[\pi_i(t+\Delta t|t) < \pi_f(t+\Delta t|t) \mid \{\ T_i^* \in (t, t+\Delta t]\} \cap \{\ T_f^* > t+\Delta t\}\big], \tag{39}$$

where $\cap$ means that both conditions hold. The formula essentially gives the probability that the future survival probability of patient i is smaller than patient f given the fact patient i experienced an event and patient f does not.

In order to extend this for the whole dataset, this is applied to all possible pairs. This leads to a proportion, where the numerator represents the number of pairs that correspond to the condition and the denominator is the total number of possible randomly chosen pairs. To give an example of the interpretation of the AUC, a value of 1 is equal to perfect discrimination and a value of 0.5 is given when the model discriminates randomly. The estimated AUC of the fitted joint model is given by

$$\widehat{AUC}(t, \Delta t) = \frac{\sum_{i=1}^{n}\sum_{f=0;f\neq i}^{n} I\{\pi_i(t+\Delta t|t) < \pi_f(t+\Delta t|t) * I\{\Omega_{if}(t)\}}{\sum_{i=1}^{n}\sum_{f=0;f\neq i}^{n} I\{\Omega_{if}(t)\}}, \tag{40}$$

where $I\{\Omega_{if}\}$ denotes the condition to determine the randomly selected pairs. This condition can be elaborated as follows

$$\Omega_{if}(t) = \big[\{T_i \in (t, t+\Delta t]\} \cap \{\delta_i = 1\}\big] \cap \big[\{T_f > t+\Delta t\}\{\delta_i = 0\}\big] \tag{41}$$

where δ is the event indicator that is equal to 1 if the event has occured, otherwise it is equal to 0. When patient i experiences an event and patient f did not experience an event before $(t+\Delta t)$, the condition holds and its indicator function $I\{\Omega_{if}\}$ has the value 1, otherwise it is 0.

The estimation of the AUC is computed at four different points in time with a time interval Δt to compare the accuracy of the joint models at different moments of follow-up. This is done to conclude if the model is able to discriminate accurately between patients with high risk and low risk of having an event during the whole spectrum of the disease. In the next subsection, the accuracy measure of calibration is explained to find out how well the fitted joint model is able to predict a future event.

3.4.2 Calibration

The expected error of prediction, denoted by $PE(\cdot)$, is a calibration measure that determines the predictive ability of the fitted joint models of the two biomarkers, Alkaline Phosphatase and Serum Bilirubin. It is important to accomodate for censoring when calculating the Prediction Error (PE). Henderson et al. (2002) introduced an estimation technique that accounts for this feature, where the sample is divided into three kinds of patients: (a) patient that are still alive at time point u, denoted by $I(T_i \geq u)L\{1-\hat{\pi}_i(u|t)\}$, (b) patient who experienced an event before u, denoted by $\delta_i I(T_i < u)L\{0-\hat{\pi}_i(u|t)\}$ and (c) patients that are censored before u, denoted by $(1-\delta_i)I(T_i < u)\big[\hat{\pi}_i(u|T_i)L\{1-\hat{\pi}_i(u|t)\} + \{1-\hat{\pi}_i(u|T_i)\}L\{0-\hat{\pi}_i(u|t)\}$. The Prediction Errors (PE) of the three groups are added and divided by the total number of patients who are at risk at time t to obtain the estimated Prediction Error (PE) of the joint model. The formula of Henderson et al. (2002) is given by

$$\widehat{PE}(u|t) = \{\mathcal{R}(t)\}^{-1} \sum_{i:T_i \geq t} I(T_i \geq u)L\{1-\hat{\pi}_i(u|t)\} \; + \; \delta_i I(T_i < u)L\{0-\hat{\pi}_i(u|t)\} \; +$$
$$(1-\delta_i)I(T_i < u)\big[\hat{\pi}_i(u|T_i)L\{1-\hat{\pi}_i(u|t)\} + \{1-\hat{\pi}_i(u|T_i)\}L\{0-\hat{\pi}_i(u|t)\}\big], \qquad (42)$$

where $\hat{\pi}_i(u|t) = \hat{p}(T_i^* \geq u | T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_n)$ and δ_i is the indicator variable taking the value of 1 if an event has occured, otherwise it is equal to zero. $L\{\cdot\}$ denotes the absolute loss function, i.e. $L\{1-\hat{\pi}_i(u|t)\}$ gives $|1-\hat{\pi}_i(u|t)|$. $\mathcal{R}(t)$ shows the number of patients who are at risk at time t. $I(\cdot)$ is an indicator function that takes on the value of 1 if the condition is true, otherwise it is equal to zero. u is a predetermined future point in time, where the prediction is based on.

When models are nested a measure of explained variation, denoted by $R^2_{PE}(\cdot)$, can be computed to determine whether a model improves the accuracy. In the equation below, model $B2$ is compared to $B1$

$$R^2_{PE}(u \mid t; B1, B2) = 1 - \frac{\widehat{PE}_{B2}(u|t)}{\widehat{PE}_{B1}(u|t)} \qquad (43)$$

In the next section, the thesis is prosecuted with a description of the used dataset, followed by a discussion on the results.

4 Data

The clinical and laboratory dataset is collected from 15 liver centers spread over eight countries in North America and Europe provided by the Global PBC Study Group. The dataset was earlier described in detail by Lammers et al. (2014) in their article on the predictive power of the biomarkers Alkaline phosphatase and Serum Bilirubin on death or liver transplantation regarding patients diagnosed with PBC. All patients in the database are identified with an authorized diagnosis of PBC in conformity with the guidelines of Europe and the United States.

In total there were 4,845 patients in the database provided by the Global PBC Study Group with in total 65,642 visits. Due to following inclusion criteria, some patients are discarded from the dataset to estimate the joint model. First, based on Appendix A.1 and A.2 it is assumed the missing data of both biomarkers, Alkaline Phosphatase and Serum Bilirubin, and censoring due to drop outs have no considerable impact. Therefor, the missing values are removed from the dataset. 345 patients were lost due to this criterion. When patients are treated with UDCA and the therapy is observed to be effective, a major drop in the values of Alkaline Phosphatase and Serum Bilirubin can be observed in the first year of treatment. Therefor, values before one year of follow-up are discarded from the dataset to ensure the decrease has no influence on the slope of the linear mixed-effects model, only on its starting point (Lammers et al. (2015)). Currently, almost all PBC patients are treated with UDCA, unless the patient is intolerant to the acid. Therefor, only patients treated with UDCA are included in the sample to fit the joint model, which provides a sample of $3,599$ patients with in total $36,782$ visits.

The dataset contains a large amount of clinical and laboratory information of the patients who are diagnosed with PBC between 1959 and 2012 including gender, age, type of medication, duration and last date of the follow-up, levels of biochemical variables (Serum Bilirubin, Alkaline Phosphatase, Albumine, Platelets, etc.), clinical outcomes (death or liver transplantion) and reasons of ending the follow-up (drop out, death, liver transplantation, end of follow-up).

Verification of the completeness, plausibility and validity of the data are thoroughly checked by members of the Global PBC Study Group. To ensure this, medical charts were extensively reviewed to recover missing data (Lammers et al. (2014)).

5 Results

5.1 Linear Mixed-Effects Model Selection

The objective of the linear mixed-effects model was to estimate the complete longitudinal history of the true levels of the two biomarkers, Alkaline Phosphatase and Serum Bilirubin, denoted by $\mathcal{M}_i$. To correct for nonlinearity, the natural logarithm of the two biomarkers has been taken (Lammers et al. (2015)).

In this thesis, eight specifications of the linear mixed-effects model are proposed to describe the evolutions of the two biomarkers. Different polynomials of time are chosen to describe the evolutions of the biomarker levels. Based on some patient-specific biomarker trajectories, these parameterizations seem to be the most eligible. Only UDCA treated patients diagnosed with PBC were included in the model using levels of Alkaline Phosphatase and Serum Bilirubin after one year of follow-up. After one year of treatment, one can identify patients with a sufficient response to treatment (Lammers et al. (2015)). UDCA has in particular a major impact in the early stage of the disease and in the beginning of the treatment (Lammers et al. (2014)). The eight parameterizations of the linear mixed-effects model are given by

1. $ln(y_i) = \beta_0 + \beta_1 * t + b_0 + b_1 * t + \epsilon_i;$
2. $ln(y_i) = \beta_0 + \beta_1 * t + \beta_2 * t^2 + b_0 + b_1 * t + b_2 * t^2 + \epsilon_i;$
3. $ln(y_i) = \beta_0 + \beta_1 * t + \beta_2 * t^4 + b_0 + b_1 * t + b_2 * t^4 + \epsilon_i;$
4. $ln(y_i) = \beta_0 + \beta_1 * t + \beta_2 * t^2 + \beta_3 * t^4 + b_0 + b_1 * t + b_2 * t^2 + b_3 * t^4 + \epsilon_i;$
5. $ln(y_i) = \beta_0 + \beta_1 * t + \beta_2 * t^3 + b_0 + b_1 * t + b_2 * t^3 + \epsilon_i;$
6. $ln(y_i) = \beta_0 + \beta_1 * t + \beta_2 * t^6 + b_0 + b_1 * t + b_2 * t^6 + \epsilon_i;$
7. $ln(y_i) = \beta_0 + \beta_1 * t + b_0 + b_1 * t + b_2 * t^2 + \epsilon_i;$
8. $ln(y_i) = \beta_0 + \beta_2 * t^2 + b_0 + b_2 * t^2 + \epsilon_i,$

where $ln(y_i)$ is the natural logarithm of the biomarker. $\beta = \{\beta_1, ..., \beta_k\}$ is the vector of coefficients of the fixed effects and $b = \{b_1, ..., b_h\}$ is the vector of coefficients for the random effects.

The independent design vectors $x_i^T(t)$ and $z_i^T(t)$, described in section 3.1.1, are containing different functional forms of time t corresponding to respectively the fixed effects and the random effects. The error terms of the linear mixed-effects model are denoted by ϵ_i. The time in years is denoted by t and starts at one year of follow-up.

As explained in the methodology section, the specifications of the biomarker trajectory are estimated using Restricted Maximum Likelihood (REML). Due to the loss in the degrees of freedom using REML and thereby the discrepancy of the number of observations, it is not appropriate to compare their likelihood statistics directly. In order to account for this feature, all models were estimated using both full Maximum Likelihood (ML) and REML. The log likelihood statistics are used to compare when the models are nested, otherwise the information criteria AIC and BIC are used.

