Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

MARIA HELENA SANTOS DE OLIVEIRA

Joint model parameterizations for longitudinal and time-to-event data

Maringá – Paraná
2023

MARIA HELENA SANTOS DE OLIVEIRA

Joint model parameterizations for longitudinal and time-to-event data

Dissertação apresentada ao Programa de Pós-graduação em Bioestatística do centro de ciências exatas da Universidade Estadual de Maringá como requisito parcial para obtenção do título de mestre em Bioestatística.

Orientadora: Profa. Dra. Isolde Terezinha Santos Previdelli, Universidade Estadual de Maringá
Membro externo: Prof. Dr. Enrico Antônio Colosimo, Universidade Federal de Minas Gerais
Membro do PBE: Prof. Dr. Edson Zangiacomi Martinez, Universidade Estadual de Maringá

Universidade Estadual de Maringá - UEM

Departamento de Estatística - DES

Programa de Pós-Graduação em Bioestatística

Maringá – Paraná

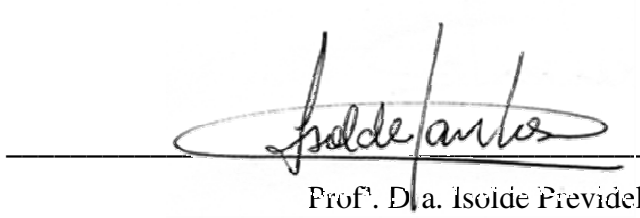
2023

MARIA HELENA SANTOS DE OLIVEIRA

Parametrizações do modelo conjunto para dados longitudinais e de sobrevivência

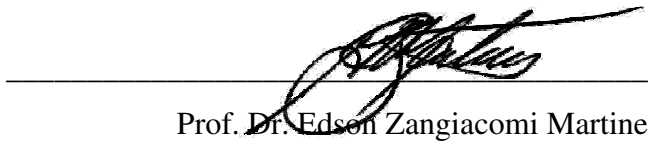
Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

BANCA EXAMINADORA



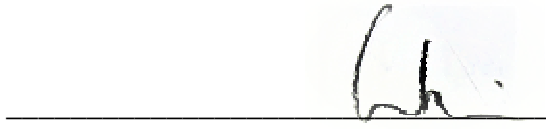
Prof^. Dr^a. Isolde Previdelli

Universidade Estadual de Maringá – PBE/UEM



Prof. Dr. Edson Zangiacomi Martinez

Universidade de São Paulo – PBE/USP



Prof. Dr. Enrico Antônio Colosimo

Universidade Estadual de Maringá - UFMG

Maringá, 14 de março de 2023.

Agradecimentos

Agradeço à minha família pelo apoio inquestionável durante a pós graduação.

Ao meu incrível companheiro Guilherme, por embarcar nessa jornada comigo e apoiar todas as minhas decisões com tanta confiança.

Às minhas grandes amigas Giovanna Bertolini, Beatriz Brito, Thaís Barbosa e Karla Sacani, que nunca duvidaram de mim.

Aos meus amigos e colegas de profissão que me tranquilizaram durante essa jornada, em especial ao Vinicius Riffel, que me ajudou a ver minhas dificuldades por outro ponto de vista.

Ao Dr. Brandon Michael Henry, que me apresentou o mundo da pesquisa, acreditou no que eu poderia fazer e fez tudo que estava ao seu alcance para me ajudar a chegar onde eu queria.

À minha orientadora, professora Isolde Previdelli, que sempre me deu liberdade e confiança, e que prezou pelo meu sucesso em todos os momentos.

À coordenação e ao corpo docente do programa de pós graduação em Bioestatística, pela sua dedicação e empenho em fazer do programa um espaço de excelência acadêmica, sem que isso custasse o acolhimento aos alunos.

Agradeço à CAPES pelo apoio financeiro, que possibilitou a minha dedicação à pós graduação.

Por fim, agradeço aos mentores, amigos e colegas que me ajudaram a ver o mundo além do mestrado.

Resumo

A observação de desfechos de sobrevivência frequentemente requer algum tipo de acompanhamento dos indivíduos em um estudo, sendo comum coletar dados longitudinal e de sobrevivência concomitantemente. Pela perspectiva da análise de dados longitudinais, a sobrevivência pode ser uma fonte de perda não ignorável, enquanto do ponto de vista de análise de sobrevivência, biomarcadores observados ao longo do tempo de acompanhamento podem se comportar como variáveis endógenas dependentes do tempo. O *framework* de modelos conjuntos se propõe a lidar com estas situações combinando modelos lineares de efeitos mistos com a regressão de riscos proporcionais através de uma função de verossimilhança conjunta. Neste estudo, diferentes possibilidades de ligações entre os dois processos são apresentadas na forma de parametrizações, e aplicadas a dois bancos de dados biomédicos. Primeiro, analisamos medidas de cloro sérico tomadas diariamente em pacientes com COVID-19 internados em unidade de tratamento intensivo. Segundo, exploramos a associação entre a contagem longitudinal de linfócitos CD4 e o tempo até a recuperação imunológica de pacientes com HIV. Cada parametrização permite uma interpretação biológica diferente sobre a dinâmica subjacente das doenças estudadas, e a parametrização de inclinação tempo-dependente, em particular, se mostra especialmente útil nas situações estudadas.

Palavras-chave: Modelos conjuntos. Análise de sobrevivência. Dados longitudinais.

Abstract

The observation of survival outcomes often requires some type of follow-up of the individuals of a study, and therefore it is common to collect both longitudinal and survival data concurrently. From the longitudinal data analysis perspective, survival may be a source of non-ignorable missingness, while from the survival standpoint, biomarkers collected across follow-up time may behave as endogenous time-varying covariates. The joint model framework proposes to deal with this situation by combining linear mixed-effects models and proportional hazards regression through a joint likelihood function. In this study, different possibilities of linking the two processes are presented in the form of parameterizations of the joint model, and applied to two sets of biomedical data. First, we analyzed serum chloride measurements taken daily from COVID-19 patients admitted to an intensive care unit and its association with patient survival. Second, we explore the association between longitudinal CD4 lymphocyte count and the time to immunologic recovery in HIV patients. Each parameterization applied to the datasets allows for a different biological interpretation of the underlying dynamics of the diseases studied, and the time-dependent slopes parameterization, in particular, appears especially useful in these settings.

Keywords: Joint models. Survival analysis. Longitudinal data.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

In biomedical research, it is important to look for statistical methods that allow not only to make precise predictions, but also to explain the underlying mechanisms that surround a biological process or a disease. In cohort studies or clinical trials that investigate the course of a disease or a treatment, it is common to have survival as an outcome of interest, as well as other time-to-event outcomes. These scenarios require longitudinal monitoring of study participants, and consequently statistical methods that accommodate longitudinal data.

When conducting a survival study, it is expected to have some type of repeated measure of each subject while waiting for the event of interest to occur. Although possible, the use of proportional hazards regression with time-dependent covariates for this situation might not be adequate, especially if the longitudinal variable happens to be an endogenous one (WU et al., 2011), meaning it is a measure of each individual and its measurement depends on the event not happening. This is the case when the event studied is death and the longitudinal covariate is a biomarker. This context also increases the complexity in modeling the longitudinal data itself, as the loss of follow-up caused by a patient's death may be the case of non-ignorable missingness, if the values of the biomarker are related to the the dropout process, which in this case is death. In the presence of data that is missing not at random (MNAR), joint models are known to provide less biased results (WU et al., 2011) for longitudinal models, and allow for survival time to be modeled in the presence of time-dependent internal covariates, usually not accommodated by traditional survival analysis (RIZOPOULOS, 2012).

The earliest developments in joint models for longitudinal and survival data were motivated by the study of Human Immunodeficiency Virus (HIV) and Acquired Immunodeficiecy Syndrome (AIDS) (GRUTTOLA; TU, 1994; TSIATIS; DEGRUTTOLA; WULFSOHN, 1995), as the white blood cell type CD4 is known to provide valuable insight into the disease progression when monitored longitudinally. In an effort to tackle the issue of survival estimation with en-

dogenous covariates, early approaches (SELF; PAWITAN, 1992) focused on two-stage methods for estimation, meaning the longitudinal model would be estimated first, independently of the survival information, and these fixed and random effects would be used to produce estimates for the longitudinal marker at each time point t, that would then be introduced as the covariate to estimate the survival model. This procedure's main advantage was its computational simplicity. However, ignoring the informative dropouts generated by the occurance of events in the estimation of longitudinal trajectories can lead to biased results (WU et al., 2011). In addition, these methods do not incorporate the uncertainty associated with the estimates of the longitudinal marker, and therefore may lead to underestimated standard errors for the parameter estimates.

The joint model framework allows two submodels, a longitudinal one and a survival one, to be connected, as the survival submodel incorporates some characteristics of the longitudinal submodel. This conection can be specified in various ways, leading to different parameterizations with different biological interpretations. What defines this framework is that both outcomes are modeled simultaneously, considering a conditional joint density (CEKIC et al., 2021). When the two outcomes are in fact correlated and the longitudinal variable is an endogenous one, this approach leads to better accuracy in the estimated parameters of each model (CEKIC et al., 2021; IBRAHIM; CHU; CHEN, 2010). By providing a measure of the effect of some aspect of the longitudinal trajectory on the survival outcome, joint modeling becomes more informative than the independent modeling of longitudinal and survival data separately.

In the present work, we will introduce the theory and usage of joint models by applying different parameterizations to two sets of data. First, we aim to examine serum chloride concentration alterations in COVID-19 patients being treated in the intensive care unit (ICU) of the Security Forces Hospital in Saudi Arabia, measured daily from admission to discharge or death, and explore the association between these longitudinal measures and patient survival. This study has been submitted as an original article to the journal *Revista Brasileira de Terapia Intensiva*. In addition, we also extend a previous analysis of HIV patients coinfected with the hepatitis B virus (HBV) and the hepatitis C virus (HCV).

1.1 Objectives

The goal of this study is to present different parameterizations of the joint model methodology for longitudinal and survival data.

1.1.1 Specific Objectives

- To intruduce the joint model methodology and each parameterization.
- To analyze serum chloride and COVID-19 survival data using different parameterizations of the joint model and compare the results
- To analyze CD4 lymphocyte count and its relationship to immunologic recovery using different parameterizations of the joint model and compare the results
- To analyze the residuals of each model fit.
- To showcase the advantages of the joint model methodology in applied research.

Chapter 2

Methods

In this chapter, we introduce the fundamental concepts of the analysis of longitudinal and survival data, from the separate treatment of each type of data to the joint modelling of the two.

2.1 Longitudinal Data

Frequently encountered in the medical field, longitudinal data are characterized by the repeated measures of one or more variables in the same set of subjects over time. It is expected that repeated measures of the same subject are positively correlated, but measures from different subjects are independent of each other. This correlation structure must be taken into account when dealing with such data. Other common features of longitudinal data are the potential for missing and unbalanced measures, meaning different subjects might have a different number of observations, as well as have been observed at different time intervals.

A common methodology used to accommodate existing correlation while modeling the changes in the response variable over time is the mixed-effects (ME) regression model. A general linear ME model can be defined as

$$\begin{cases} \boldsymbol{y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \\ \boldsymbol{b}_i \sim \mathcal{N}(0, D), \\ \boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}_{n_i}), \end{cases} \tag{2.1.1}$$

where $\boldsymbol{y}_i$ is a vector of responses of dimension n_i that assumes values y_{ij} for the ith individual at the jth time point. X_i and Z_i are known design matrices, for the fixed-effects regression coefficients $\boldsymbol{\beta}$ and the random-effects regression coefficients $\boldsymbol{b}_i$. A multivariate normal distribution with mean zero and variance-covariance matrix D is assumed for the random

effects, which are independent of the error terms ε_i, also normally distributed with mean zero and variance matrix $\sigma^2 I_{n_i}$. Responses from the same subject at different time points are conditionally independent, given the covariates and random effects, and have conditional normal distributions.

Given that the marginal density of the observed response variable for the ith subject is an n_i-dimensional Normal distribution with mean $X_i\boldsymbol{\beta}$ and variance-covariance matrix $V_i = Z_i D Z_i' + \sigma^2 I_{n_i}$, and the maximum likelihood estimator for the vector of fixed effects $\boldsymbol{\beta}$ is dependent on V_i, the linear ME model parameters can be estimated through Restricted Maximum Likelihood (REML), where

$$\hat{\boldsymbol{\beta}} = \left(\Sigma_{i=1}^{n} X_i' V_i^{-1} X_i\right)^{-1} \Sigma_{i=1}^{n} X_i' V_i^{-1} \boldsymbol{y}_i \tag{2.1.2}$$

corresponds to the generalized least squares estimator, and $\hat{V}_i$ is obtained by maximizing the modified log-likelihood function, corrected by the term:

$$-\frac{1}{2} \log |\Sigma_{i=1}^{n} X_i' V_i^{-1} X_i|, \tag{2.1.3}$$

usually with the aid of a numerical optimization algorithm such as the quasi-Newton Broyden–Fletcher–Goldfarb–Shanno method implemented in the nlme package (PINHEIRO; BATES; R Core Team, 2022), given the great complexity of obtaining the parameter estimates analitically. More detail on the REML method can be found in Rizopoulos (2012) and Diggle et al. (2013)

Once estimated, the fixed effects can be interpreted similarily to the coefficients of regular linear regression, and represent covariate effects at the population level, while an individual's random effects represent that subject's deviation from the population mean. To accomodate longitudinal trajectories that start at different points for different individuals, a random intercept term, b_{0i}, is added to the fixed intercept term β_0 and therefore allows individuals baseline measurements to vary. This structure will induce a correlation structure for observations of the same individual that is compound symmetric, meaning each pair of two observations of the same individual are equally correlated independent of their distance in time. For $b_i \sim \mathcal{N}(0, \sigma_b^2)$, the implied marginal covariance structure takes the form

$$V_i = \sigma_b^2 1_{n_i} 1_{n_i}' + \sigma^2 I_{n_i}, \tag{2.1.4}$$

1_{n_i} denoting the n_i-dimensional unit vector. This structure assumes constant variance over time as well as equal positive correlation ρ between the measurements of any two time points, which can be referred to as intra-class correlation coefficient.

To allow for random slopes over time as well as random intercepts, another random effect b_{1i} is introduced to the model, accomodating trajectories that have different evolutions over

time. The presence of both random effects will induce a different marginal covariance function for observations of the same individual,

$$cov(y_{ij}, y_{ij'}) = d_{22}t_{ij}t_{ij'} + d_{12}(t_{ij} + t_{ij'}) + d_{11} + \sigma^2, \tag{2.1.5}$$

where d_{kl} represent the elements of the random effects covariance matrix D. For the same time point $t_{ij} = t_{ij'} = t$ the variance function is still dependent on t and therefore this model has heteroscedastic error terms, and it is expected that variance increases over time, while the correlations decrease (RIZOPOULOS, 2012). The linear ME model can also be extended to account for further correlation in the data by allowing a more general covariance matrix for subject-specific errors Σ_i, such that $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \Sigma_i)$. This matrix can have various structures which lead to different types of serial correlation functions.

According to Rizopoulos (2012), ME models are commonly used in the joint modeling framework for longitudinal and time to event data because of their ability to predict individual trajectories of the response variable over time, as well as their flexibility when it comes to unbalanced data. Not only does this methodology account for correlation within individuals in a parsimonious way, it also does not require different individuals to have the same number of observations, or for the observations to be made at the same set of time points.

2.2 Survival Analysis

In the field of biostatistics, it is common to have interest in studying the time until the occurrence of a certain event, a variable commonly referred to as failure time. The main characteristic of this type of data is the presence of incomplete information, known as censoring. Censoring occurs when a subject's follow up is interrupted and the event that is being evaluated is not in fact observed, leading to a partial observation of the actual failure time (COLOSIMO; GIOLO, 2006).

Extending the use of traditional statistics, the field of Survival Analysis provides methods that appropriately incorporate censored observations. This study focuses on the most common type of censoring, known as right censoring, which refers to the situation where a censored subject's true failure time is unknown, but is known to be greater than the observed time (RIZOPOULOS, 2012). This type of censoring occurs, for example, when the data collection period of a study ends before every subject has experienced the event. It can also be caused by subjects who choose to drop out of the study or experience the event for a reason different than the one being studied.

When characterizing survival data, an indicator variable δ_i is defined, which assumes value 1 when the i-th subject has experienced the event, and zero when it is right-censored. Then,

$$\delta_i = \begin{cases} 1 & \text{if subject } i \text{ has failed} \\ 0 & \text{if subject } i \text{ is censored, for } i = 1, 2, ..., n. \end{cases} \tag{2.2.1}$$

In addition, the survival function is defined in terms of a continuous random variable T^*, which denotes the failure times, and $p(.)$, the corresponding probability density function. $S(t)$ expresses the probability that the event occurs after t, or that a subject survives time t,

$$S(t) = Pr(T^* > t) = \int_t^\infty p(s)ds. \tag{2.2.2}$$

A survival function $S(t)$ must be nonincreasing as t increases, and $S(t = 0)$ always equals one unit, meaning at time zero none of the subjects has experienced the event, and the probability of surviving cannot increase over time.

The hazard function also plays an important role in survival analysis, describing the instantaneous risk of an event in the time interval $[t, t + dt)$, given survival up to time t. The hazard function is defined as

$$h(t) = \lim_{dt \to 0} \frac{Pr(t \leq T^* < t + dt | T^* \geq t)}{dt}, \tag{2.2.3}$$

and can also be referred to as risk function. The survival and the risk functions can be expressed in terms of each other, as

$$S(t) = \exp\{-H(t)\} = \exp\Big\{ - \int_0^t h(s)ds \Big\}, \tag{2.2.4}$$

where $H(t)$ describes the accumulated risk until time t and is known as the cumulative risk or cumulative hazard function.

The general interest when conducting survival data analysis is to estimate the survival function or the hazard function from the available data. Nonparametrically, the survival function can be estimated through the Kaplan-Meier (KM) estimator, proposed in 1958 (KAPLAN; MEIER, 1958). This estimator does not assume any underlying probability distribution for the failure times, and provides a survival curve that is based solely on the observed $\{T_i^*, \delta_i\}$. Given r_i the number of subjects at risk at each distinct observed failure time t_i, and d_i the number of events that occur at t_i, the KM estimator is

$$\hat{S}_{KM}(t) = \prod_{i:t_i \leq t} \frac{r_i - d_i}{r_i}. \tag{2.2.5}$$

Estimation of survival function can also be based on the maximum likelihood method, when $S(t)$ is assumed to have a specific parametric form. When constructing the likelihood function, censoring must be taken into account, and subjects who experience the event contribute more

information than censored observations. Let's assume $\{T_i, \delta_i\}, i = 1, ..., n$, a random sample from a distribution function parameterized by $\boldsymbol{\theta}$, with probability density function $p(t; \boldsymbol{\theta})$. A subject i contributes $p(T_i; \boldsymbol{\theta})$ to the likelihood when the event occurs for that subject on time T_i, and contributes $S_i(T_i; \boldsymbol{\theta})$ to the likelihood when the subject is censored at time T_i. Thus, we obtain the log-likelihood function

$$\ell(\theta) = \sum_{i=1}^{n} \delta_i \log p(T_i; \boldsymbol{\theta}) + (1 - \delta) \log S_i(T_i; \boldsymbol{\theta}), \tag{2.2.6}$$

which can also be rewritten in terms of the hazard function as

$$\ell(\theta) = \sum_{i=1}^{n} \delta_i \log h_i(T_i; \boldsymbol{\theta}) - \int_0^{T_i} h_i(s; \boldsymbol{\theta}) ds. \tag{2.2.7}$$

Maximum likelihood estimates for $\boldsymbol{\theta}$ can be achieved through iterative procedures such as the Newton-Raphson algorithm, and inference can be made under classical asymptotic maximum likelihood theory (COX; HINKLEY, 1979).

In addition to parametric and non-parametric approaches, a semi-parametric method known as the proportional hazards model has been widely utilized in the medical field for the analysis of survival data. The model assumes that covariates have multiplicative effects on the event's hazard, such that

$$h_i(t|\boldsymbol{w}_i) = h_0(t) \exp(\boldsymbol{\gamma}' \boldsymbol{w}_i), \tag{2.2.8}$$

where $\boldsymbol{w}_i' = (w_{i1}, ..., w_{ip})$ corresponds to the covariate vector and $\boldsymbol{\gamma}$ the vector of respective regression coefficients. The $h_0(t)$ function is the baseline hazard function, the hazard function of a subject whose $\boldsymbol{\gamma}' \boldsymbol{w}_i = 0$, and is treated nonparametrically, while a parametric form is assumed for the covariate effects. Given the model in log scale,

$$\log h_i(t|\boldsymbol{w}_i) = \log h_0(t) + \gamma_1 w_{i1} + \gamma_2 w_{i2} + ... + \gamma_p w_{ip}, \tag{2.2.9}$$

a regression coefficient γ_j denotes the change in the log hazard at any time point t caused by an increase of one unit in w_j and no change in other predictors, and $\exp(\gamma_j)$ denotes the hazard ratio for a one unit change in the corresponding predictor at any time t.

2.3 Missing Data in Longitudinal Studies

It is a common occurrence that participants of a longitudinal study will not be available for all follow-up visits, resulting in some or many missing values in covariates. Dropouts due to many reasons are also common, leading to unknown outcomes in the dataset. This is especially true when dealing with longitudinal data in the survival context, since censoring usually prevents

measurements of any variables to be taken, and the occurrence of the event of interest, when the event of interest is death, can also prevent the measurement of endogenous covariables, such as laboratory examinations.

Missing data can be summarized by three categories according to the mechanism that leads to their missing status. To define these categories, as done by Rizopoulos (2012), let's introduce an observed data indicator r_{ij}, which assumes value 1 when y_{ij} is observed, and 0 when it is missing. The vector $\boldsymbol{r}_i = (r_{i1}, ..., r_{in_i})'$ contains information for the response vector $\boldsymbol{y}_i$, which can be partitioned into two subvectors $\boldsymbol{y}_i^o$, containing the observed data, and $\boldsymbol{y}_i^m$, containing the missing data.

Missing data mechanisms are defined according to the probability of $\boldsymbol{r}_i$, conditional to the response vector $\boldsymbol{y}_i = \boldsymbol{y}_i^o + \boldsymbol{y}_i^m$ and the corresponding parameter vector $\boldsymbol{\theta}_r$.The Missing Completely at Random (MCAR) mechanism is present when the probability of observations being missing is not related to either $\boldsymbol{y}_i^m$ or $\boldsymbol{y}_i^o$, that is, not dependent on observed values or on the unobserved ones. This mechanism allow for the observed data $\boldsymbol{y}_i^o$ to be considered a random sample of the complete data $\boldsymbol{y}_i$, such that the distribution of the observed data and the distribution of the complete data are the same. These results imply that there is no harm in ignoring the process generating the missing data, as any common statistical methods applied to the data should provide valid inferences.

Missing at Random (MAR) differs to MCAR in that this mechanism assumes the probability of an observation being missing is related to the set of available observations, although like in MCAR it is unrelated to the values that are missing. Therefore, when data are MAR,

$$p(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{\theta}_r) = p(\boldsymbol{r}_i|\boldsymbol{y}_i^o, \boldsymbol{\theta}_r), \tag{2.3.1}$$

meaning $\boldsymbol{r}_i$ is conditionally independent of $\boldsymbol{y}_i^m$ given $\boldsymbol{y}_i^o$. MAR data is the case of random dropouts, when an individual's lack of follow up is related only to the observed values of $\boldsymbol{y}_i$. In these cases, observed data cannot be considered a random sample of the complete vector of observations $\boldsymbol{y}_i$, as their distributions do not coincide. Likelihood-based analyses that only take into account the observed data can still provide valid inferences, given that the model for $\boldsymbol{y}_i$ has been specified correctly.