Estimation of the linear mixed-effects models concerning the trajectory of Serum Bilirubin is summarized in table 1, for both ML and REML. Starting with Maximum Likelihood (ML) estimation, model 2 seems to have the best fit based on its log likelihood statistic. The restricted model 7 shows slightly better information criteria under ML. A likelihood ratio test points out that model 2 and 7 are significantly indifferent with a likelihood ratio of respectively 0.3660 and a p-value of 0.5468. When the goodness-of-fit statistics obtained by REML are used to compare the models, model 7 suggests to be the model with the best fit. Again, the goodness-of-fit statistics of model 7 and 2 are relatively close. Although model 2 provides an insignificant value for the fixed effects coefficient β_2, it has a slightly higher log likelihood statistic under ML and it is unlikely that β_2 is exactly equal to zero as model 7 suggests. So, the continuous logarithmic scaled levels of Serum Bilirubin are defined by model 2.

(a) LME models using Maximum Likelihood (ML)

Model	**L**	**AIC**	**BIC**
1	-19719.86	39451.73	39502.80
2	**-19383.71**	38787.43	38872.55
3	-19609.59	39239.19	39324.31
4	-20432.53	40654.92	40823.39
5	-19517.03	39054.06	39139.19
6	-19644.86	39309.73	39394.86
7	**-19383.89**	**38785.79**	**38862.40**
8	-20114.06	40240.12	40291.19

(b) LME models using Restricted Maximum Likelihood (REML)

Model	**L**	**AIC**	**BIC**
1	-19728.73	39469.46	39520.54
2	-19399.81	38819.62	38904.75
3	-19586.90	39193.80	39278.93
4	-25803.64	51637.29	51768.12
5	-19536.55	39093.09	39178.22
6	-19673.77	39367.54	39452.67
7	**-19392.70**	**38803.40**	**38880.01**
8	-20125.41	40262.81	40313.89

Table 1: Goodness-of-fit statistics of different linear mixed-effects models for Serum Bilirubin

In order to select a linear mixed-effects model for the logarithmic scaled levels of Alkaline Phosphatase, the same approach is conducted. The full Maximum Likelihood (ML) estimation directly leads to model 2 using the goodness-of-fit statistics provided in table 2. A likelihood ratio test between model 7 and model 2 provides a likelihood ratio of 2.8610 and a p-value of 0.0908, which indicates no substantial gain in fit. When REML is used to estimate the trajectory of the biomarker Alkaline Phosphatase, model 7 provides the best log likelihood statistic and information criteria. Using the same argument as stated above, that it is unlikely for the β_2 coefficient of model 2 to be exactly equal to zero, combined with the fact that both models show no significant difference in fit, it is suggested to use model 2 in the joint modeling framework.

The other parameterizations of the linear mixed-effects model suggest a substantially lower fit for the trajectory of both Serum Bilirubin and Alkaline Phosphatase based on their goodness-of-fit statistics. So, model 2 is adopted in the joint modeling framework for both biomarkers. Model 2 describes the trajectory of the biomarkers by a second degree polynomial of time for both the fixed effects as the random effects.

(a) LME models using Maximum Likelihood (ML)

Model	L	AIC	BIC
1	-10689.21	21390.42	21441.50
2	**-9555.89**	**19131.78**	19216.91
3	-10096.2	20212.29	20297.42
4	-9698.933	19327.87	19455.56
5	-9857.96	19735.91	19821.04
6	-10388.73	20797.45	20882.58
7	-9557.32	19132.64	**19209.26**
8	-11516.12	23044.25	23095.33

(b) LME models using Restricted Maximum Likelihood (REML)

Model	L	AIC	BIC
1	-10698.57	21409.14	21460.22
2	-9572.64	19165.3	19250.4
3	-10119.3	20258.7	20343.8
4	-9699.148	20088.3	20115.99
5	-9860.75	19741.51	19826.63
6	-10359.34	20738.67	20823.80
7	**-9566.70**	**19151.40**	**19228.02**
8	-11528.02	23068.04	23119.12

Table 2: Goodness-of-fit statistics of different linear mixed-effects models for Alkaline Phosphatase

5.2 Selection of the Joint Model

5.2.1 Different Parameterizations of the Baseline Covariates

There are considered twelve different parameterizations of the baseline covariates of the proportional hazards function, mainly based on the components of the GLOBE score (Lammers et al. (2015)). The GLOBE score is a risk score to support physicians in their development of a diagnosis for PBC patients. This risk score consists of logarithmic scaled values of Alkaline Phosphatase and Serum Bilirubin times the Upper Limit of Normal (ULN) at 1 year of follow-up, age at the beginning of the UDCA treatment, Albumine times the ULN at 1 year of follow-up and the patients' platelet count per 109/L at 1 year of follow-up (Lammers et al. (2015)). To summarize the different models, an enumeration is given below.

1. $h_0(t)\ exp(\alpha m_i(t))$;
2. $h_0(t)\ exp(\gamma_{0i} * FUstartAge + \alpha m_i(t))$;
3. $h_0(t)\ exp(\gamma_{0i} * AgeDiagnosed + \alpha m_i(t))$;
4. $h_0(t)\ exp(\gamma_{0i} * AgeDiagnosed + \gamma_{1i} * ln(bili12) + \gamma_{2i} * ln(alp12) + \alpha m_i(t))$;
5. $h_0(t)\ exp(\gamma_{0i} * FUstartAge + \gamma_{1i} * ln(bili12) + \gamma_{2i} * ln(alp12) + \alpha m_i(t))$;
6. $h_0(t)\ exp(\gamma_{0i} * AgeDiagnosed + \gamma_{1i} * FUstartAge + \gamma_{2i} * ln(bili12) + \gamma_{3i} * ln(alp12) + \alpha m_i(t))$;
7. $h_0(t)\ exp(\gamma_{0i} * FUstartAge + \gamma_{1i} * ln(bili12) + \gamma_{2i} * ln(alp12) + \gamma_{3i} * alb12 + \alpha m_i(t))$;
8. $h_0(t)\ exp(\gamma_{0i} * FUstartAge + \gamma_{1i} * ln(bili12) + \gamma_{2i} * ln(alp12) + \gamma_{3i} * alb12 + \gamma_{4i} * Sex + \gamma_{5i} * AMA + \gamma_{6i} * plat12 + \alpha m_i(t))$;
9. $h_0(t)\ exp(\gamma_{0i} * FUstartAge + \gamma_{1i} * alb12 + \gamma_{2i} * plat12 + \alpha m_i(t))$;
10. $h_0(t)\ exp(\gamma_{0i} * FUstartAge + I[y_i = alp]\gamma_{1i} * ln(bili12) + I[y_i = bili]\gamma_{2i} * ln(alp12) + \gamma_{3i} * alb12 + \gamma_{4i} * plat12 + \alpha m_i(t))$;
11. $h_0(t)\ exp(\gamma_{0i} * FUstartAge + \gamma_{1i} * alb12 + \alpha m_i(t))$;
12. $h_0(t)\ exp(\gamma_{0i} * alb12 + \alpha m_i(t))$,

where the first model contains no baseline covariates, only the patient-specific *current value* of the biomarker without its error term, denoted by $m_i(t)$. In the rest of the models, this *current value* of the biomarker is accompanied by baseline covariates with their corresponding coefficients, denoted by γ_{pi}, where p is the amount of baseline covariates. The baseline covariates $FUstartAge$ and $AgeDiagnosed$ are respectively the age at the beginning of the follow-up study and the age at diagnosis of PBC. $ln(bili12)$, $ln(alp12)$ and $alb12$ are the biomarker values times the Upper Limit of Normal (ULN) at 1 year of follow-up, where the two biomarkers, Alkaline Phosphatase and Serum Bilirubin, are scaled logarithmic to correct for nonlinearity and Albumine is not. Sex and AMA are categorical variables to indicate the patients' gender and presence of the Anti-mitochondrial antibodies (AMA). AMA are autoantibodies, found in approximately 90-95% of the patients diagnosed with PBC (Gershwin et al. (1987)). AMA could be observed when the liver results are normal and the patient shows no signs or symptoms. The problem is that 0.5% of the population not diagnosed with PBC also exhibits positive test results on AMA. So, a positive test result on AMA is not able to indicate PBC on its own. $plat12$ is the platelet count per 109/L at 1 year of follow-up. Model 10 uses the logarithmic scaled levels of Alkaline Phosphatase times the ULN at 1 year of follow-up, when $m_i(t)$ represents the *current value* of Serum Bilirubin. This condition is denoted by the indicator function $I[\cdot]$. When the interest lies in the relation between the levels of Alkaline Phosphatase without an error m_i and the hazard ratio, the logarithmic scaled level of Serum Bilirubin times the ULN at 1 year of follow-up is used in model 10 as baseline covariate.

Table 4 shows the goodness-of-fit statistics regarding the joint models including the true values $m_i(t)$ of Serum Bilirubin. The first model $B1$ contains no baseline covariates and its goodness-of-fit statistics indicates the least fit compared to the other joint models. Adding the age at the beginning of the follow-up, denoted by $FUstartAge$, provides an substantially improved fit compared to $B1$. Model $B2$ is significantly preferred to model $B1$ with a likelihood ratio of 270.7 and a p-value smaller than 0.0001. The age at diagnosis of PBC is used in model $B3$, indicating a less improved fit compared to model $B2$. In models $B4$ and $B5$ the logarithmic scaled levels of Serum Bilirubin and Alkaline Phosphatase times the ULN at 1 year of follow-up are included, but they show no significant relation with the hazard ratio. This insignificance is probably due to the multicollinearity with the levels of Serum Bilirubin $m_i(t)$. To test the added value of the age at diagnosis, model $B5$ and $B6$ are compared using the likelihood ratio test. The fit of the two models are expected to be significantly indifferent with a likelihood ratio of 1.28 and a p-value of 0.2587. The second substantial improvement in fit is provided by the addition of Albumine times the ULN at 1 year of follow-up, shown in the models $B7$ up to $B11$. These goodness-of-fit statistics and the likelihood ratio's of the models $B7$ up to $B11$ show no significant difference between their fit. In order to choose between these five models, the significance and thereby clinical relevance has been examined. Only the level of Albumine times the ULN at 1 year of follow-up and the platelets at 1 year of follow-up are determined with a significant relation towards the hazard ratio. Excluding the follow-up start age from model $B11$ diminishes its fit as can be observed by the goodness-of-fit statistics of $B12$. Thereby, model 9 is suggested to use in case of a joint modeling framework where the longitudinal submodel contains the biomarker values of Serum Bilirubin. Parameter estimates of model $B9$ are shown in Appendix A.4.