Lastly, when the probability of not observing a longitudinal response is related to the missing values themselves, the missing data mechanism is called Missing Not at Random (MNAR), also referred to as nonrandom dropout. Similarly to MAR cases, under a MNAR process the observed data cannot be considered a random sample of the complete observations, and the distribution of $\boldsymbol{y}_i^m$ given $\boldsymbol{y}_i^o$ depends on both $\boldsymbol{y}_i^o$ and $p(\boldsymbol{r}_i|\boldsymbol{y}_i)$. In these cases, the missingness process must be considered in the analysis, and we can only obtain valid inferences from analyses that consider the joint distribution of the response process and missing process, such

as shared parameter models, which introduce random effects that capture the association between the two processes. In this framework, given $\boldsymbol{\theta}$ the parameter vector of the joint distribution, $\boldsymbol{\theta}_y$ the parameter vector of the measurement model and $\boldsymbol{\theta}_b$ the parameters of the random effects covariance matrix,

$$p(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{r}_i; \boldsymbol{\theta}) = \int p(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | \boldsymbol{b}_i; \boldsymbol{\theta}_y) p(\boldsymbol{r}_i | \boldsymbol{b}_i; \boldsymbol{\theta}_r) p(\boldsymbol{b}_i; \boldsymbol{\theta}_b) d\boldsymbol{b}_i, \tag{2.3.2}$$

meaning the two processes are assumed conditionally independent given the random effects. Shared parameter models can also be refferred to as Joint Models, as has been done in this text.

When in the specific situation of jointly modelling a longitudinal biomarker and a survival outcome, it may not be conceptually reasonable to consider the values of the longitudinal outcome after the occurrence of the event, that is, after the subject is deceased (KURLAND et al., 2009). However, Rizopoulos (2012) point out that when assuming a mixed effects model for the observed longitudinal responses, the joint model implicitly makes assumptions for the complete response vector $\boldsymbol{y}_i$, where $\boldsymbol{y}_i^o$ contains every longitudinal measure of subject i before the event time and $\boldsymbol{y}_i^m$ the longitudinal measures that would have been observed had the event not occurred. Given T_i^* the failure time,

$$p(T_i^* | \boldsymbol{y}_i^o, \boldsymbol{y}_i^m; \boldsymbol{\theta}) = \int p(T_i^* | \boldsymbol{b}_i, \boldsymbol{\theta}) p(\boldsymbol{b}_i | \boldsymbol{y}_i^o, \boldsymbol{y}_i^m; \boldsymbol{\theta}) d\boldsymbol{b}_i, \tag{2.3.3}$$

indicating the time to dropout (event occurance) depends on $\boldsymbol{y}_i^m$ and the corresponding missing data mechanism is MNAR. This is conditional on the two processes, longitudinal and survival, being dependent on each other, and sharing random effects, a characteristic that can be inferred on through the association parameters introduced to the survival model. When these parameters are not significant, i.e. equal zero, the dropout process corresponds to MCAR, and the two outcomes can be modeled separately. Additionally, dropouts in the longitudinal measuring process may also come from censoring, which in this framework is assumed to be noninformative and therefore dependent only on the observed history of the longitudinal biomarker that precedes censoring, characterizing a MAR mechanism.

2.4 Joint Models

Joint models estimated by a joint likelihood method can be a more effective approach to study both the longitudinal and time-to-event outcomes simultaneously (WU et al., 2011). These models are composed of two sub-models, which are connected by either a shared random effect or a coefficient, and all parameters are estimated simultaneously.

To specify the joint model, we introduce a term $m_i(t)$ which denotes the true value of the longitudinal variable at time t. This term is different than the observed values $y_i(t)$, given that

these are contaminated with measurement error and may not have been observed for every time point t. A relative risk model that quantifies the association between $m_i(t)$ and the risk for an event can be written as

$$\begin{aligned} h_i(\mathcal{M}_i(t), \boldsymbol{w}_i) &= \lim_{dt \to 0} Pr\{t \leq T_i^* < t + dt | T_i^* \geq t, \mathcal{M}_i(t), \boldsymbol{w}_i\}/dt \\ &= h_0(t) \exp\{\boldsymbol{\gamma}' \boldsymbol{w}_i + \alpha m_i(t)\}, t > 0, \end{aligned} \tag{2.4.1}$$

where $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ is the history of the unobserved longitudinal process up to time point t, $h_0(.)$ is the baseline risk function and $\boldsymbol{w}_i$ a vector of baseline covariates with corresponding regression coefficients $\boldsymbol{\gamma}$. Parameter α quantifies the effect of the longitudinal outcome on the risk for an event, such that $\exp\{\alpha\}$, in this parameterization, represents the relative increase in the risk of an event that results from one unit of increase in $m_i(t)$ at a certain time point. Additionally, parameter α represents the connection through which the survival and the longitudinal submodels share the same random effects. When this parameter is equal to zero, the survival outcome no longer depends on the longitudinal outcome or the random effects, and the parameters for each model can be estimated separately (RIZOPOULOS, 2012), although joint model estimates will still be valid.

While in semi-parametric proportional hazards modeling we can usually leave the baseline risk function $h_0(.)$ completely unspecified without any issue, that is not true for joint modelling. Although possible (WULFSOHN; TSIATIS, 1997), leaving the baseline risk function unspecified may result in underestimated standard errors for the parameter estimates (HSIEH; TSENG; WANG, 2006). To avoid this, we can attribute a parametric distribution to $h_0(.)$, such as the Weibull distribution, or opt for a more flexible specification based on splines.

2.4.1 Maximum Likelihood Estimation

Estimation of the joint model is based on the joint distribution of the observed outcomes $\{T_i, \delta_i, \boldsymbol{y}_i\}$, where the vector $\boldsymbol{y}_i$ contains the observations of the ith individual at each available time point j, and assuming that the random effects $\boldsymbol{b}_i$ account not only for the correlation between the repeated measurements in the longitudinal data but also the association between the longitudinal and the event outcomes. Formally, we define $\boldsymbol{\theta} = (\boldsymbol{\theta}_t', \boldsymbol{\theta}_y', \boldsymbol{\theta}_b')'$ the full parameter vector, where $\boldsymbol{\theta}_t$ corresponds to the parameters for the event time outcome, $\boldsymbol{\theta}_y$ to the parameters for the longitudinal outcome and $\boldsymbol{\theta}_b$ the parameters of the random-effects covariance matrix. Therefore, the event outcomes and the longitudinal outcome are conditionally independent, given the random effects and parameters, and longitudinal measurements of the same subject are also independent given the random effects and parameters,

$$p(T_i, \delta_i, \boldsymbol{y}_i | \boldsymbol{b}_i; \boldsymbol{\theta}) = p(T_i, \delta_i | \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{y}_i | \boldsymbol{b}_i; \boldsymbol{\theta}), \text{and} \tag{2.4.2}$$

$$p(\boldsymbol{y}_i|\boldsymbol{b}_i;\boldsymbol{\theta}) = \prod_j p\{y_i(t_{ij})|\boldsymbol{b}_i;\boldsymbol{\theta}\}. \tag{2.4.3}$$

The log-likelihood contribution for the ith subject can be defined as

$$\begin{aligned}\log p(T_i,\delta_i,\boldsymbol{y}_i;\boldsymbol{\theta}) &= \log \int p(T_i,\delta_i,\boldsymbol{y}_i,\boldsymbol{b}_i;\boldsymbol{\theta})d\boldsymbol{b}_i \\ &= \log \int p(T_i,\delta_i|\boldsymbol{b}_i;\boldsymbol{\theta}_t,\boldsymbol{\beta})\Big[\prod_j p\{y_i(t_{ij})|\boldsymbol{b}_i;\boldsymbol{\theta}_y\}\Big]p(\boldsymbol{b}_i;\boldsymbol{\theta}_b)d\boldsymbol{b}_i,\end{aligned} \tag{2.4.4}$$

where the conditional density for the survival part $p(T_i,\delta_i|\boldsymbol{b}_i;\boldsymbol{\theta}_t,\boldsymbol{\beta})$ takes the form

$$\begin{aligned}p(T_i,\delta_i|\boldsymbol{b}_i;\boldsymbol{\theta}_t,\boldsymbol{\beta}) &= h_i(T_i|\mathcal{M}_i(T_i);\boldsymbol{\theta}_t,\boldsymbol{\theta})^{\delta_i} S_i(T_i|\mathcal{M}_i(T_i);\boldsymbol{\theta}_t,\boldsymbol{\beta}) \\ &= \Big[h_0(T_i)\exp\{\boldsymbol{\gamma}'\boldsymbol{w}_i + \alpha m_i(T_i)\}\Big]^{\delta_i} \\ &\quad \exp\Big(-\int_0^{T_i} h_0(s)\exp\{\boldsymbol{\gamma}'\boldsymbol{w}_i + \alpha m_i(s)\}ds\Big),\end{aligned} \tag{2.4.5}$$

and the joint density for the longitudinal responses together with the random effects is given by

$$\begin{aligned}p(\boldsymbol{y}_i|\boldsymbol{b}_i;\boldsymbol{\theta})p(\boldsymbol{b}_i;\boldsymbol{\theta}) &= \prod_j p\{y_i(t_{ij})|\boldsymbol{b}_i;\boldsymbol{\theta}_y\}p(\boldsymbol{b}_i;\boldsymbol{\theta}_b) \\ &= (2\pi\sigma^2)^{-\frac{n_i}{2}}\exp\{-||\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b}_i||^2/2\sigma^2\} \\ &\quad (2\pi)^{\frac{q_b}{2}}\det(D)^{-\frac{1}{2}}\exp(-\boldsymbol{b}_i'D^{-1}\boldsymbol{b}_i/2),\end{aligned} \tag{2.4.6}$$

where q_b denotes the dimensionality of the random-effects vector and $||x|| = [\sum_i x_i^2]^{1/2}$ the Euclidean vector norm.

Maximization of the log-likelihood function $\ell(\boldsymbol{\theta}) = \sum_i \log p(T_i,\delta_i,\boldsymbol{y}_i;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is usually done through the Expectation-Maximization (EM) algorithm.

Particularly, assuming a Weibull distribution for the baseline hazard:

$$h_0(t) = (\nu/\rho)(t/\rho)^{\nu-1}, \tag{2.4.7}$$

where $\rho > 0$ is a scale parameter and $\nu > 0$ a shape parameter, leads to a cumulative baseline hazard rate of $\Lambda_0(t) = (t/\rho)^\nu$ and the conditional density for the survival part is expressed as

$$\begin{aligned}p(T_i,\delta_i|\boldsymbol{b}_i;\boldsymbol{\theta}_t,\boldsymbol{\beta}) =& \Big[h_0(T_i)\exp\{\boldsymbol{\gamma}'\boldsymbol{w}_i + \alpha m_i(T_i)\}\Big]^{\delta_i} \\ & \exp\Big(-\Lambda_0(t)\exp\{\boldsymbol{\gamma}'\boldsymbol{w}_i + \alpha m_i(s)\}ds\Big).\end{aligned} \tag{2.4.8}$$

2.4.2 The Likelihood Ratio Test

Since the joint model is fit by maximizing the joint likelihood of the longitudinal and survival data, the likelihood ratio test can be used to test hypothesis regarding its parameters (RIZOPOULOS, 2012).

The test statistic is defined as

$$LRT = -2\{\ell(\hat{\boldsymbol{\theta}}_0 - \ell(\hat{\boldsymbol{\theta}})\} \tag{2.4.9}$$

where $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}$ represent the parameter estimates under the null and under the alternative hypothesis, respectively. This test is appropriate for the comparison of nested models, where a significant p-value indicates that the model under the alternative hypothesis provides a better fit to the data than the model under the null hypothesis.

2.4.3 Diagnostics

Evaluation of model fit and validation of assumptions can be done for joint models using different types of residuals, as well as measures of model fitness, as described in Rizopoulos (2012). In this section we present these residuals and their usability in joint model diagnostics, followed by the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) definitions.

2.4.3.1 Standardized Conditional Residuals

Conditional residuals, also referred to as subject-specific residuals, can be used to evaluate the hierarchical version of the longitudinal submodel. These residuals predict the conditional errors $\boldsymbol{\epsilon}_i \sim N(0, \sigma^2)$ in the presence of the random effects for each subject. Their standardized version is defined as

$$\boldsymbol{r}_i^{yss}(t) = \boldsymbol{y}_i(t) - \boldsymbol{x}_i'(t)\hat{\boldsymbol{\beta}} - \boldsymbol{z}_i'(t)\hat{\boldsymbol{b}}_i/\hat{\sigma}, \tag{2.4.10}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ are the maximum likelihood estimates, and $\hat{\boldsymbol{b}_i}$ the empirical Bayes estimates for the random effects.

These residuals can be used to verify the assumptions of homoscedasticity and normality.

2.4.3.2 Standardized Marginal Residuals

Marginal residuals come from the marginal longitudinal model, where the random effects are omitted from the linear predictor, representing the marginal errors $\boldsymbol{y}_i - X_i\boldsymbol{\beta} = Z_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$.

The standardized version of the marginal residuals is defined as

$$\boldsymbol{r}_i^{ysm} = \hat{V}_i^{-1/2}(\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}), \tag{2.4.11}$$

where $\hat{V}_i = Z_i \hat{D} Z_i' + \sigma^2 I_{n_i}$ represents the estimated marginal covariance matrix of $\boldsymbol{y}_i$. These residuals are useful when verifying the specification of the mean structure of the longitudinal submodel, as well as normality and heteroscedasticity.

2.4.3.3 Martingale Residuals

To evaluate the survival submodel, martingale residuals can be calculated. These residuals represent the difference between the observed and the expected number of events for the ith subject at each time point based on the fitted model, and can be defined as

$$\boldsymbol{r}_i^{tm}(t) = N_i(t) - \int_0^t R_i(s)\hat{h}_0(s)\exp\{\hat{\gamma}'\boldsymbol{w}_i + \hat{\alpha}\hat{m}_i(s)\}ds, \tag{2.4.12}$$

for joint models fit with the current-value parameterization, where $N_i(t)$ is the counting process denoting the number of events for subject i at time t and $R_i(t)$ is the risk indicator that assumes value 1 if the subject is at risk at time t, and zero otherwise.

These residuals are useful when evaluating if the appropriate functional form has been used to add the longitudinal process as a covariate in the survival model. Ideally, when plotted against subject-specific fitted values of the longitudinal outcome, these residuals should present a linear trend parallel to the horizontal axis.

2.4.3.4 Cox-Snell Residuals

Commonly used to evaluate the fit of survival models, the Cox-Snell residuals represent the estimated cumulative risk function at each observed event time T_i, and for the current-value parameterization can be defined as

$$\boldsymbol{r}_i^{tcs} = \int_0^{T_i} \hat{h}_0(s)\exp\{\hat{\boldsymbol{\gamma}}'\boldsymbol{w}_i + \hat{\alpha}\hat{m}_i(s)\}ds. \tag{2.4.13}$$

When the model fits the data well, it is expected that the Cox-Snell residuals will have a unit exponential distribution. However, as the residuals are evaluated at the observed event times T_i and these are censored, the residuals are censored as well, meaning they should represent a censored sample from a unit exponential distribution. The fit can be assessed by comparing the survival function of the unit exponential distribution with the Kaplan-Meier estimate of the survival function of $\boldsymbol{r}_i^{tcs}$.

2.4.3.5 Measures of Model Fit

The AIC and BIC are two goodness-of-fit statistics that summarize a model's ability to describe a set of data, commonly used for model selection when adding covariables or testing different distributions to the data (VRIEZE, 2012). The two methods are based on the model's

likelihood function and therefore can only be used to compare models that are fit to the same data and assuming probability distributions that may be different, but nested, meaning they are particular cases of another, more general, distribution.

The AIC is defined as

$$AIC = 2\ell(\hat{\boldsymbol{\theta}}) + 2p, \tag{2.4.14}$$

where p is the number of estimated parameters (the number of elements in $\hat{\boldsymbol{\theta}}$). Similarly, the BIC is defined as

$$BIC = 2\ell(\hat{\boldsymbol{\theta}}) + \log(n)p, \tag{2.4.15}$$

n the number of observations contributing to the sum in the likelihood equation.

By reccomending the model with the lowest value for the criteria, both AIC and BIC seek to penalize the model's likelihood by adding a function of the number of estimated parameters, in an effort to benefit a model that is parsimonious and avoid overfitting. An important difference in the use of each criterion is the properties they have. The BIC is known to be consistent, meaning that as the sample grows large, BIC will select the correct model with probability that approaches 1, assuming that the true model is under consideration and has a constant and finite number of parameters. That property is not true for the AIC, which is known to be an efficient statistic instead, meaning it minimizes a loss function, the mean square error of prediction, and therefore may be a better option when the true model we are seeking is one of infinite parameters, when the number of parameters increases with the sample size or when the true model is not a candidate for selection (VRIEZE, 2012).

2.5 Parameterizations of the Joint Model

In this section, we introduce four different parameterizations of the joint model for longitudinal and time-to-event data, as described in Rizopoulos (2012), which allow for different aspects of the longitudinal trajectory to be introduced as covariates in the survival model, leading to more personalized estimates and different biological interpretations. These parameterizations are later utilized in the analysis of two different datasets.

2.5.1 Current value

We have previously presented a standard joint model where the exponential of the association parameter α denotes the change in the risk for an event at a certain time t that is related to a unit of increase in the value of the longitudinal outcome at that same time point. This is often referred to as the "current value" parameterization, since the connection between the two outcomes is based on the value of the longitudinal marker at a certain time point. This

standard joint model has been widely utilized, in scenarious such as the prediction of prostate cancer diagnosis using longitudinal PSA values (Plè et al., 2015) and the use of biomarkers such as blood urea nitrogen and creatinine to predict kidney transplant graft failure (ALIMI et al., 2020). Other parameterizations have been described by Rizopoulos (2012) and Cekic et al. (2021), among others.

2.5.2 Time-dependent slopes

One other useful way to specify a joint model is through the "time-dependent slopes" parameterization, where the association parameters $\boldsymbol{\alpha}$ are associated not only with $m_i(t)$, but also its derivative $m_i'(t)$, and the survival submodel is defined as

$$h_i(t) = h_0(t) \exp\{\boldsymbol{\gamma}'\boldsymbol{w}_i + \alpha_1 m_i(t) + \alpha_2 m_i'(t)\}. \tag{2.5.1}$$

In this parameterization, the interpretation of parameter α_1 is the same as α in the standard joint model, and α_2 can be interpreted as a measure of the association between the slope, or the rate of change, of the longitudinal trajectory at time t and the risk for an event at that same time. This joint model has been utilized in the study of HIV infected patients' disease progression (WU et al., 2011), as well as the risk for preterm birth in women with type 1 diabetes (GUPTA et al., 2020).

2.5.3 Cumulative Effects

Another parameterization of the joint model can be done considering the integral of the longitudinal trajectory up to a time point t. This specification differs from the previous ones by taking into account the entire previous history of the longitudinal outcome, instead of assuming the risk for an event at time t depends only on the value of the longitudinal outcome at that same time point. Referred to as "cumulative effects" parameterization, the survival submodel takes the form

$$h_i(t) = h_0(t) \exp\{\boldsymbol{\gamma}'\boldsymbol{w}_i + \alpha_3 \int_0^t m_i(s)ds\}, \tag{2.5.2}$$

and parameter α_3 now measures the association between the risk for an event at time t and the area under the longitudinal trajectory up to that same time t. Mauff et al. (2017) explored applications of this parameteriztion in the prediction of the survival of patients with primary biliary cirrhosis using longitudinal serum billirubin measurements, and Brown, Ibrahim e DeGruttola (2005) applied this parameterization to the modeling of viral load and time to event data from an AIDS clinical trial.

2.5.4 Lagged effects

Lastly, the "lagged effects" parameterization is one especially useful when a treatment or patient characteristic may present delayed or ongoing effect even after it is no longer present. In such cases, a joint model can be specified such that the risk of an event at time t is dependent on the true value of the longitudinal marker at time $t - c$, where c represents the time lag of interest:

$$h_i(t) = h_0(t) \exp\{\boldsymbol{\gamma}'\boldsymbol{w}_i + \alpha_4 m_i\{\max(t - c, 0)\}. \tag{2.5.3}$$

Each of the $h_i(t)$ parameterizations presented is incorporated into expression 2.4.5 for the definition of the joint likelihood function and estimation process.

Chapter 3

Applications

The joint model methodology has been demonstrated in sections 3.1 and 3.2 with two separate applications to datasets that contain longitudinal biomarkers and time to event outcomes. Each section begins with an introduction regarding the data's clinical importance and includes descriptive analyses, joint modeling in several parameterizations, and subsections for the discussion of the results and conclusion.

All statistical analysis was done using the R software (R Core Team, 2022), and the packages survival (THERNEAU, 2022), nlme (PINHEIRO; BATES; R Core Team, 2022) and JM (RIZOPOULOS, 2010).

3.1 Longitudinal chloride and COVID-19 dataset

In this study, the aforementioned methods are applied to the data of 58 patients admitted to an intensive care unit for the treatment of COVID-19. These patients spent a minimum of two and a maximum of 58 days in the ICU, and the median ICU time was 12.5 days. A total of 21 patients died during the course of their ICU stay.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus responsible for coronavirus disease 2019 (COVID-19), has been a major health concern worldwide since its emergence in December of 2019. As a respiratory disease that can progress to severe stages, affecting multiple organs and systems, it is important to understand the dynamics of laboratory parameters that indicate disease progression and future prognosis (TEZCAN et al., 2020).

Electrolytes are essential for human life and have important roles in maintaining and regulating cellular functions in the human body (SHRIMANKER; BHATTARAI, 2022). Imbalances in serum electrolytes can have great consequences if not promptly dealt with, especially when

associated with severe diseases such as COVID-19. Different imbalances in electrolytes have been associated with severe illness and poor outcomes in COVID-19 patients (SULTANA et al., 2020; TEZCAN et al., 2020), and therefore their monitoring may have important implications in the management and prognosis of critically ill patients. Chloride is an important electrolyte, found predominantly in the extracellular fluid. Chloride levels are regulated by kidney function, and its imbalance can lead to excess water gain conditions such as congestive heart failure (SHRIMANKER; BHATTARAI, 2022). The presence of low chloride levels, known as hypochloraemia, in particular, has been associated with higher frequencies of ICU admission, use of mechanical ventilation and mortality (TEZCAN et al., 2020), as well as with the development of acute kidney injury (KIMURA et al., 2020).

Although electrolyte imbalances reported in cross-sectional studies might be related to underlying patient characteristics, the pathological processes that occur as a consequence of COVID-19 itself can also be related to imbalances during the disease course. Particularly, SARS-Cov-2 acts directly on the renin-aldosterone-angiotensin system, which regulates electrolyte homeostasis (POURFRIDONI et al., 2021). It has been suggested that electrolyte levels may be sucessful indicators of disease progression (ATILA et al., 2021), and that the correction of unbalanced levels may improve patient outcomes (TAN et al., 2020; FLOR et al., 2021).

Abnormal chloride measures at hospital admission have been found to be associated with poor prognosis and overall mortality (TEZCAN et al., 2020) through logistic regression analysis. However, these results only take into account the status of chloride deregulation at the time of presentation, given the cross sectional nature of the studies. By evaluating disease progression over time, longitudinal measures can be used to increase the performance of regression models.

In this study we aim to examine serum chloride concentration alterations of COVID-19 patients being treated in the intensive care unit (ICU) of the Security Forces Hospital in Saudi Arabia, measured daily from admission to discharge or death, and explore the association between these longitudinal measures and patient survival via joint models.

A version of this study has been submitted to the *Revista Brasileira de Terapia Intensiva* for peer review and publication.

3.1.1 Descriptive Analysis

The dataset contains 16 female and 42 male patients, ranging from 27 to 87 years of age. The median age was 57 years. 29 patients presented with a diagnosis of hypertension and 3 with coronary artery disease. 2 patients had been previously diagnosed with heart failure, 17 with hyperlipidemia and 33 with diabetes. Chronic obstructive pulmonary disease was present in one patient, chronic kidney disease in 7, and 4 had a history of stroke. These characteristics

can be observed according to patients outcomes in Table 1.

Table 1 – Baseline sample characteristics according to patient's outcome.