Model	**L**	**AIC**	**BIC**
B1	-22332.72	44701.43	44812.82
B2	-22197.37	44432.74	44550.32
B3	-22226.93	44491.86	44609.44
B4	-22226.94	44495.87	44625.83
B5	-22197.41	44436.83	44566.78
B6	-22196.41	44436.82	44572.97
B7	-22175.08	44394.17	44530.31
B8	-22174.25	44398.5	44553.21
B9	**-22173.21**	**44388.42**	**44518.38**
B10	-22173.12	44390.24	44526.38
B11	-22175.81	44391.62	44515.38
B12	-22295.70	44629.39	44746.97

Table 3: Goodness-to-fit statistics of the different joint models for Serum Bilirubin

Next, the joint model concerning the longitudinal measurements of Alkaline Phosphatase is discussed. The parameterizations 1 till 12 are corresponding to the earlier described enumeration. Table 5 summarizes the goodness-of-fit statistics corresponding to its joint models.

In the joint models of Alkaline Phosphatase, it can be observed that the effects of the age at the beginning of the follow-up and the age at diagnosis are similar to those in the models where the true levels of Serum Bilirubin are included. A likelihood ratio test exhibits a significant improved fit in model $A2$ compared to $A1$ with a likehood ratio of 209.89 and a p-value smaller than 0.0001. Again, $A3$ has a smaller log likelihood statistic than $A2$.

Adding both logarithmic scaled biomarker levels times the ULN at 1 year of follow-up improves the fit substantially, shown in model $A4$ and $A5$. The likelihood ratio test between $A5$ and $A6$ indicate no significant difference in fit, wherefrom it is concluded that the added value of the age at diagnosis is of minor importance than the age at the beginning of the follow-up. Albumine times the ULN at 1 year of follow-up shows just like the model with Serum Bilirubin a substanstial improvement in the fit in model $A7$. Model $A7$ is significantly favored over model $A5$ with a likelihood ratio of 81.62 and a p-value smaller than 0.0001. Model $A8$ shows an increase in its log likelihood statistic compared to $A7$ and $A9$. Its high log likelihood statistic is probably due to its large number of baseline covariates, where some covariates show highly insignificant results. The presence of the auto-antibody AMA and the patients' gender show no significant association, with p-values of 0.2712 and 0.2592. Their association with the hazard ratio is assumed to be of less medical relevance. The logarithmic scaled level of Alkaline Phosphatase at 1 year of follow-up is less significant than the other parameters due to its correlation with the true level of the Alkaline Phosphatase, denoted by $m_i(t)$. Hence, model $A10$ is preferred due to it parsimonious feature, highly significant parameters and high value of its log likelihood statistic. The logarithmic scaled level of Serum Bilirubin times the ULN at 1 year of follow-up shows no multicollinearity with the true levels of Alkaline Phophatase and improves the fit substantially, next to $alb12$, $FUstartAge$ and the number of platelets at 1 year of follow-up. Parameter estimates of joint model $A10$ are shown in the Appendix A.4. The fact Serum Bilirubin is not correlated with levels of Alkaline Phosphatase is due to its later evolving changes compared to Alkaline Phosphatase. Exluding the age at the beginning of the follow-up is not recommended as shown by the major drop in the log likelihood statistic of model $A12$.

Model	L	AIC	BIC
A1	-13024.97	26085.95	26197.34
A2	-12920.07	25878.14	25995.72
A3	-12946.94	25931.88	26049.46
A4	-12699.86	25441.72	25571.68
A5	-12678.19	25398.38	25528.34
A6	-12676.8	25397.59	25533.74
A7	-12637.38	25318.76	25454.9
A8	**-12615.29**	**25280.58**	**25435.29**
A9	-12756.22	25554.45	25684.41
A10	**-12620.59**	**25285.18**	**25421.33**
A11	-12821.47	25682.95	25806.72
A12	-12903.29	25844.57	25962.15

Table 4: Goodness-to-fit statistics of the different joint models for Alkaline Phosphatase

So, models $B9$ and $A10$ are recommended based on the goodness-of-fit statistics, the log likelihood ratio test, the significance of the covariates and the amount of parameters.

In the next section, different association structures between the biomarkers and the hazard ratio are discussed using goodness-of-fit statistics to compare their fit. Thereafter, individual dynamic predictions are derived from the fitted model with the association structure that has the best fit, followed by the accuracy measures of the fitted model and its alternative structure associations.

5.2.2 Parameterizations of the Longitudinal Data

The different parameterizations of the longitudinal data represent alternative association structures between the risk of an event and the true trajectory of the two biomarkers. In this thesis, four alternative association structures of the biomarker trajectory are suggested: (1) a two years lagged value of the true *current value* of Alkaline Phophatase or Serum Bilirubin, (2) the addition of the slope of the biomarkers' trajectory, (3) the inclusion of the true complete history of the biomarker by an equally weighted cumulative function and (4) the addition of the weighted cumulative function to allocate different weight to the available true biomarker levels. The weights are determined by a standard normal density as explained in section 3.2.2.

In table 5 and 6, the goodness-of-fit statistics recarding the longitudinal parameterizations of the fitted models $B9$ and $A10$ are shown. Among the alternative association structures to relate Serum Bilirubin to the risk of an event, a 2 years lagged parameter value of Serum Bilirubin shows no gain in fit compared to the joint model with the *current value.* Adding the slope of the *current value* of Serum Bilirubin improves the fit significantly with a likelihood ratio test statistic of 7.48 and a p-value of 0.0062. When all available information of the biomarker is related to the hazard ratio in the functional form of a cumulative function, the weighted cumulative function is preferred and substantially improves the fit compared to the model with the *current value* of Serum Bilirubin. A likelihood ratio test prefers this parameterization significantly above that of only its *current value* with a likelihood ratio statistic of 28.19 and a p-value smaller than 0.0001. Based on the goodness-to-fit statistics, Serum Bilirubin is related in the best way to the hazard ratio through a weighted cumulative function.

For the fitted joint model including all available observations of Alkaline Phosphatase, linking the risk of an event to the weighted integral of the trajectory of Alkaline Phosphatase (earlier described in section 3.2.2), provides the largest gain in fit. Likelihood ratio test for comparing the basic fitted model with the weighted cumulative parameterization leads to a likelihood ratio of 47.68 with a p-value smaller than 0.0001. In case of Alkaline Phosphatase, the addition of a slope results in a significant improved fit as well with a likelihood ratio of 7.32 and a p-value 0.0068. Relating the hazard ratio to the two year lagged marker value or the equally weighted cumulative function show no improvements in fit. Including all available information with carefully determined weights improves the fit of the joint model the most.

In the end both biomarkers are showing a significant improved fit when they are related to the hazard ratio through a weighted cumulative function. This indicates that the inclusion of all available information of the biomarkers in the proportional hazards function improves the performance of the joint model. In the next subsection, the result recarding the dynamic individual predictions of three different patient is discussed.

Model	**L**	**AIC**	**BIC**
B9 *current value*	-22173.21	44388.42	44518.38
B9 *lagged*	-22234.03	44510.07	44640.02
B9 *value + slope*	-22169.47	44382.94	44519.08
B9 *cumulative*	-22426.31	44894.62	45024.58
B9 *weighted cumulative*	**-22159.12**	**44360.23**	**44490.19**

Table 5: Goodness-to-fit statistics of the fitted joint model of Serum Bilirubin and its four alternative association structures of m_i

Model	L	AIC	BIC
A10 *current value*	-12620.59	25285.18	25421.33
A10 *lagged*	-12625.23	25294.46	25430.6
A10 *value + slope*	-12616.93	25279.86	25422.2
A10 *cumulative*	-12634.33	25312.67	25448.81
A10 *weighted cumulative*	**-12596.75**	**25237.5**	**25373.64**

Table 6: Goodness-to-fit statistics of the fitted joint model of Alkaline Phosphatase and its four alternative association structures of m_i

5.2.3 Individual Dynamic Predication

To illustrate the ability to derive individual dynamic predictions using the Monte Carlo scheme explained in section 3.3, several plots are shown in Appendix A.5. The plots are created using model $B9$ and $A10$ with the weighted cumulative parameterization for three different patients at four different points in time (2 years, 5 years, 10 years and 15 years). The first patient is still alive at the end of follow-up, the second has been deceased before the end of the follow-up and the last one received a liver transplantation before the end of the follow-up.

In figure 1, the individual dynamic predictions of the patient who is still alive at the end of follow-up show for both biomarkers almost the same predicted survival curves. The logarithmic scaled levels of Serum Bilirubin seems to be considerably constant, while the levels Alkaline Phosphatase exhibit a small decrease. These graphs provide no aberrant results with the literature in the sense the biomarker levels are constant or decreasing as well as its survival curve.

Figure 2 shows more aberrant prognoses. The levels of Serum Bilirubin exhibit an increase over time, which leads to more steep survival curves. However, just before 10 years of follow-up a sudden decrease is observed, what improves the patients predictive survival probability. Thereafter, the status of the patient is detoriated, which can be observed by the increase in Serum Bilirubin. Shortly after this increase the patient is deceased. The Alkaline Phosphatase levels show an opposite trend, what results in too optimistic survival curves. This indicates the importance of the use of both biomarkers.

In figure 3, the individual dynamic predictions of a patient who received a liver transplantation is shown. The situation is worse than for the patient who is deceased before the end of follow-up. Both levels of Alkaline Phosphatase and Serum Bilirubin are high, which indicates steeper survival curves. In the beginning of the follow-up study, individual dynamic predictions using the joint model of Serum Bilirubin provide too optimistic survival curves compared to the situation indicated by the joint model of Alkaline Phosphatase. This shows again the importance of using both biomarkers for the prognosis of the PBC patient and corresponds to literature in the sense Serum Bilirubin generally demonstrate meaningful changes in a later stadium of PBC (Lammers et al. (2015)).

As can be observed in the individual dynamic prediction plots, it is of great importance to use both biomarkers to obtain a prognosis for the patient. As earlier explained in literature, Serum Bilirubin demonstrates mostly significant changes in a later stage of the disease compared to Alkaline Phophatase, which shows meaningfull changes during the whole spectrum of PBC (Lammers et al. (2015)). Changes of Alkaline Phosphatase levels are assumed to be more spread over the course of the disease and are more important for the diagnosis of PBC. In a later stage, it could be harder to discriminate between patient due to the incline of the overall biomarker level. As Serum Bilirubin only shows meaningfull changes in a further developed stage of the disease, it is expected that its prognostic strength is better in a later stage of disease than

Alkaline Phosphatase. To confirm this hypothesis and to compare the alternative association structures of the biomarkers in discrimination and calibration, accuracy measures are discussed in the next two subsections.