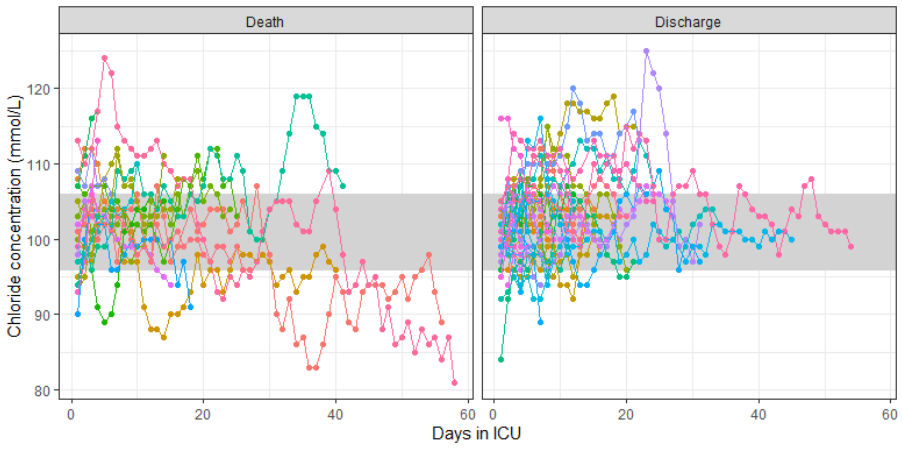
Variable		Died	Survived	p-value
Age, years		65.0 (56.0-72.0)	52.0 (43.0-62.0)	0.015
Sex	Male	17 (40.5%)	25 (59.5%)	0.365
	Female	4 (25.0%)	12 (75.0%)	
Hypertension		12 (41.4%)	17 (58.6%)	0.585
Coronary Artery Disease		2 (66.7%)	1 (33.3%)	0.546
Heart Failure		2 (100.0%)	0 (0.0%)	0.127
Hyperlipidemia		6 (35.3%)	11 (64.7%)	1
Diabetes		12 (36.4%)	21 (63.6%)	1
Chronic Obstructive Pulmonary Disease		1 (100.0%)	0 (0.0%)	0.362
Chronic Kidney Disease		3 (42.9%)	4 (57.1%)	0.695
History of Stroke		1 (25.0%)	3 (75.0%)	1
Days in ICU		14.0 (6.0-24.0)	12.0 (7.0-21.0)	0.852

Values are absolute and relative frequencies for categorical variables, median and interquartile range for numerical variables. p-values refer to Fisher's exact test or Mann-Whitney's U test, accordingly.

3.1.1.1 Longitudinal Measurements of Chloride Concentration

Given the normal range of chloride between 98 mmol/L and 106 mmol/L, fifteen patients were hypochloraemic when admitted to the ICU, while 6 presented with hyperchloraemia. 32 patients measured Chloride below the lower limit of the normal range at some point in their ICU stay, while 34 measured above the upper limit on at least one occasion. The profile charts allow us to visualize each patient's sodium trajectory over the course of their treatment in the ICU, both for deceased and discharged patients (Figure 3.1.1).

Figure 3.1.1 – Profile chart of chloride over time according to patient outcome. Shaded area represents values within the normal range.

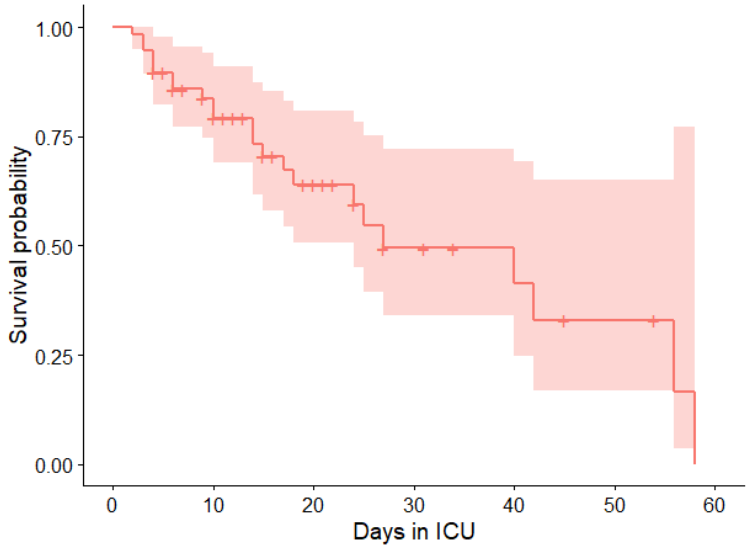


As can be seen in the charts, chloride measurements show reasonable variation over time for each subject. Each patient can present to the ICU with different baseline measurements as well as different trends over time. A slight difference in overall trajectories is suggested by the profile chart, given that patients whose outcome was death appear to have decreasing levels of chloride over time, while discharged patients show stable or increasing measurements throughout their ICU stay. These observations suggest that a linear ME model for these longitudinal trajectories might benefit from a random intercept term as well as a random slope term.

3.1.1.2 Patient Survival

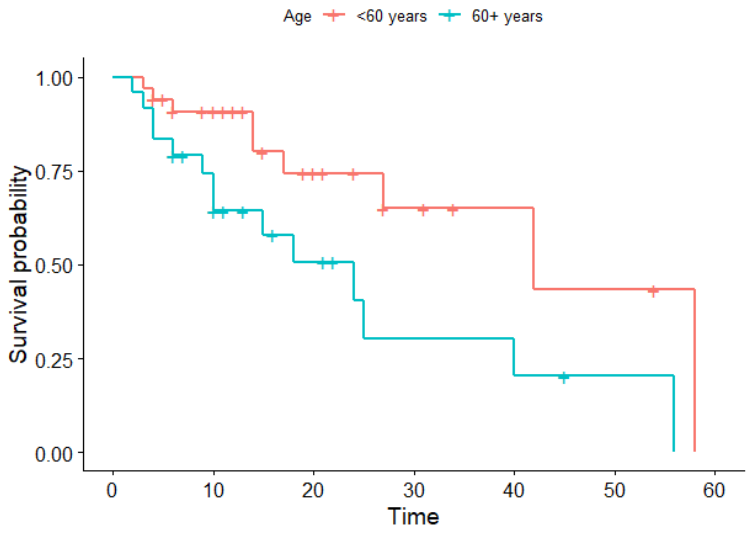
As for the survival outcome, we can use the Kaplan-Meier estimator to estimate a median survival time of 27 days for the overall sample, as well as observe the rate at which the survival probability declines (Figure 3.1.2).

Figure 3.1.2 – Kaplan-Meier curve for overall survival



The Kaplan-Meier estimator was also used to plot different survival curves according to patient's age, and a large gap can be observed between the two curves in Figure 3.1.3, which suggests that age is a determining factor when estimating risk of death in the ICU. Median survival time was estimated at 42 days for those under 60 years of age, and 24 days for those 60 years or older.

Figure 3.1.3 – Kaplan-Meier curve for survival according to age



3.1.2 Longitudinal and Survival Submodels

The longitudinal submodel for Chloride measurements is a linear mixed effects model with random intercept and random slopes. Patients' scaled age was used as a covariable, such that one unit of $AgeScaled$ corresponds to one standard deviation of 14.4 years from the mean of 56.7 years, and time in the icu was also reparameterized so that one unit of $Time$ corresponds to 3 days in the ICU.

The model expression is defined,

$$Chloride_{ij} = \hat{\beta}_0 + \hat{b}_{0i} + (\hat{\beta}_1 + \hat{b}_{1i}) * Time_{ij} + \hat{\beta}_2 * AgeScaled_i + e_{ij}, \tag{3.1.1}$$

and the REML estimates, as well as corresponding tests of significance and p-values are present in Table 2. Standard deviations for the intercept and slope random effects were estimated as 4.895 and 1.179, respectively, and the correlation between them was -0.627.

Table 2 – REML estimates for the longitudinal submodel

Fixed Effects	Coefficient	Std. Error	DF	t-value	p-value
Intercept	102.245	0.705	897	144.964	0.0000
AgeScaled	-0.114	0.538	56	-0.212	0.833
Time	0.198	0.203	897	0.976	0.329

The survival submodel is a proportional hazards regression model, where patients who did not die during their ICU stay are considered censored. Individuals' age was also included in the model as a covariable, scaled in the same way as mentioned for the ME model. The model with unspecified baseline hazard function was estimated via maximum likelihood, resulting

in a coefficient of 0.519 (standard error: 0.227, z-value: 2.284, p-value: 0.022) for age. The exponentiated coefficient resulted in 1.68, meaning the hazard of death in patients 14.4 years older was 68% higher than in patients with the mean age of 56.7 years.

These two submodels were used to estimate the joint models described in the next section.

3.1.3 Joint Models

For the joint model specification, to avoid underestimation of the standard errors, the baseline hazard function of the survival part was assumed to follow a Weibull distribution. Notably, changes in parameter estimates can occur due to the nature of the missing mechanism being assumed for the data in the longitudinal case, and due to adjustment to the time varying covariate in the survival case.

3.1.3.1 Current Value Parameterization

Under the "current value" parameterization, we can observe changes in the estimated coefficients for both parts of the model (Table 3), in spite of a non-significant association parameter $\alpha = -0.010$. In this model, patients' older age was associated with lower Chloride measurements, and time progression was generally associated with increase in Chloride, on average. Age was also associated with increased risk of death during ICU stay. The $log(shape)$ and $Scale$ estimates refer to the baseline Weibull risk function parameters.

Table 3 – Current value joint model estimates

Variance Components	Std. Deviation	Corr
Intercept	5.258	
Time	1.167	-0.642
Residual	4.001	

	Coefficient	Std. Error	p-value
Longitudinal Process			
Intercept	100.629	0.605	0.0000
AgeScaled	-0.764	0.344	0.026
Time	0.387	0.085	0.0000
Event Process			
Intercept	-3.759	1.209	0.002
AgeScaled	0.497	0.225	0.027
α	-0.010	0.011	0.352
log(shape)	0.268	0.164	0.102
Scale: 1.307			

3.1.3.2 Time-Dependent Slope Parameterization

Adding an association parameter that corresponds to the effect of the slope of the longitudinal trajectory in the survival process provides a significant relationship between the two processes. Although other estimates remain relatively stable, the results from this parameterization imply that negative slopes are associated with increased risk of death during ICU, while positive slopes are associated with lower risk of death (Table 4). These results are in agreement with the previous graphical analysis that showed decreasing levels of Chloride in deceased patients in the course of their ICU stay, and also characterizes the dropout process generated by patients' deaths as MNAR, contrary to what would have been inferred by the previous parameterization.

Table 4 – Time-dependent slope joint model estimates

Variance Components	Std. Deviation	Corr
Intercept	5.166	
Time	1.064	-0.623
Residual	4.015	

	Coefficient	Std. Error	p-value
Longitudinal Process			
Intercept	100.612	0.533	0.0000
AgeScaled	-0.732	0.299	0.015
Time	0.384	0.075	0.0000
Event Process			
Intercept	-6.934	1.907	0.003
AgeScaled	0.542	0.244	0.026
α_1	-0.016	0.014	0.259
α_2	-1.024	0.486	0.035
log(shape)	0.372	0.182	0.040
Scale: 1.45			

3.1.3.3 Cumulative Effects Parameterization

While in the previous model we included the longitudinal trajectory's derivative with respect to time in the linear predictor of the event process, in this parameterization we include the integral of that same trajectory with respect to time instead, this way taking the entire previous history of the longitudinal biomarker into account when estimating the risk of death at each time point. As presented in Table 5, The association parameter corresponding to the area under the previous longitudinal trajectory was not significant in estimating risk of death.

Table 5 – Cumulative effects joint model estimates

Variance Components	Std. Deviation	Corr
Intercept	5.227	
Time	1.154	-0.635
Residual	4.003	

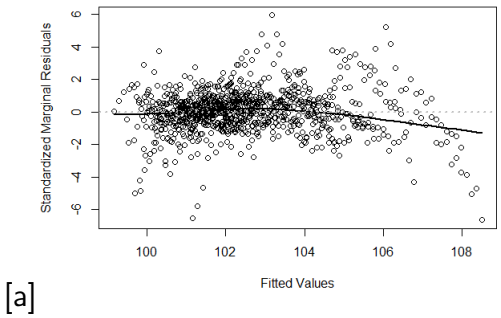
	Coefficient	Std. Error	p-value
Longitudinal Process			
Intercept	100.525	0.891	0.0000
AgeScaled	-0.746	0.478	0.118
Time	0.403	0.127	0.001
Event Process			
Intercept	-4.498	0.794	0.0000
AgeScaled	0.475	0.219	0.031
α_3	0.0001	0.0002	0.693
log(shape)	0.144	0.265	0.587
Scale: 1.155			

3.1.4 Analysis of Residuals and Model Diagnostics

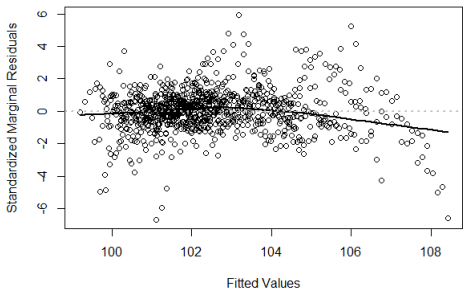
To assess the fit of the longitudinal part of the joint model, standardized marginal(Figure 3.1.4) and conditional (Figure 3.1.5) residuals were plotted against fitted values. The first plot represents individuals deviation from the fixed part of the model, while the second represents their deviation from their individual predictions, taking subject-specific random effects into account. These residuals should appear randomly distributed around the mean of zero, which is more clear in the conditional residuals plots than in the marginal residuals plots. When not accounting for random effects, higher chloride values seem to be somewhat underestimated by the fixed part of the model, but this lack of fit is resolved when the random effects are taken into consideration.