5.2.4 Discrimination

For the comparison in terms of discriminitive strength in identifying patients with an high risk of having an event or a low risk, the 10-fold cross validated estimates of the Area Under the Curve (AUC) are calculated for the fitted models and its parameterizations of the true biomarker level explained in Section 3.2.2. To determine whether the ranking of Alkaline Phosphatase or Serum Bilirubin correspond with their health status, the AUC is computed at four different points in time, respectively at 2, 5, 10 and 15 years of follow-up. The length of its prediction interval is in this case two years.

The best predictive performance within the different parameterizations are not necessarily corresponding to the model with the best goodness-of-fit statistic (Proust-Lima and Taylor (2009)). The accuracy measures are based on a part of the used sample to make a dynamic prediction up to a predetermined point in time, while the goodness-of-fit statistics are based on all information of the complete sample.

In table 8 and 9, the AUC values are summarized for both fitted joint models at four different points in time. The joint model with the longitudinal observations of Serum Bilirubin is fitted using three baseline covariates, respectively the age at starting the follow-up, Albumine times the ULN at 1 year of follow-up and platelet count 109/L at 1 year of follow-up. For the joint model of Alkaline Phosphatase, a fourth baseline covariate is added in the form of the logarithmic scaled Serum Bilirubin level times the ULN at 1 year of follow-up. Next to the relation between the *current value* of the biomarker and its hazard ratio, the four different association structures described in section 3.2.2 are conducted to find out which has the best discrimination. As can be observed, the AUC values tend to decrease over time for both biomarkers. The explanation for this decrease is due to the fact that the biomarker values rise over time, while the number of patients decreases. In the beginning of the follow-up study, many patients show relatively low levels of the markers. If a patient experiences an event in the beginning of the follow-up and shows increased values of Alkaline Phosphatase or Serum Bilirubin, it is much easier to identify the patient. In a later stage of the follow-up, marker levels have increased over time for all patients, what makes it harder to discriminate between patients.

In general, the joint model involving the levels of Alkaline Phosphatase show lower AUC values than the models with Serum Bilirubin. This has to do with the fact that the levels of Alkaline Phosphatase exhibit changes across the whole spectrum of the disease, while the Serum Bilirubin shows meaningfull changes in a later stage of the hepatobiliary disease.

The accuracy measures for discrimination show for both biomarker models and its parameterizations no remarkable differences cross sectionally, although the equally weighted cumulative is consequently lower than the other parameterizations. This indicates that the equally weighted cumulative parameterization of the joint model has an overall worse performance than the other models. For both biomarkers the weighted cumulative parameterization shows the best discriminative performance at all points in time, with exception of the AUC at 15 years of follow-up for the joint model involving the longitudinal values of Serum Bilirubin. Here, the AUC leads to the model with the slope of the biomarkers' trajectory. In a later stage of the disease, the biomarker levels have all gone up and thereby the slope of the true and unobserved biomarker levels constitute a more important role.

In Section 5.3.2, the addition of the slope of the marker trajectory already showed a substantial gain in fit. Based on the goodness-of-fit statistics of Section 5.3.2 and the values of the AUC, the weighted cumulative parameterization is preferred up to 10 years of follow-up for Serum Bilirubin and thereafter the slope parameterization. For Alkaline Phosphatase the weighted cumulative parameterization is recommended at all predetermined points in time.

In the next Section, the prediction error is computed to discuss the ability whether the model is accurate in the prediction of a future event.

model	$\widehat{AUC}(4\|2)$	$\widehat{AUC}(7\|5)$	$\widehat{AUC}(12\|10)$	$\widehat{AUC}(17\|15)$
B9 *value*	0.8381	0.8246	0.8095	0.7569
B9 *lagged value*	0.8378	0.8172	0.8017	0.7324
B9 *value + slope*	0.8376	0.8213	0.8115	**0.7628**
B9 *cumulative*	0.7952	0.7875	0.7844	0.7229
B9 *weigted cumulative*	**0.8385**	**0.8248**	**0.8118**	0.7572
Number of patients at risk	3,392	2,592	1,378	525

Table 7: Discrimination measures of different parameterizations of the fitted joint model of Serum Bilirubin

model	$\widehat{AUC}(4\|2)$	$\widehat{AUC}(7\|5)$	$\widehat{AUC}(12\|10)$	$\widehat{AUC}(17\|15)$
A10 *value*	0.8373	0.7901	0.7602	0.6914
A10 *lagged value*	0.8369	0.7873	0.7581	0.6896
A10 *value + slope*	0.8375	0.7908	0.761	0.6928
A10 *cumulative*	0.7776	0.7776	0.7561	0.6746
A10 *weigted cumulative*	**0.8376**	**0.7929**	**0.7637**	**0.6949**
Number of patients at risk	3,392	2,592	1,378	525

Table 8: Discrimination measures of different parameterizations of the fitted joint model of Alkaline Phophatase

5.2.5 Calibration

Calibration measures are computed using a 10-fold cross validation as well. The situation at time points 2, 5, 10 and 15 are again examined with an prediction interval of two years. Just like the AUC values, the prediction errors are computed with a subset of the complete sample, which means it does not necessarily leads to the same model as the goodness-of-fit statistics.

Table 10 and 11 show the estimated prediction errors for the different fitted models. The estimated errors increase over time due to the fact that the number of patients has decreased and the overall level of the biomarker has inclined, what makes it harder to predict events accurately. Hence, an increase in the estimated prediction errors over time is observed.

At two and five years of follow-up, the weighted cumulative parameterization has the smallest prediction error regarding the joint model of Serum Bilirubin. At 10 years of follow-up the slope parameterization leads to the smallest prediction error for Serum Bilirubin. However, if the measure of explained variation R^2 is used to compute how much the accuracy has increased for the weighted cumulative parameterization and the slope parameterization, both parameterizations are not resulting in a considerable different gain in accuracy. Both models are resulting in a gain of explained variation of approximately 0.01 relative to the model with the *current value* of Serum Bilirubin. So, based on the AUC and goodness-of-fit statistics the weighted cumulative parameterization can be used as well at 10 years of follow-up without a substantial difference in calibration. The last time point at 15 years of follow-up leads to the equally weighted cumulative parameterization. This is probably a random call due to the fact the model is not performing well in general at 15 years of follow-up. Combining the goodness-of-fit statistics, the AUC values and the prediction errors, the slope parameterization of model $B9$ is recommended at 15 years of follow-up.

In case of the joint models regarding Alkaline Phosphatase, the estimated prediction errors designate the model with the slope parameterization at 2 years of follow-up. However, this model gains in accuracy by approximately 0.0018 of explained variance, while the weighted cumulative parameterization declines in accuracy by 0.01. This difference is quite small and thereby the weighted cumulative parameterization is still recommended based on its AUC value and goodness-of-fit statistics. At time point 5 and 10, joint model $A10$ of Alkaline Phosphatase with the weighted cumulative parameterization has the smallest prediction error for Alkaline Phosphatase. Again, the equally weighted cumulative function has the smallest prediction error at 15 years of follow-up. However, if the goodness-of-fit statistics, AUC values and prediction errors are again combined, the weighted cumulative parameterization has the preference above the other.

model	$\widehat{PE}(4\|2)$	$\widehat{PE}(7\|5)$	$\widehat{PE}(12\|10)$	$\widehat{PE}(17\|15)$
B9 *value*	0.0509	0.0642	0.0813	0.1006
B9 *lagged value*	0.0514	0.0672	0.0863	0.1117
B9 *value + slope*	0.0508	0.064	**0.0803**	0.1011
B9 *cumulative*	0.0596	0.0723	0.0828	**0.0996**
B9 *weigted cumulative*	**0.0504**	**0.0629**	0.0807	0.1043
# of patients at Risk	3,392	2,592	1,378	525

Table 9: Calibration measures of different parameterizations of the fitted model of Serum Bilirubin

model	$\widehat{PE}(4\|2)$	$\widehat{PE}(7\|5)$	$\widehat{PE}(12\|10)$	$\widehat{PE}(17\|15)$
A10 *value*	0.0558	0.0729	0.0933	0.1272
A10 *lagged value*	0.0559	0.0732	0.0935	0.1282
A10 *value + slope*	**0.0557**	0.0726	0.093	0.1261
A10 *cumulative*	0.0567	0.0734	0.0929	**0.1227**
A10 *weigted cumulative*	0.0564	**0.0721**	**0.0924**	0.1273
# of patients at Risk	3,392	2,592	1,378	525

Table 10: Calibration measures of different parameterizations of the fitted model of Alkaline Phophatase

When the goodness-of-fit, the AUC's and the prediction errors are combined to designate a model, the weighted cumulative parameterization of model $B9$ and $A10$ is suggested up to and including time point 15, excluding the joint model of Serum Bilirubin at 15 years of follow-up. Joint model $B9$ leads to a preference of the slope parameterization at 15 years of follow-up.

6 Conclusion

In this thesis, the selection of a joint model is examined to discuss the discriminative ability of the biomarkers Alkaline Phosphatase and Serum Bilirubin over time to the risk of an event (death or liver transplantation). Upon fitting it to the most comprehensive database regarding patients diagnosed with PBC, it is revealed that the trajectory of the biomarkers is captured by a linear mixed-effects model where both the random effects as the fixed effects are determined by a second degree polynomial of time. Alkaline Posphatase and Serum Bilirubin were analyzed in their natural logarithmic values to correct for nonlinearity. The focus has been on patients treated exclusively with Ursodeoxycholic acid (UDCA) because this constitutes the current standard.

To know the dependence between the trajectory of the two biomarkers and its risk of experiencing an event, the linear mixed-effects model has been related to the survival process using a shared random effects approach. This means the random effects explain the interdependencies between the longitudinal data and the time-to-event data. The clinical outcomes (death or liver transplantation) of the patient are depending on two types of covariates, the time-independent baseline covariates and the time-dependent endogenous covariates (Alkaline Phophatase and Serum Bilirubin). To test the baseline covariates, different parameterizations of the joint model are estimated in combination with the non parameterized true *current value* of Alkaline Phophatase or Serum Bilirubin. Goodness-of-fit statistics shows a substantially improved fit, when the level of Albumine times the Upper Limit of Normal (ULN) at one year of follow-up, the platelet count per 109/L at 1 year of follow-up and the age at the beginning of the follow-up are included as baseline covariates. All are highly significant. In case of the jointly modeled levels of Alkaline Phosphatase, the logarithmic scaled level of Serum Bilirubin times the ULN at 1 year of follow-up gains the fit substantially next to the latter three baseline covariates. The other less parsimonious joint models including these baseline covariates show a similar fit, but contain insignificant parameters presumably due to multicollinearity or medical irrelevance.

To extend the possibilities of relating the longitudinal trajectory to the clinical outcomes (death or liver tranplantation), four alternative association structures of the true levels of Alkaline Phophatase and Serum Bilirubin are suggested. The use of lagged parameters and the inclusion of the slope and (weighted) cumulative effects have been examined. For both Alkaline Phosphatase and Serum Bilirubin, a substantially improved fit can be observed using the weighted cumulative parameterization of the true marker levels. This is due to the inclusion of the complete true history of Alkaline Phophatase and Serum Bilirubin. In a more general sense, the preference that is found in the joint model with the weighted cumulative parameterization points out that the inclusion of the complete true history of the biomarkers through the proportional hazards function indicates a more efficient use of the available information.

To determine the predictive ability of the fitted joint model and its different association structures of the true marker levels, a 10-fold cross validated estimate of the Area Under the Curve (AUC) and the Prediction Error (PE) are computed in this thesis. Although the weighted cumulative parameterization shows a substantially improved fit, its validated accuracy measure is not so different from the other parameterizations. The difference of designation of a model between the goodness-of-fit statistics and the accuracy measures are explained by their use of information. The accuracy measures are based on a part of the used sample to make a dynamic prediction up to a predetermined point in time, while the goodness-of-fit statistics are based on all information of the complete sample. Up to 15 years of follow-up the weighted cumulative parameterization shows for both biomarkers the best discriminative strength using the Area Under the Curve (AUC) with exception of the joint model of Serum Bilirubin at 15 years of follow-up, where the slope seems to contain more importance. In addition, the slope parameterization shows an improved fit as well by the goodness-of-fit statistics, which indicate increased importance of the trajectory slope over time. This is due to the incline of the overall biomarker values over time.

How well the model can predict an event is quantified by the estimated prediction error of the fitted model and its alternative association stuctures. Generally, the estimated prediction errors have the smallest values with the weighted cumulative parameterization up to 10 years of follow-up for both biomarkers. Thereafter, it leads to the equally weighted cumulative parameterization, which is assumed to be associated with the detoriated accuracy in general of the model. Combining the goodness-of-fit statistics with the accuracy measures designates the model with the second smallest prediction error at 15 years of follow-up for the joint model with Serum Bilirubin; the model with the slope parameterization. For the model with longitudinal measurements of Alkaline Phosphatase, the weighted cumulative parameterization remains the preferred model.

So, when the goodness-of-fit statistics, the AUC's and the prediction errors (PE) are combined to designate a model, the weighted cumulative parameterization of both biomarkers can predict clinical outcomes (death or liver transplantation) in the most accurate way up to 15 years of follow-up; with exclusion of the joint model involving longitudinal measurements of Serum Bilirubin at 15 years of follow-up. This model benefits from the slope parameterization at 15 years of follow-up. Both biomarkers are accompanied by three baseline covariates, respectively Albumine times the ULN at 1 year of follow-up, platelet count per 109/L at 1 year of follow-up and the age at the beginning of the follow-up. In case of the joint model regarding Alkaline Phosphatase, the logarithmic scaled levels of Serum Bilirubin times the ULN at 1 year of follow-up improve the performance of the model significantly next to the other three earlier mentioned. Although the joint model of Serum Bilirubin generally indicates more accuracy over time, it is recommended to use both models to develop a diagnosis for the remaining course of Primary Biliary Cholangitis (PBC) as both biomarkers are able to predict clinical outcomes.

There are various suggestions for further research. Currently, there are many applied risk scores to provide physicians an appropriate prognosis, like the Paris-1 or the GLOBE score. It would be interesting to compare the fitted joint model with these widely used scoring systems by their accuracy or to adopt a risk score into the joint modeling framework. In this thesis, the biomarker levels of Alkaline Phosphatase and Serum Bilirubin are individually related to the risk of an event. Further research could relate both markers simultaniously to the risk of an event. The importance of the use of both biomarker is already observed in the individual dynamic predictions of this thesis. Hence, relating both simultaniously could provide a better prognosis.

The research on the predictive performance of Alkaline Phosphatase and Serum Bilirubin in the context of a joint modeling framework might show improved results using a Bayesian approach, although literature is still inconclusive about both methods. But since the dynamic individual predictions and the accuracy measures make use of Bayesian theorem, it seems reasonable to use a Bayesian approach for the estimation of the joint model.

There exists a classification method based on the levels of Serum Bilirubin and Albumine, dividing the course of the disease in four different stages. Normal levels of Serum Bilirubin and Albumine are associated with stage one. If both Serum Bilirubin and Albumine show abnormal levels, the disease is in the fourth stage (ter Borg et al. (2006)). In this thesis, Albumine times the ULN at 1 year of follow-up shows a substantially improve in fit for the joint model. Examine the prognostic significance of Albumine as a longitudinal biomarker might be an interesting study in the future.

Acknowledgements

Prof. Dr. Paap of the department of Econometrics at the Erasmus University Rotterdam and Dr. Hansen of the department of Gastroenterology and Hepatology are gratefully acknowledged for their helpfull comments and suggestions. I also would like to thank Prof. Dr. Fok as co-reader at the department of Econometrics for his time to revise this thesis and the PBC Study Group for the acces to their large database on PBC.

A Appendix

A.1 Missing data

In order to clarify the missing data by a logistic regression, the biomarker values are transformed into categorical variables. When an observation is available, it takes the value of zero and when the observation is missing, it gets the value of 1. With the logistic regression and several other available covariates, an attempt is made to explain the categorical variable of missingness in the biomarker values.

The age over time and the age at the beginning of the follow-up show a significant (p-values of 0.001 and 0.0001) relation with the missingness of respectively 0.0033 and 0.0086. However, the association is considerably small and is expected to be more related to elderness than with a deteriorated health condition due to PBC.

If the missing values of Alkaline Phosphatase are used to explain the missing data of Serum Bilirubin and vice versa, the model shows opposite relations. The missing values related to the Serum Bilirubin are significantly negatively correlated (-0.13120 with p-value smaller than $2e-16$) with the levels of Alkaline Phosphatase. The missing data of Alkaline Phosphatase show a small positive relation with levels of Serum Bilirubin (0.03482 with a p-value smaller than $2e-16$). The negative relationship of Alkaline Phosphatase and the considerably small association with Serum Bilirubin are not corresponding to the current literature. This indicates that it is not related to the clinical outcomes, corresponding to non-informative missingness or Missing Completely At Random (MCAR).

If the missingness of the biomarker is stratified by City only two cities show large differences in percentages of missing values of Serum Bilirubin and Alkaline Phosphatase. This concerns the two Italian cities Padova and Milan as shown in table 3. The other cities show similar percentages, suggesting simultanious missing biomarker values. To extend the research of missingness, the missing values of Serum Bilirubin are regressed on Alkaline Phosphatase for the city of Padova and the missing values of Alkaline Phosphatase are regressed on the levels of Serum Bilirubin for the city of Milan. In the city of Milan, the levels of Serum Bilirubin show no significant relation with the missing values of Alkaline Phosphatase. For Padova, the missing values of Serum Bilirubin are again negatively associated with the levels of Alkaline Phosphatase. Although, higher levels of Alkaline Phosphatase are associated with a higher hazard ratio (Lammers et al. (2014)). Thereby, it is assumed that the missingness of both Alkaline Phospatase and Serum Bilirubin is not associated with increased levels of each other.

City	Missing Bili (%)	Missing Alp (%)
The Netherlands, Rotterdam	16	6
Belgium, Leuven	21	8
Spain, Barcelona	9	6
French, Paris	17	16
Italia, Padova	35	14
Italia, Milan	9	47
UK, London	2	2
Canada, Toronto	33	19
USA, Texas	3	3
USA, Rochester	16	17
USA, Seattle	24	13
Canada, Edmonton	17	9
UK, Birmingham	15	15

Table 11: Percentage missing data of the total number of observations for the biomarkers Alkaline Phosphatase and Serum Bilirubin

Albumine is not a significant regression parameter for the missing data of Serum Bilirubin or Alkaline Phosphatase with p-values of respectively 0.951 and 0.989. It is assumed that the significant association between the missing data and other available covariates are mainly determined by the size of the dataset and not associated with the clinical outcomes (death or liver transplantation). Hence, the missing values of the two biomarkers, Alkaline Phosphatase and Serum Bilirubin, are discarded from the dataset and not included in the sample to fit the joint model.

A.2 Censoring

During follow-up a total of 184 drop outs have occured. The fitted models $B9$ and $A10$ are computed without the patients that drop out to illustrate their influence on the inference regarding the fitted joint model. Discarding the drop outs does not lead to another inference about the joint models. The tables below show no large differences in the parameter estimates between the fitted models (shown in appendix A.5) with or without the drop outs.

The longitudinal process is captured by a second degree polynomial, where the fixed effects are denoted by β_p. β_0 is the intercept, β_1 the fixed coefficient corresponding to the time in years t (0 denotes the starting point of follow-up) and β_2 is corresponding to the squared time in years t^2. The individual parameter deviations, the co-called random effects, are denoted by b_p, where b_0 denotes the random effects of the intercept, b_1 and b_2 are the random effect related to respectively β_1 and β_2. In the tables below are the standard deviations and correlation structures presented. The event process is described by the coefficients of the baseline covariates, denoted by γ, the coefficient corresponding to the longitudinal measurment of the biomarker, denoted by α, and the values of the $log(\xi_q)$ are the estimated hazard ratios for the seven intervals determined by the seven knots of the piecewise-constant model. γ_0 is coefficient related to the age at the beginning of follow-up, γ_1 corresponds to the Albumine level times the Upper Limit of Normal (ULN) at 1 year of follow-up, γ_2 is associated with the platelet count per 109/L at 1 year of follow-up and γ_4 represents the logarithmic scaled level of Serum Bilirubin times the ULN at 1 year of follow-up.