When it comes to the comparison between the three fitted models, they all seem to be considerably similar, which is expected, since the difference between them is the linear predictor of the survival part. Subject-specific residuals were also plotted against the theoretical quantiles of a normal distribution (Figure 3.1.6), where they should appear to closely follow the diagonal line, which is true for all three parameterizations.

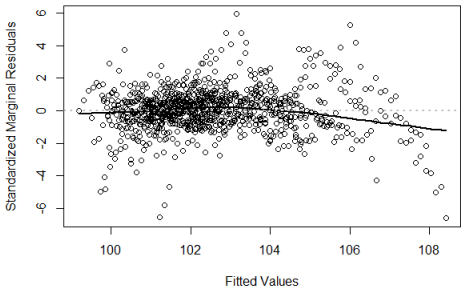
Figure 3.1.4 – Standardized marginal residuals and fitted values for (a) current value (b) time-dependent slope and (c) cumulative effects models. Solid black lines represent loess curves.



[a]



[b]



[c]

Figure 3.1.5 – Standardized subject-specific residuals and fitted values for (a) current value (b) time-dependent slope and (c) cumulative effects models. Solid black lines represent loess curves.

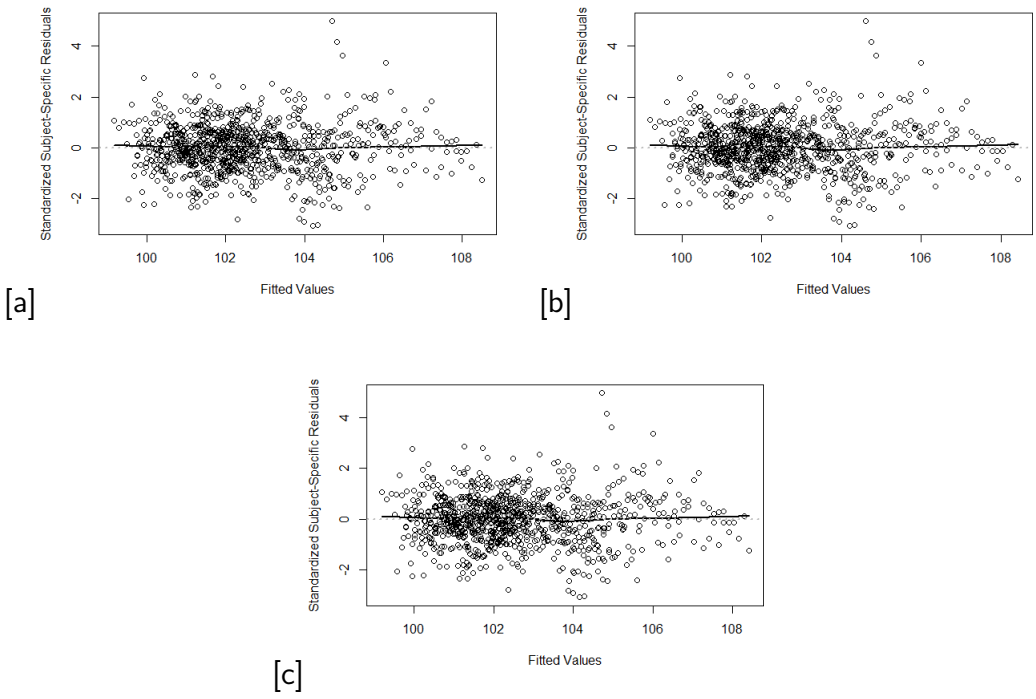
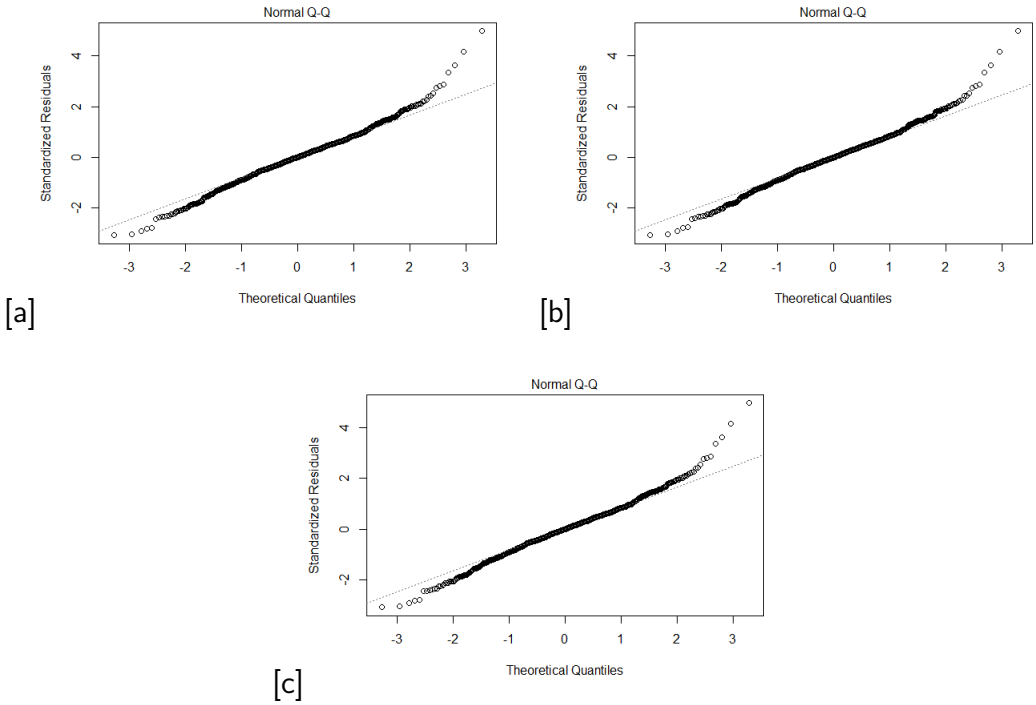
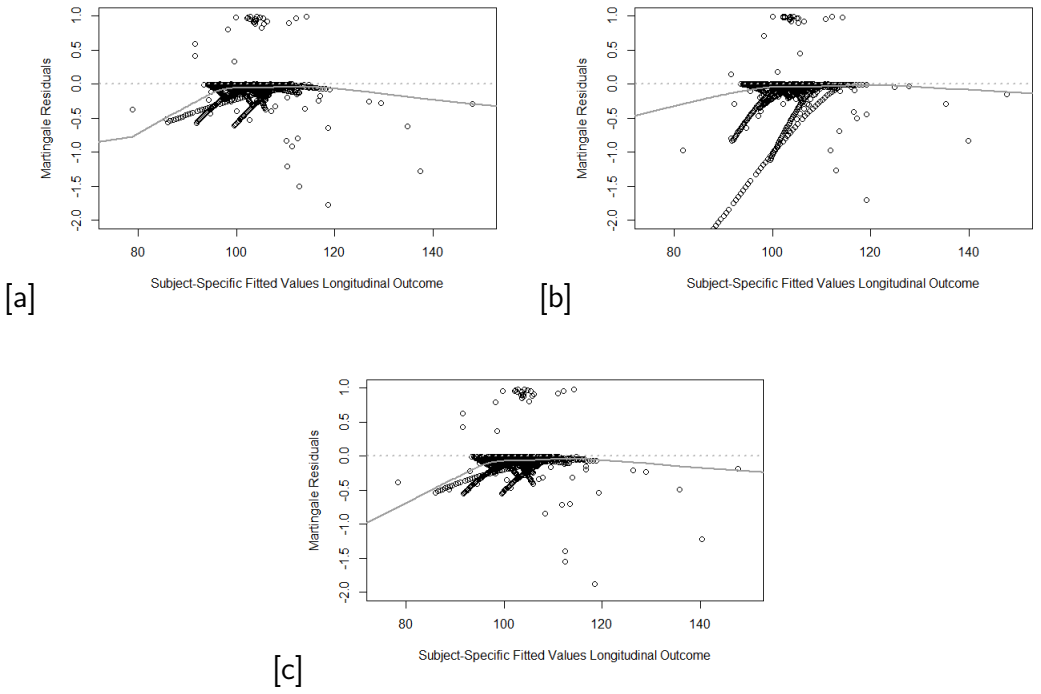


Figure 3.1.6 – Normal Q-Q plots for (a) current value (b) time-dependent slope and (c) cumulative effects models



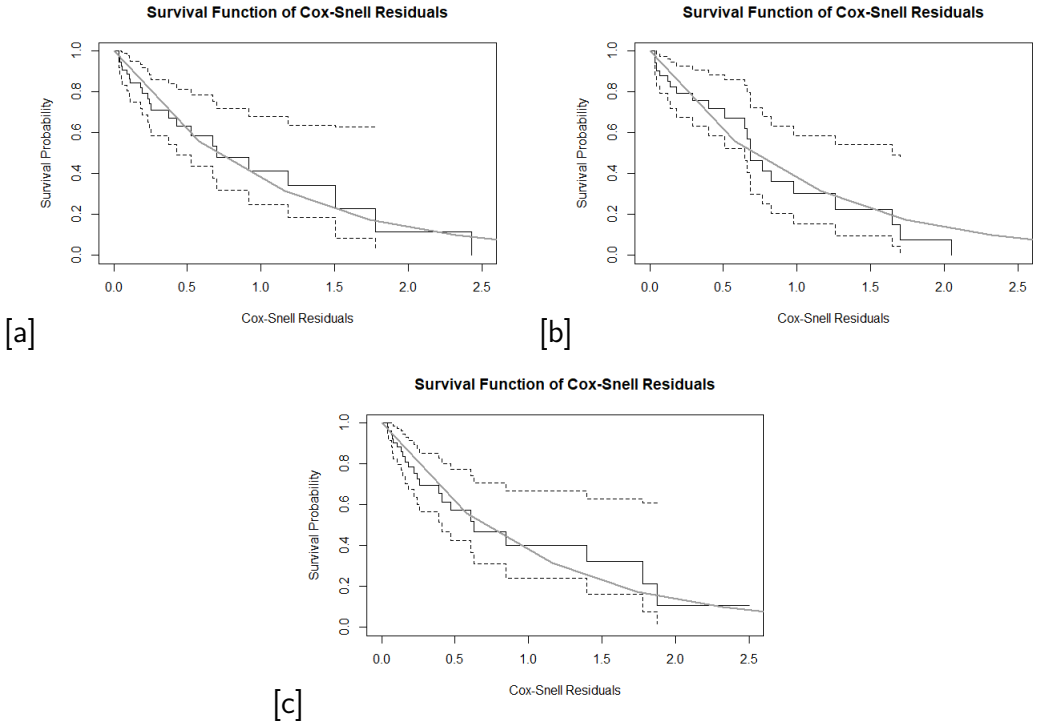
Martingale residuals and Cox-Snell residuals were used to verify model assumptions for the survival part. In the martingale residuals represented in Figure 3.1.7, we expect to see a relatively straight horizontal line in the loess curve, paralell to the horizontal axis, indicating that the relationship between the longitudinal measures and the survival process had been correctly specified. However, an apparent lack of fit for smaller values of chloride can be explained by the imbalance in observations caused by the missing process itself, as explained in detail by Rizopoulos (2012). It is reasonable to assume that the smoothed curve would appear much straighter if the longitudinal trajectories had not been truncated by death or censoring. When it comes to comparing the three model specifications, the time-dependent slope parameterization seems to be the one that leads to the straighter smoothing curve, and therefore the relationship between the two processes might be better specified.

Figure 3.1.7 – Martingale residuals and fitted values for (a) current value (b) time-dependent slope and (c) cumulative effects models. Solid grey lines represent loess curves.