Longitudinal Process					Variance Components				Event Process				
parameter	**value**	**st. error**	**Z-value**	**p-value**	**parameter**	**st. dev.**	**Corr**		**parameter**	**value**	**st. error**	**Z-value**	**p-value**
β_0	-0.4992	0.0138	-36.1361	< 0.0001	b_0	0.7294	b_0	b_1	γ_0	0.0475	0.0037	12.8921	< 0.0001
β_1	0.0210	0.0037	5.6304	< 0.0001	b_1	0.1538	-0.3800		γ_1	-2.1955	0.2757	-7.9621	< 0.0001
β_2	0.0001	0.0002	0.6893	0.4906	b_2	0.0091	0.4065	-0.8817	γ_3	-0.0012	0.0004	-2.7157	0.0066
					ϵ_i	0.3218			α	1.4954	0.0459	32.5756	< 0.0001
									$log(\xi_1)$	-4.4378	0.3921	-11.3183	
									$log(\xi_2)$	-3.6609	0.3976	-9.2080	
									$log(\xi_3)$	-3.4668	0.4005	-8.6553	
									$log(\xi_4)$	-3.1889	0.4024	-7.9248	
									$log(\xi_5)$	-3.1223	0.4078	-7.6562	
									$log(\xi_6)$	-2.8960	0.4073	-7.1104	
									$log(\xi_7)$	-2.6829	0.4140	-6.4801	

Table 12: Parameter estimates of joint model $B9$ discarding patients that drop out from the follow-up study

Longitudinal Process					Variance Components				Event Process				
parameter	**value**	**st. error**	**Z-value**	**p-value**	**parameter**	**st. dev.**	**Corr**		**parameter**	**value**	**st. error**	**Z-value**	**p-value**
β_0	0.4126	0.0139	29.7141	< 0.0001	b_0	0.7212	b_0	b_1	γ_0	0.0415	0.0035	11.9868	< 0.0001
β_1	-0.0194	0.0035	-5.6074	< 0.0001	b_1	0.1526	-0.5299		γ_1	-2.0095	0.2686	-7.4818	< 0.0001
β_2	0.0005	0.0002	2.2464	0.0247	b_2	0.0096	0.3415	-0.9309	γ_2	-0.0030	0.0005	-6.6475	< 0.0001
					ϵ_i	0.2404			γ_3	0.8805	0.0512	17.1979	< 0.0001
									α	0.7189	0.0689	10.4290	< 0.0001
									$log(\xi_1)$	-4.1484	0.3880	-10.6927	
									$log(\xi_2)$	-3.2612	0.3862	-8.4442	
									$log(\xi_3)$	-3.0421	0.3864	-7.8736	
									$log(\xi_4)$	-2.7303	0.3863	-7.0676	
									$log(\xi_5)$	-2.6639	0.3896	-6.8379	
									$log(\xi_6)$	-2.3874	0.3879	-6.1539	
									$log(\xi_7)$	-2.1160	0.3912	-5.4088	

Table 13: Parameter estimates of joint model $A10$ discarding patients that drop out from the follow-up study

A.3 EM Algorithm

In the context of the joint modeling framework, the EM algorithm represents an iterative method to optimize the joint log likelihood function. In this section of the appendix, the numerical optimization technique is explained for the joint model with the *current value* of the true and unobserved levels of the biomarker. The longitudinal and survival submodel are in this context illustrated by

$$\begin{aligned} y_i &= X_i^T \beta + Z_i^T b_i + \epsilon_i, \\ h_i(t|\mathcal{M}_i(t), w_i) &= h_0(t)\ exp(\gamma^T w_i + \alpha m_i(t)), \\ b_i &\sim \mathcal{N}(0, D), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2). \end{aligned}$$

The EM algorithm alternates between the so-called E-step and M-step to find the parameter estimates. After each iteration, the log likelihood function increases handling the random effects b_i as if it are missing data, i.e. $log\ p(y_i, T_i, \delta_i; \theta^{(it+1)}) > log\ p(y_i, T_i, \delta_i; \theta^{(it)})$. The E-step computes the expected complete data log likelihood, denoted by $Q(\theta|\theta^{it})$. In case of the joint model this is given by

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

The integrals to estimate the complete data log likelihood $Q(\theta|\theta^{(it)})$ involves two integrals that generally lead to no closed-form solution. This are the integrals regarding the survival function and the random effects. These two are approximated by the pseudo adaptive Gauss Hermite rule. After the E-step, the estimated complete data log likelihood function is maximized in the M-step to find parameter estimates, i.e. the values of $\theta_t = (\gamma^T, \alpha, \theta_{h_0}^T)^T$, $\theta_y = (\beta^T, \sigma^2)^T$ and $\theta_b = vech(D)$.

The computation of the covariance matrices regarding the measurement errors and random effects have closed-form solutions and are given by

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{N}\sum_i \int (y_i - X_i\beta - Z_ib_i)^T(y_i - X_i\beta - Z_ib_i)p(b_i|T_i, \delta_i, y_i; \theta), \\
&= \frac{1}{N}\sum_i \int (y_i - X_i\beta)^T(y_i - X_i\beta - 2Z_i\tilde{b_i}) + tr(Z_i^T Z_i v\tilde{b}_i) + \tilde{b_i}^T Z_i^T Z_i\tilde{b_i}, \\
\hat{D} &= \frac{1}{n}\sum_i v\tilde{b}_i + \tilde{b_i}\tilde{b}_i^T,
\end{aligned}$$

where $N = \sum_i n_i$, $\tilde{b_i} = E(b_i|T_i, \delta_i, y_i; \theta^{(it)}) = \int b_ip(b_i|T_i, \delta_i, y_i; \theta^{(it)})\, db_i$, and $v\hat{b}_i = var(b_i|T_i, \delta_i, y_i; \theta^{(it)}) = \int (b_i - \tilde{b_i})^2 p(b_i|T_i, \delta_i, y_i; \theta^{(it)})\ db_i$.

The other parameters' optimization do not have closed-form solutions. Hence, the score vectors of β, γ, α and θ_{h_0} are obtained using a one-step Newton-Raphson update mechanism, given by

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

where $\theta_t^{(it)}$ is the parameter value at the current iteration. The one-step Newton-Raphson update is using the Hessian matrix of the score vectors and the score vectors to obtain the new parameter estimates. The Hessian matrices $\partial S(\theta)/\partial\theta)$ are calculated with the central difference approximation technique (Press (2007)). The score vectors of β, γ, α and θ_{h_0} are computed as follows

$$\begin{aligned}
s(\beta) &= \sum_i X_i^T\{y_i - X_i\beta - Z_i\hat{b}_i\}/\sigma^2 + \alpha\delta_i x_i(Ti) \\
&\quad - exp(\gamma^T w_i)\int\int_0^{T_i} h_0(s)\alpha x_i exp\big[\alpha\{x_i^T(s)\beta + z_i^T(s)b_i\}\big] * p(b_i|T_i, \delta_i, y_i; \theta)\ ds\ db_i \\
S(\gamma) &= \sum_i w_i\left[\delta_i - exp(\gamma^T w_i)\int\int_0^{T_i} h_0(s)\alpha x_i exp\big[\alpha\{x_i^T(s)\beta + z_i^T(s)b_i\}\big] * p(b_i|T_i, \delta_i, y_i; \theta)\ ds\ db_i\right] \\
S(\alpha) &= \sum_i \delta_i\{x_i^T(T_i)\beta + z_i^T(T_i)\tilde{b_i}\} - exp(\gamma^T w_i)\int\int_0^{T_i} h_0(s)\alpha x_i exp\big[\alpha\{x_i^T(s)\beta + z_i^T(s)b_i\}\big] \\
&\quad * p(b_i|T_i, \delta_i, y_i; \theta)\ ds\ db_i \\
S(\theta_{h_0}) &= \sum_i \delta\frac{\partial h_0(T_i; \theta_{h_0})}{\partial h_0^T} exp(\gamma^T w_i)\int\int_0^{T_i} h_0(s)\alpha x_i exp\big[\alpha\{x_i^T(s)\beta + z_i^T(s)b_i\}\big] \\
&\quad * p(b_i|T_i, \delta_i, y_i; \theta)\ ds\ db_i
\end{aligned}$$

The EM algorithm alternates between the E-step and M-step untill the parameter estimates or likelihood function satisfies one of the two criteria stated in Section 3.2.1.

A.4 Parameter Estimates Joint Models

This section of the appendix contain table 14 up to table 16, which summarize the parameter estimates of the joint models ($B9$ and $A10$) with the *current value* parameterization of the longitudinal process and the weighted cumulative parameterization. The longitudinal process is captured by a second degree polynomial, where the fixed effects are denoted by β_p. β_0 is the intercept, β_1 the fixed coefficient corresponding to the time in years t (0 denotes the starting point of follow-up) and β_2 is corresponding to the squared time in years t^2. The individual parameter deviations, the co-called random effects, are denoted by b_p, where b_0 denotes the random effects of the intercept, b_1 and b_2 are the random effect related to respectively β_1 and β_2. In the tables below are the standard deviations and correlation structures presented. The event process is described by the coefficients of the baseline covariates, denoted by γ, the coefficient corresponding to the longitudinal measurment of the biomarker, denoted by α, and the values of the $log(\xi_q)$ are the estimated hazard ratios for the seven intervals determined by the seven knots of the piecewise-constant model. γ_0 is coefficient related to the age at the beginning of follow-up, γ_1 corresponds to the Albumine level times the Upper Limit of Normal (ULN) at 1 year of follow-up, γ_2 is associated with the platelet count per 109/L at 1 year of follow-up and γ_4 represents the logarithmic scaled level of Serum Bilirubin times the ULN at 1 year of follow-up.

The tables show that increased levels of Alkaline Phosphatase and Serum Bilirubin are associated with an increased hazard ratio. Hence, increased levels indicate an increased risk of experiencing an event. Lower levels of Albumine and platelet counts at 1 year of follow-up are associated with an increased risk of having an event. Being older at the beginning of follow-up increases the patients' risk of having an event. The logarithmic scaled hazard ratios of the piecewise-constant model exhibit an incline over time and their summation constitutes the baseline hazard function.

Longitudinal Process					Variance Components				Event Process				
parameter	**value**	**st. error**	**Z-value**	**p-value**	**parameter**	**st. dev.**	**corr**		**parameter**	**value**	**st. error**	**Z-value**	**p-value**
β_0	−0.4945	0.0134	−36.7860	< 0.0001	b_0	0.7223	b_0	b_1	γ_0	0.0527	0.0037	14.1309	< 0.0001
β_1	0.0198	0.0036	5.5059	< 0.0001	b_1	0.1516	−0.3717		γ_1	−1.6630	0.2763	−6.0192	< 0.0001
β_2	0.0001	0.0002	0.6823	0.4950	b_2	0.0091	0.3782	−0.8796	γ_2	−0.0010	0.0004	−2.3498	0.0188
					ϵ_i	0.3212			α	1.5675	0.0467	33.5879	< 0.0001
									$log(\xi_1)$	−5.4292	0.3987	−13.6179	
									$log(\xi_2)$	−4.6403	0.4032	−11.5088	
									$log(\xi_3)$	−4.4311	0.4055	−10.9279	
									$log(\xi_4)$	−4.1784	0.4072	−10.2606	
									$log(\xi_5)$	−4.0763	0.4122	−9.8893	
									$log(\xi_6)$	−3.8960	0.4124	−9.4466	
									$log(\xi_7)$	−3.6975	0.4194	−8.8163	

Table 14: Parameter estimates of joint model $B9$

Longitudinal Process					Variance Components				Event Process				
parameter	**value**	**st. error**	**Z-value**	**p-value**	**parameter**	**st. dev.**	**corr**		**parameter**	**value**	**st. error**	**Z-value**	**p-value**
β_0	0.4078	0.0135	30.1148	< 0.0001	b_0	0.7196	b_0	b_1	γ_0	0.0414	0.0035	11.9872	< 0.0001
β_1	−0.0185	0.0034	−5.4289	< 0.0001	b_1	0.1515	−0.5308		γ_1	−1.9409	0.2660	−7.2956	< 0.0001
β_2	0.0004	0.0002	1.9316	0.0534	b_2	0.0095	0.3436	−0.9312	γ_2	−0.0030	0.0004	−6.6828	< 0.0001
					ϵ_i	0.2404			γ_3	0.9012	0.0514	17.5430	< 0.0001
									α	0.7183	0.0691	10.3965	< 0.0001
									$log(\xi_1)$	−4.2779	0.3845	−11.1265	
									$log(\xi_2)$	−3.3849	0.3824	−8.8511	
									$log(\xi_3)$	−3.1409	0.3817	−8.2281	
									$log(\xi_4)$	−2.8655	0.3818	−7.5055	
									$log(\xi_5)$	−2.7482	0.3843	−7.1508	
									$log(\xi_6)$	−2.5104	0.3832	−6.5515	
									$log(\xi_7)$	−2.1932	0.3863	−5.6776	

Table 15: Parameter estimates of joint model $A10$

Longitudinal Process					Variance Components				Event Process				
parameter	**value**	**st. error**	**Z-value**	**p-value**	**parameter**	**st. dev.**	**corr**		**parameter**	**value**	**st. error**	**Z-value**	**p-value**
β_0	-0.4932	0.0136	-36.2952	< 0.0001	b_0	0.7316	b_0	b_1	γ_0	0.0507	0.0037	13.6833	< 0.0001
β_1	0.0186	0.0036	5.2057	< 0.0001	b_1	0.1526	-0.3947		γ_1	-1.8980	0.2705	-7.0161	< 0.0001
β_2	0.0002	0.0002	0.8963	0.3701	b_2	0.0092	0.4038	-0.8836	γ_2	-0.0012	0.0004	-2.8585	0.0043
					ϵ_i	0.3213			α	3.4344	0.1035	33.1692	< 0.0001
									$log(\xi_1)$	-4.7534	0.3942	-12.0587	
									$log(\xi_2)$	-4.0941	0.3975	-10.2996	
									$log(\xi_3)$	-3.8637	0.4005	-9.6473	
									$log(\xi_4)$	-3.5903	0.4028	-8.9123	
									$log(\xi_5)$	-3.4540	0.4081	-8.4635	
									$log(\xi_6)$	-3.2264	0.4102	-7.8647	
									$log(\xi_7)$	-2.9192	0.4225	-6.9089	

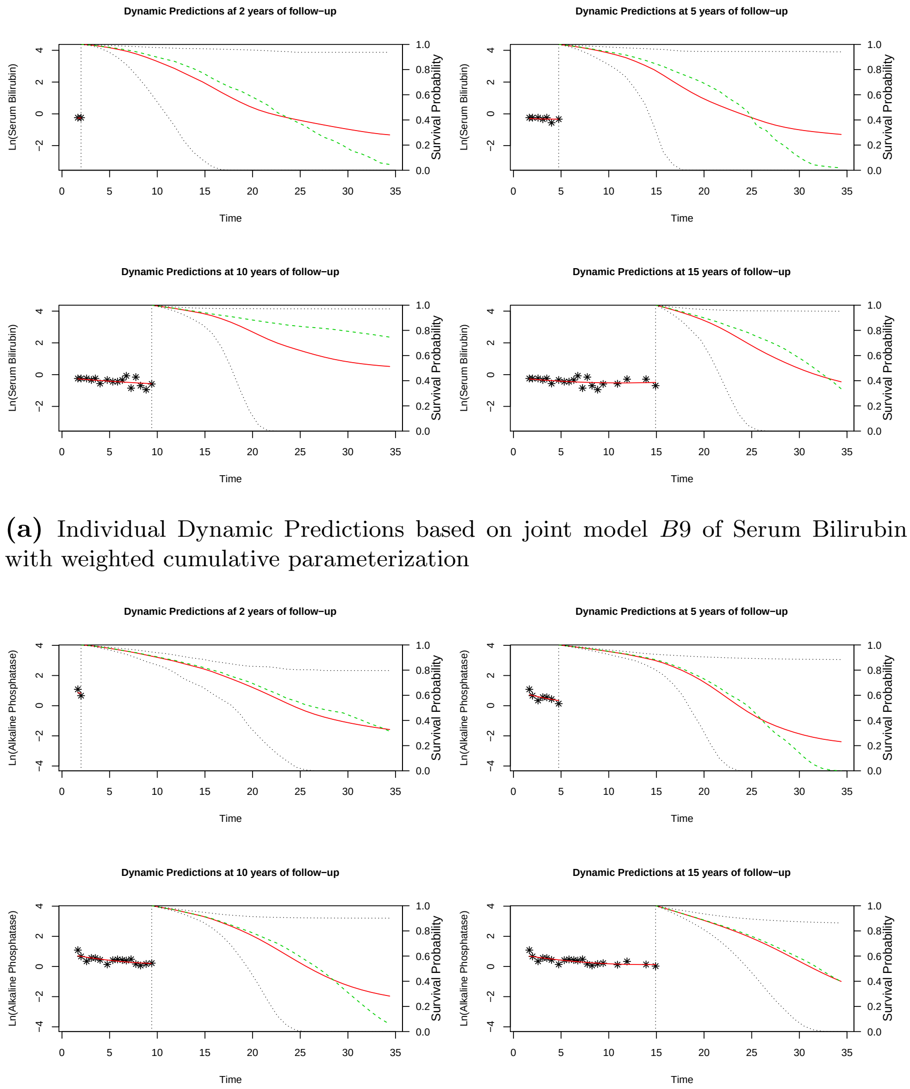
Table 16: Parameter estimates of joint model $B9$ with the cumulative parameterization

Longitudinal Process					Variance Components				Event Process				
parameter	**value**	**st. error**	**Z-value**	**p-value**	**parameter**	**st. dev.**	**corr**		**parameter**	**value**	**st. error**	**Z-value**	**p-value**
β_0	0.4068	0.0137	29.6204	< 0.0001	b_0	0.7204	b_0	b_1	γ_0	0.0435	0.0035	12.5127	< 0.0001
β_1	-0.0180	0.0036	-4.9998	< 0.0001	b_1	0.1518	-0.5336		γ_1	-1.9382	0.2671	-7.2561	< 0.0001
β_2	0.0004	0.0002	1.5393	0.1237	b_2	0.0095	0.3468	-0.9314	γ_2	-0.0032	0.0004	-7.0135	< 0.0001
					ϵ_i	0.2404			γ_3	0.8901	0.0513	17.3573	< 0.0001
									α	1.7867	0.1408	12.6902	< 0.0001
									$log(\xi_1)$	-4.3321	0.3831	-11.3072	
									$log(\xi_2)$	-3.6108	0.3851	-9.3774	
									$log(\xi_3)$	-3.3382	0.3841	-8.6913	
									$log(\xi_4)$	-3.0387	0.3838	-7.9173	
									$log(\xi_5)$	-2.8972	0.3864	-7.4983	
									$log(\xi_6)$	-2.6215	0.3857	-6.7973	
									$log(\xi_7)$	-2.2489	0.3913	-5.7472	

Table 17: Parameter estimates of joint model $A10$ with the cumulative parameterization

A.5 Plots of the Dynamic Individual Predictions

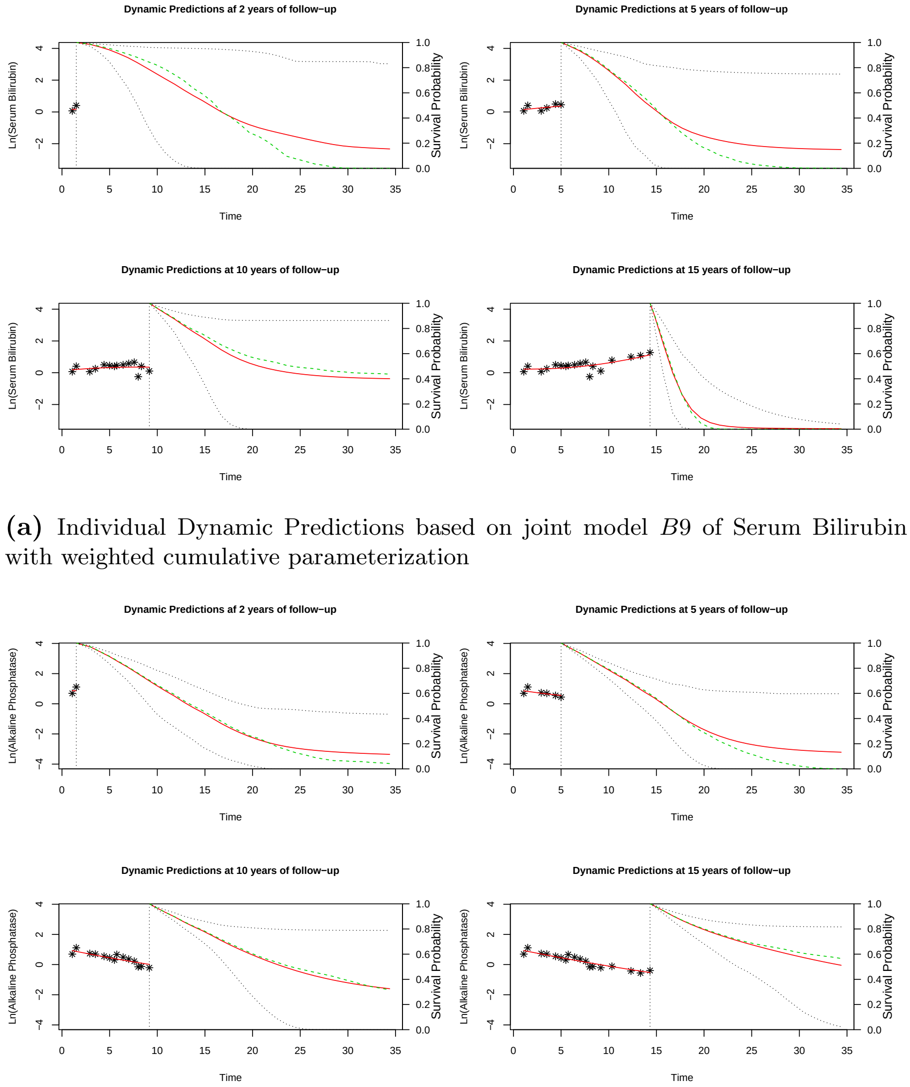
The individual dynamic predictions are obtained by the Monte Carlo simulation scheme described in section 3.3. The red line represents the mean of the 200 Monte Carlo simulations and the green line the median. The dashed line represents the 95 % confidence interval of the predicted survival curve. The time on the x-axis is in years. The left y-axis denotes the values of the logarithmic scaled biomarkers, while the right y-axis denotes the survival probabilities.