Cox-Snell residuals, represented in Figure 3.1.8, should have a survival function that closely resembles a unit exponential, when the model is correctly specified. The residuals kaplan-meier estimator and 95% confidence interval are represented so that the unit exponential must be contained in the interval, therefore not exhibiting evidence that this assumption is violated. All three model specifications show a reasonable resemblance between the estimated survival curve and the unit exponential curve.

Figure 3.1.8 – Kaplan-Meier estimator of Cox-Snell residuals for (a) current value (b) time-dependent slope and (c) cumulative effects models. Dashed lines represent the estimator's 95% confidence interval, solid grey lines represent the unit exponential.



Given that all three model specifications seem to fit the data reasonably well, the Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC) and log-likelihood values for each model are presented in Table 6. Ideally, the model that best fits the data while remaining parsimonious would have the highest log-likelihood value, as well as the lowest AIC and BIC values. The model with the time-dependent slope parameterization has all of those characteristics, and its improvement in fit when compared to the current value model is further confirmed by the likelihood ratio test (Table 7), which can be employed in this case given that the two models are nested. Therefore, the time-dependent slope model was chosen for further interpretation of its practical value.

Table 6 – Measures of model fitness

Parameterization	Log-likelihood	AIC	BIC
Current Value	-2896.728	5715.457	5838.122
Time-Dependent Slope	-2892.396	5808.792	5833.517
Cumulative Effects	-2895.863	5815.725	5840.451

Table 7 – Likelihood ratio test results

First model	Second model	Test statistic	p-value
Current Value	Time-Dependent Slope	8.66	0.003

3.1.5 Discussion

This study's results have found that, as an essential electrolyte, chloride has an important role in describing the progression of COVID-19 disease in severely ill patients being treated in the ICU. Chloride imbalances have previously been associated with mortality in critically ill patients (JI; LI, 2021; MARTTINEN et al., 2016), and in patients with severe acute conditions (GRODIN et al., 2015; MAATEN et al., 2016). A study of hospitalized patients suffering from acute heart failure found that newly developed or persistent hypochloraemia was associated with increased mortality, while baseline hypochloraemia that resolved within 14 days was not (MAATEN et al., 2016). These findings are in accordance with the results of the present study, where the rate of decrease in chloride concentration was significantly associated with lower survival time, while increase in chloride concentration was associated with increased survival. These results also highlight the importance of longitudinal studies that take into account the dynamics of these biomarkers over time in hospitalized patients.

Hypochloraemia in ICU patients may be related to gastrointestinal or renal losses of chloride ions, which can occur in the presence of renal disorders, gastrointestinal symptoms such as vomiting, and congestive heart failure (BANDAK; KASHANI, 2017). Acute renal involvement is not unexpected in COVID-19 patients, and is correlated with poor outcomes and higher mortality in these patients (POURFRIDONI et al., 2021). Gastrointestinal symptoms such as abdominal pain and vomiting have also been widely reported in relation to this disease (HENRY et al., 2020). COVID-19 disease has also been reported to increase the odds of development of acute heart failure in both previously healthy patients (BADER et al., 2021) and patients with previous history of heart failure (REY et al., 2020). Combined with the present study, previous investigations suggest that the effects of COVID-19 infection on kidney and heart function may translate into lowering chloride levels, making it an important marker of disease progression and poor prognosis.

3.1.6 Conclusion

Although limited by its small sample and observational nature, this longitudinal study can provide valuable insight into the dynamics of COVID-19 infection in severely ill patients. Our results bring attention to the increase in information that can be found when collecting data from patients at many time points, instead of limiting data collection to the moment of hospital admission, especially considering traditional regression models, other than the joint

model presented in this work, would not be able to evaluate the significance of the longitudinal trajectory's slope on the survival time. This information can lead to a better understanding of the disease and its consequences. Moreover, monitoring patients over time, when possible, can be extremely useful to identify changes in their prognosis, which, in collaboration with well trained dynamic algorithms, can even be done automatically, relieving some of the burden of healthcare professionals by aiding them in informed decision making, consequently reducing costs and contributing to patient-focused quality healthcare services.

3.2 HIV coinfection dataset

In this study, the joint model methodology is used to explore the relationship between the quantity of a specific type of white blood cell, CD4 lymphocytes, and the time of treatment necessary to reach a healthy CD4/CD8 cell count ratio in HIV patients coinfected with the hepatitis B and hepatitis C viruses (HBV and HCV). While absolute CD4 cell count has been an established predictor of HIV disease progression, evidence suggests that the ratio between CD4 and CD8 cells might be an even better marker of immune dysfunction in HIV patients treated using antiretroviral therapy (ART). While long-term use of ART is known to increase CD4 cell count to a normal range on up to 80% of HIV patients, in many cases CD8 cells remain elevated through years of treatment, resulting in low CD4/CD8 ratios (LU et al., 2015; HELLEBERG et al., 2015). A lower CD4/CD8 cell ratio in HIV patients has been associated with higher non-AIDS related mortality even when CD4 cell count is at a normal range (LU et al., 2015).

HBV and HCV share common routes of transmission with HIV, leading to a prevalence of up to 30% of coinfection among HIV patients (BONACINI et al., 2004). Individuals with these conditions in addition to HIV are at a larger risk of hospitalization and liver-related mortality (BONACINI et al., 2004; ANDREONI et al., 2012). HBV and HCV have also been reported to prevent immunological recovery in HIV patients, although the impact of these diseases on CD4 cell count and CD4/CD8 ratio are still a subject of ongoing investigation (SILVA et al., 2018).

3.2.1 Descriptive analysis

This data originates from a retrospective cohort study of individuals diagnosed with HIV and coinfected with HBV and HCV between 2002 and 2016 in the cities of Cascavel and Maringá, in the south region of Brazil. The sample consisted of 147 individuals with a minimum of 3 and a maximum of 16 measures of CD4 and CD8 cells available for a follow up of up to 4 years. The event of interest was defined as the patient reaching a ratio of CD4/CD8 cells above or

equal to 0.9, indicating a healthy immune system in HIV infected patients.

108 individuals belonged to the control group, meaning they were infected exclusively by HIV, and 34 of them experienced the event. 21 individuals were infected by HIV and HBV simultaneously, and 3 of those experienced the event, and 18 were infected by HIV and HCV simultaneously, of which 8 experienced the event during follow up.

As previously described by (BRUM; PREVIDELLI, 2018), CD4 cell count presents a distribution that is asymmetric, but the square root transformation corrects this behaviour and any heterokedasticity that may occur as a consequence of larger values being accompanied by larger variability. Each group's mean trajectory suggests that HCV and HBV coinfections are associated with differences in CD4 cell count when compared to the control group and with each other (Figure 3.2.1). Figure 3.2.2 also provides evidence that between-subject variability is present at baseline, which could be accommodated by a random intercept term, and that slopes vary between subjects during follow-up, suggesting the model for this data can benefit from a random slope term.

Figure 3.2.1 – Mean $\sqrt{CD4}$ trajectories by group

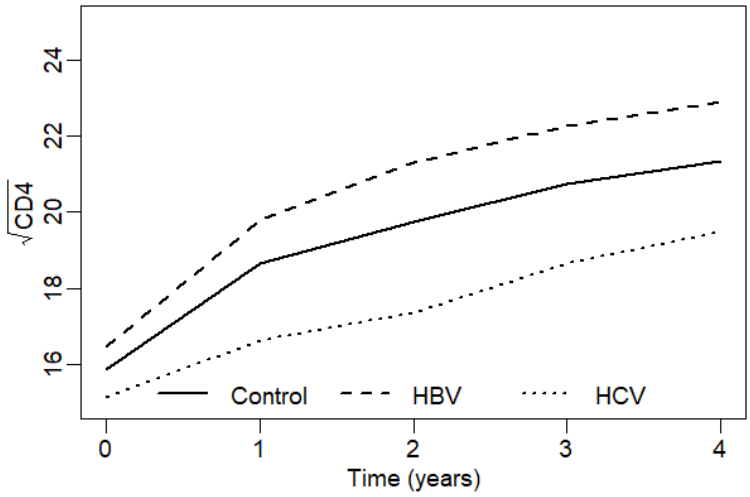
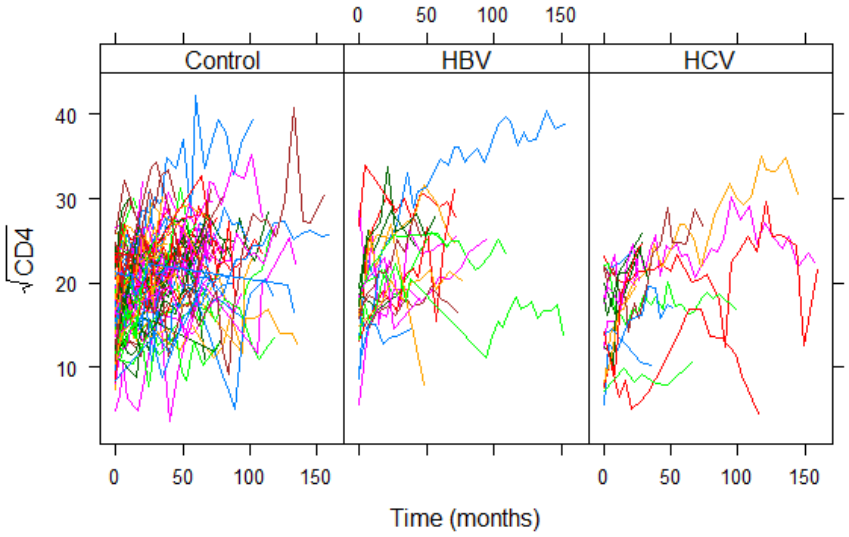


Figure 3.2.2 – Profile charts of $\sqrt{CD4}$ over time



3.2.2 Longitudinal and survival submodels

The longitudinal submodel for $\sqrt{CD4}$ cell count over time is a mixed-effects linear model where the patient's group is a covariate that interacts with time, the number of months of follow-up. Two random effects are specified, one for the intercept and one for the slope. This model's coefficients are presented on Table 8 and random effects' standard deviations were estimated at 3.906 for the intercept and 0.126 for the slope. Time of follow-up was the only statistically significant variable at the 5% level of significance, indicating there was an average increase in $\sqrt{CD4}$ values over time.

Table 8 – REML estimates for the longitudinal submodel of $\sqrt{CD4}$

Fixed Effects	Coefficient	Std. Error	DF	t-value	p-value
Intercept	17.107	0.409	897	41.855	<0.001
HBV	1.336	1.011	144	1.322	0.188
HCV	-1.989	1.080	144	-1.842	0.068
Time (months)	0.138	0.015	897	9.105	<0.001
HBV*Time (months)	-0.014	0.036	897	-0.390	0.696
HCV*Time (months)	0.008	0.040	897	0.191	0.849

For the survival submodel, patients who did not reach a CD4/CD8 cell count ratio of 0.9 had their observations censored. The coinfection group was used as a covariable in this model but did not show statistical significance (HR = 0.373, p = 0.102 for HBV and HR = 1.302, p = 0.510 for HCV).

These submodels were used in the subsequent estimation of the joint models presented in the next section.

3.2.3 Joint models

Four different parameterizations of the joint model were fit to this data. Each of them has in common the two submodels utilized and the baseline hazard function, which was assumed to follow a Weibull distribution.

Under the current value parameterization (Table 9), we found a significant association parameter $\alpha_1 = 0.048$, corresponding to a Hazard Ratio (HR) of 1.049, which can be interpreted as a risk of event 4.9% higher at a certain time point for each unit of increase in the square root of CD4 cell count at that time. Taking this association into account also factored into the estimation of the longitudinal process coefficients, and the HCV coinfected group is now significantly different from the control group regarding mean square root CD4 cell counts. Along with the longitudinal marker, the HBV coinfection was also a significant factor in determining the risk of acquiring a healthy CD4/CD8 ratio, and the HR of 0.258 indicates patients infected with HBV were around 25% as likely as controls to achieve a healthy CD4/CD8 ratio.