(a) Individual Dynamic Predictions based on joint model $B9$ of Serum Bilirubin with weighted cumulative parameterization

(b) Individual Dynamic Predictions based on joint model $A10$ of Alkaline Phosphatase with weighted cumulative parameterization

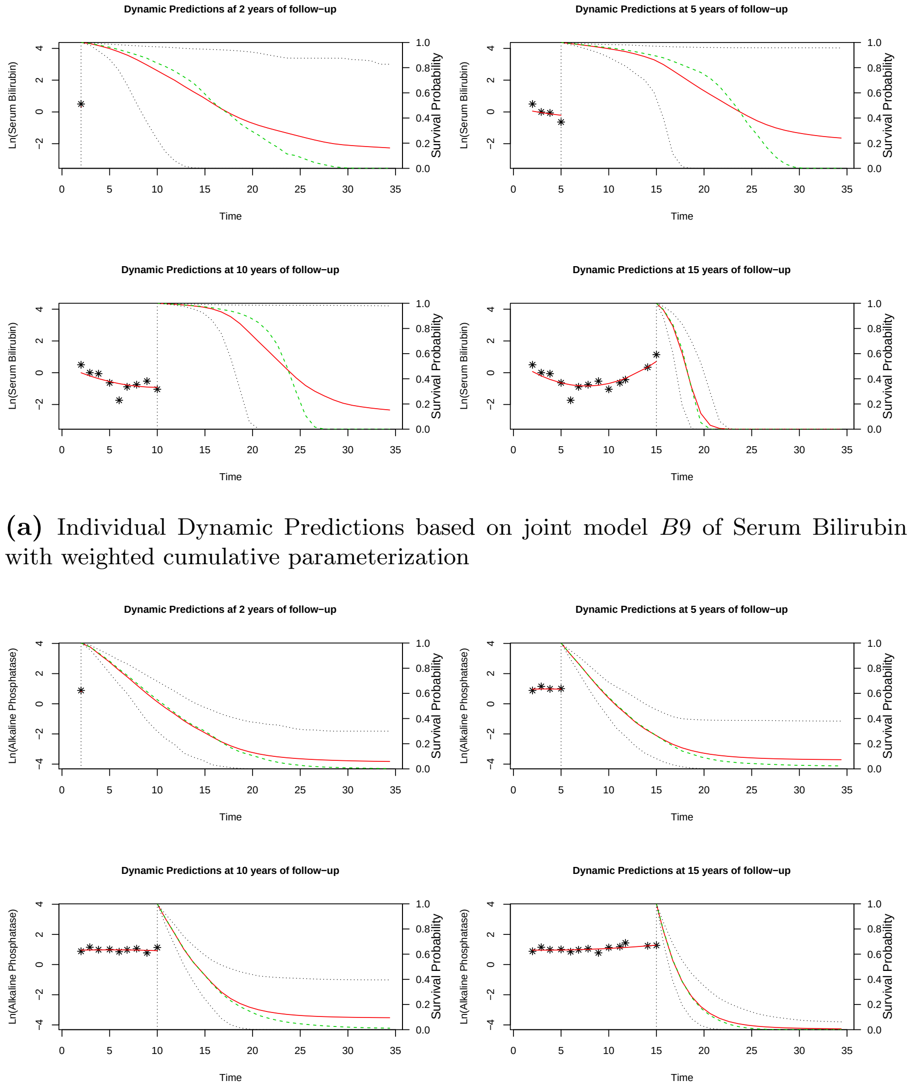
Figure 1: Individual dynamic predictions for a patient who is still alive at the end of follow-up



(a) Individual Dynamic Predictions based on joint model $B9$ of Serum Bilirubin with weighted cumulative parameterization

(b) Individual Dynamic Predictions based on joint model $A10$ of Alkaline Phosphatase with weighted cumulative parameterization

Figure 2: Individual dynamic predictions for a patient who is deceased before the end of follow-up



(a) Individual Dynamic Predictions based on joint model $B9$ of Serum Bilirubin with weighted cumulative parameterization

(b) Individual Dynamic Predictions based on joint model $A10$ of Alkaline Phosphatase with weighted cumulative parameterization

Figure 3: Individual dynamic predictions for a patient who received a liver transplantation before the end of follow-up

References

Andrinopoulou, E.-R., Rizopoulos, D., Geleijnse, M. L., Lesaffre, E., Bogers, A. J., and Takkenberg, J. J. (2015). Dynamic prediction of outcome for patients with severe aortic stenosis: application of joint models for longitudinal and time-to-event data. *BMC cardiovascular disorders*, 15(1):1.

Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, pages 45–57.

Brilleman, S. L., Crowther, M. J., May, M. T., Gompels, M., and Abrams, K. R. (2016). Joint longitudinal hurdle and time-to-event models: an application related to viral load and duration of the first treatment regimen in patients with hiv initiating therapy. *Statistics in Medicine*.

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

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

Gershwin, M. E., Mackay, I., Sturgess, A., and Coppel, R. (1987). Identification and specificity of a cdna encoding the 70 kd mitochondrial antigen recognized in primary biliary cirrhosis. *The Journal of Immunology*, 138(10):3525–3531.

Gurka, M. J. (2006). Selecting the best linear mixed model under reml. *The American Statistician*, 60(1):19–26.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338.

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

Henderson, R., Diggle, P., and Dobson, A. (2002). Identification and efficacy of longitudinal markers for survival. *Biostatistics*, 3(1):33–50.

Hsieh, F., Tseng, Y.-K., and Wang, J.-L. (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*, 62(4):1037–1043.

Kaplan, M. M. and Gershwin, M. E. (2005). Primary biliary cirrhosis. *New England Journal of Medicine*, 353(12):1261–1273.

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

Lammers, W. J., Hirschfield, G. M., Corpechot, C., Nevens, F., Lindor, K. D., Janssen, H. L., Floreani, A., Ponsioen, C. Y., Mayo, M. J., Invernizzi, P., et al. (2015). Development and validation of a scoring system to predict outcomes of patients with primary biliary cirrhosis receiving ursodeoxycholic acid therapy. *Gastroenterology*, 149(7):1804–1812.

Lammers, W. J., van Buuren, H. R., Hirschfield, G. M., Janssen, H. L., Invernizzi, P., Mason, A. L., Ponsioen, C. Y., Floreani, A., Corpechot, C., Mayo, M. J., et al. (2014). Levels of alkaline phosphatase and bilirubin are surrogate end points of outcomes of patients with primary biliary cirrhosis: an international follow-up study. *Gastroenterology*, 147(6):1338–1349.

Lindor, K. D., Gershwin, M. E., Poupon, R., Kaplan, M., Bergasa, N. V., and Heathcote, E. J. (2009). Primary biliary cirrhosis. *Hepatology*, 50(1):291–308.

Lindstrom, M. J. and Bates, D. M. (1988). Newtonraphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Liver, E. A. F. T. S. O. T. et al. (2009). Easl clinical practice guidelines: management of cholestatic liver diseases. *Journal of Hepatology*, 51(2):237–267.

Parés, A., Caballería, L., and Rodés, J. (2006). Excellent long-term survival in patients with primary biliary cirrhosis and biochemical response to ursodeoxycholic acid. *Gastroenterology*, 130(3):715–720.

Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6(3):289–296.

Prentice, R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69(2):331–342.

Press, W. H. (2007). *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press.

Proust-Lima, C. and Taylor, J. M. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach. *Biostatistics*, page kxp009.

Rizopoulos, D. (2012a). Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive gaussian quadrature rule. *Computational Statistics & Data Analysis*, 56(3):491–501.

Rizopoulos, D. (2012b). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.

Rizopoulos, D. and Ghosh, P. (2011). A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in medicine*, 30(12):1366–1380.

Rizopoulos, D., Murawska, M., Andrinopoulou, E.-R., Molenberghs, G., Takkenberg, J. J., and Lesaffre, E. (2013). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *arXiv preprint arXiv:1306.6479*.

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2008). Shared parameter models under random effects misspecification. *Biometrika*, 95(1):63–74.

Selmi, C., Bowlus, C. L., Gershwin, M. E., and Coppel, R. L. (2011). Primary biliary cirrhosis. *The Lancet*, 377(9777):1600–1609.

Sène, M., Taylor, J. M., Dignam, J. J., Jacqmin-Gadda, H., and Proust-Lima, C. (2014). Individualized dynamic prediction of prostate cancer recurrence with and without the initiation of a second treatment: Development and validation. *Statistical methods in medical research*, page 0962280214535763.

Shapiro, J., Smith, H., and Schaffner, F. (1979). Serum bilirubin: a prognostic factor in primary biliary cirrhosis. *Gut*, 20(2):137–140.

Song, X., Davidian, M., and Tsiatis, A. A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*, 58(4):742–753.

ter Borg, P. C., Schalm, S. W., Hansen, B. E., and van Buuren, H. R. (2006). Prognosis of ursodeoxycholic acid-treated patients with primary biliary cirrhosis. results of a 10-yr cohort study involving 297 patients. *The American journal of gastroenterology*, 101(9):2044–2050.

Trivedi, P. J., Lammers, W. J., van Buuren, H. R., Parés, A., Floreani, A., Janssen, H. L., Invernizzi, P., Battezzati, P. M., Ponsioen, C. Y., Corpechot, C., et al. (2015). Stratification of hepatocellular carcinoma risk in primary biliary cirrhosis: a multicentre international study. *Gut*, pages gutjnl–2014.

Tsiatis, A., Degruttola, V., and Wulfsohn, M. (1995). Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association*, 90(429):27–37.

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

Vacek, P. M. (1997). Assessing the effect of intensity when exposure varies over time. *Statistics in Medicine*, 16(5):505–513.

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer Science & Business Media.

Warnes, T. (1972). Alkaline phosphatase. *Gut*, 13(11):926–937.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339.