Table 9 – Current value joint model for $\sqrt{CD4}$ and time to event

Variance Components	Std. Deviation	Corr
Intercept	3.860	
Time	0.126	-0.232
Residual	2.489	

	Coefficient	Std. Error	p-vale
Longitudinal Process			
Intercept	17.187	0.394	<0.001
HBV	1.343	0.695	0.053
HCV	-2.906	0.725	<0.001
Time (Months)	0.139	0.012	<0.001
HBV * Time (Months)	-0.016	0.027	0.555
HCV * Time (Months)	0.021	0.028	0.464
Intercept	-6.799	0.684	<0.001
HBV	-1.355	0.636	0.033
HCV	0.219	0.401	0.585
α_1	0.048	0.015	0.002
log(shape)	0.081	0.156	0.605
Scale: 1.084			

Similarly, when using the time-lagged parameterization (Table 10), with a time lag of 6

months, both time and HCV coinfection had significant effects on the square root of CD4 cell counts, and the cell count and HBV coinfection were significantly associated with the risk of event.

Table 10 – Six months lagged joint model $\sqrt{CD4}$ and time to event

Variance Components	Std. Deviation	Corr
Intercept	3.863	
Time	0.126	-0.234
Residual	2.490	

	Coefficient	Std. Error	p-vale
Longitudinal Process			
Intercept	17.189	0.394	<0.001
HBV	1.336	0.694	0.054
HCV	-2.894	0.733	<0.001
Time (Months)	0.139	0.012	<0.001
HBV * Time (Months)	-0.016	0.027	0.565
HCV * Time (Months)	0.021	0.028	0.468
Intercept	-6.754	0.681	<0.001
HBV	-1.313	0.629	0.037
HCV	0.219	0.399	0.583
α_1 (lag = 6)	0.047	0.016	0.003
log(shape)	0.085	0.154	0.579
Scale: 1.089			

The time-dependent slopes parameterization introduces a new parameter α_2, representing the effect of the slope of the longitudinal trajectory of the square root of the CD4 cell count on the time to event outcome. The inclusion of this parameter caused the current value association parameter to loose significance, while a positive slope of the CD4 trajectory was associated with a significant increase in the hazard of the event. Considering the standard deviation of the random effect associated with the slope was estimated at 0.131, an increase of one standard deviation in the slope was associated with over 3.25 times the hazard of reaching a healthy CD4/CD8 ratio at a certain time point. HBV coinfection remained associated with a decreased risk of the event.

Table 11 – Time-dependent slopes joint model for $\sqrt{CD4}$ and time to event

Variance Components	Std. Deviation	Corr
Intercept	3.742	
Time	0.131	-0.167
Residual	2.493	

	Coefficient	Std. Error	p-vale
Longitudinal Process			
Intercept	17.090	0.398	<0.001
HBV	1.467	0.707	0.038
HCV	-2.945	0.664	<0.001
Time (Months)	0.144	0.012	<0.001
HBV * Time (Months)	-0.021	0.027	0.437
HCV * Time (Months)	0.021	0.031	0.501
Intercept	-9.447	1.394	<0.001
HBV	-1.605	0.696	0.021
HCV	0.0001	0.451	0.999
α_1	-0.015	0.027	0.572
α_2	9.017	2.525	<0.001
log(shape)	0.528	0.199	0.008
Scale: 1.695			

Finally, the cumulative effects parameterization was used (Table 12), in an effort to take into account each patients' previous history of the longitudinal biomarker. HCV coinfection and follow-up time remained significantly associated with the CD4 cell count, but there were no associations with the time-to-event outcome in this scenario.

Table 12 – Cumulative effects joint model with cumulative effects for $\sqrt{CD4}$ and time to event

Variance Components	Std. Deviation	Corr
Intercept	3.866	
Time	0.125	-0.239
Residual	2.492	

	Coefficient	Std. Error	p-vale
Longitudinal Process			
Intercept	17.180	0.399	<0.001
HBV	1.331	0.690	0.054
HCV	-2.826	0.786	<0.001
Time (Months)	0.139	0.012	<0.001
HBV * Time (Months)	-0.015	0.027	0.595
HCV * Time (Months)	0.018	0.028	0.523
Intercept	-6.602	0.888	<0.001
HBV	-0.969	0.606	0.110
HCV	0.215	0.395	0.586
α_3	0.001	0.001	0.954
log(shape)	0.252	0.179	0.157
Scale: 1.287			

3.2.4 Analysis of residuals and model selection

In order to evaluate each model's fit to the data, we conducted an analysis of relevant types of residuals. Since the longitudinal part of the model is specified in the same way for each parameterization, we do not expect to see any relevant changes in the residuals. However, there may be an increase or decrease in model adequacy when considering the survival process residuals.

The standardized marginal (Figure 3.2.3) and subject-specific (Figure 3.2.4) residuals behave as expected, with no clear indication of assumption violations. The inclusion of the random effects in the model corrects most of the heterokedasticity that is apparent on the marginal residuals. In addition, the normal quantile plots (Figure 3.2.5) for each model also do not show differences between each other.

Figure 3.2.3 – Standardized marginal residuals for (a) current value, (b) 6-months lagged, (c) time-dependent slope and (d) cumulative effects models.

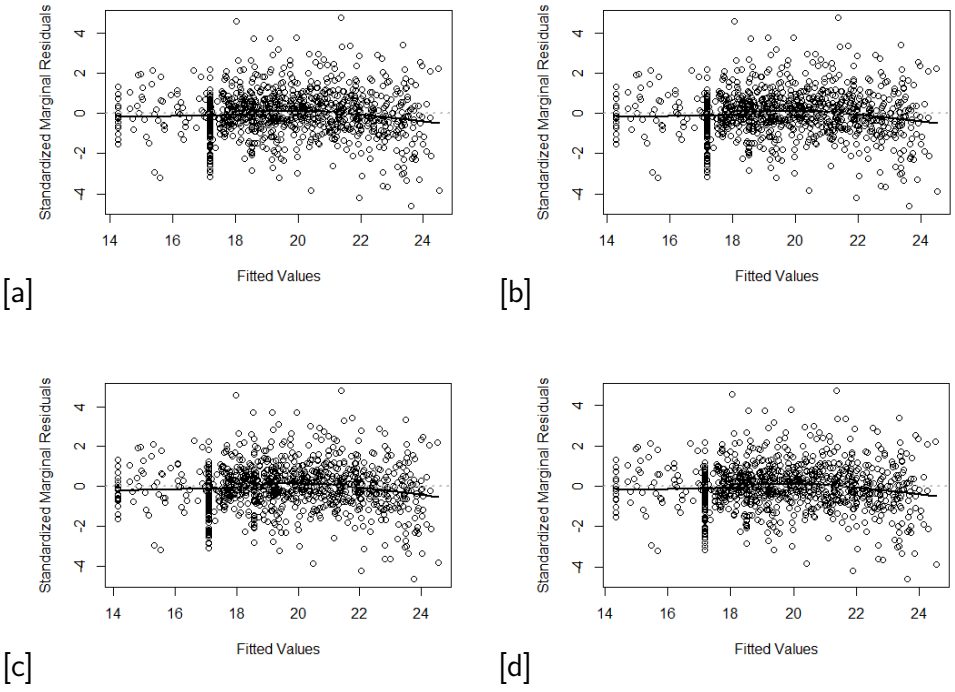


Figure 3.2.4 – Subject-specific residuals for (a) current value, (b) 6-months lagged, (c) time-dependent slope and (d) cumulative effects models.

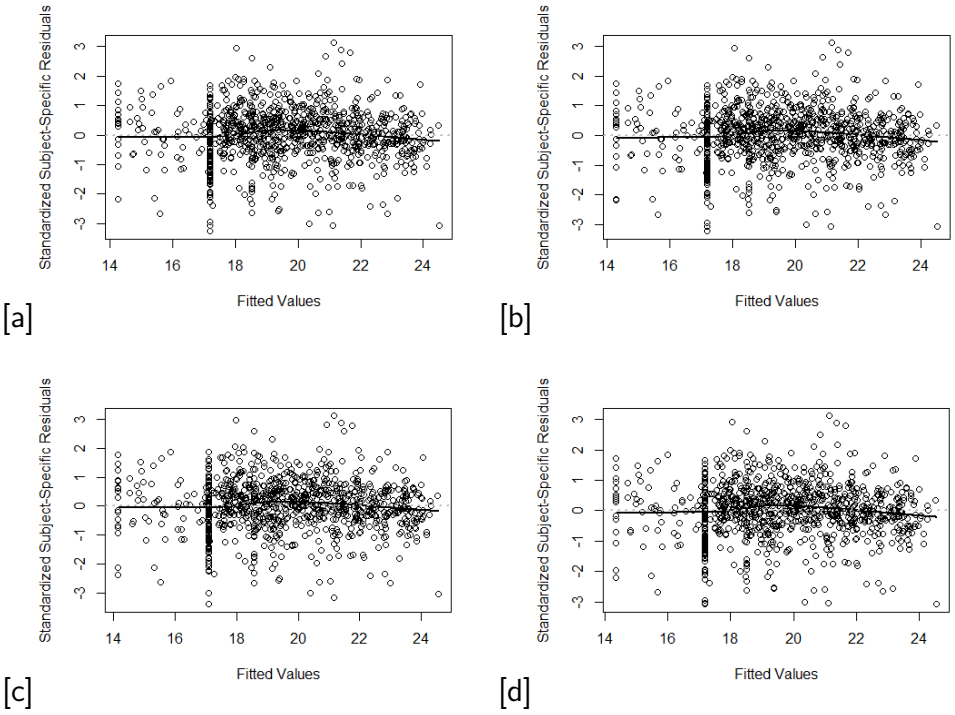
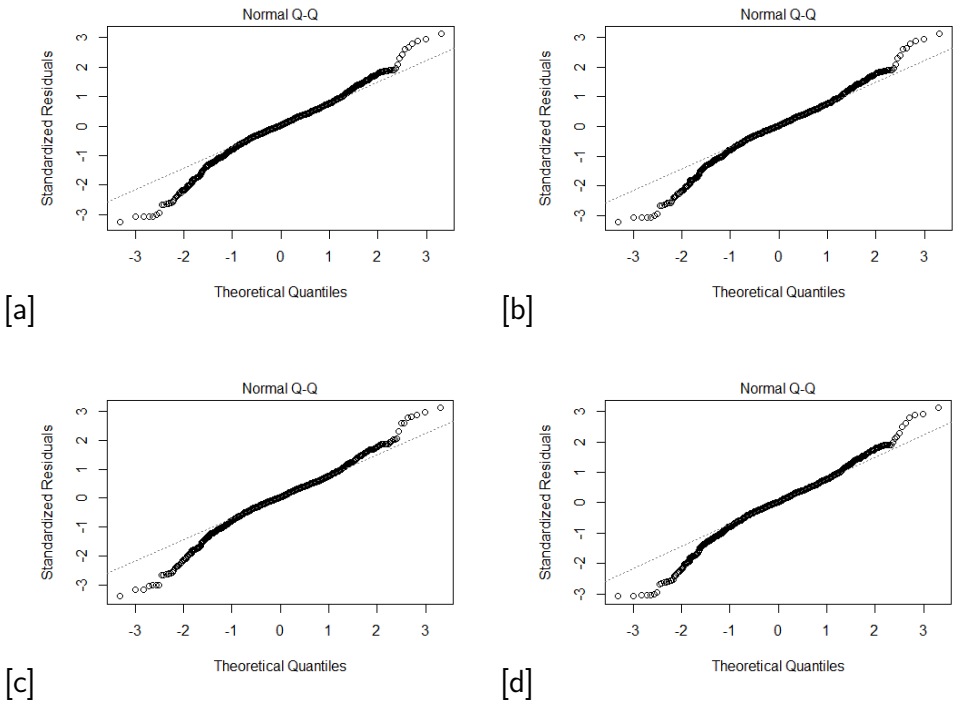


Figure 3.2.5 – Normal Q-Q plots of subject-specific residuals for (a) current value, (b) 6-months lagged, (c) time-dependent slope and (d) cumulative effects models. Solid grey line represents loess curve.



Martingale and Cox-Snell residuals were used to evaluate the fit of the survival part of each model. Martingale residuals (Figure 3.2.6) show reasonable adequacy, considering the loess curves follow a straight line mostly parallel to the horizontal axis. Cox-Snell residuals, however, did not fully correspond to the expected behaviour, a Kaplan-Meier curve that does not significantly differ from an exponential distribution. The model where these residuals resemble the expected curve the least is the time-dependent slopes parameterization, while the current value and 6-months lagged parameterizations seem to have almost identical fit (Figure 3.2.7).

Figure 3.2.6 – Martingale residuals for (a) current value, (b) 6-months lagged, (c) time-dependent slope and (d) cumulative effects models. Solid grey line represents loess curve.

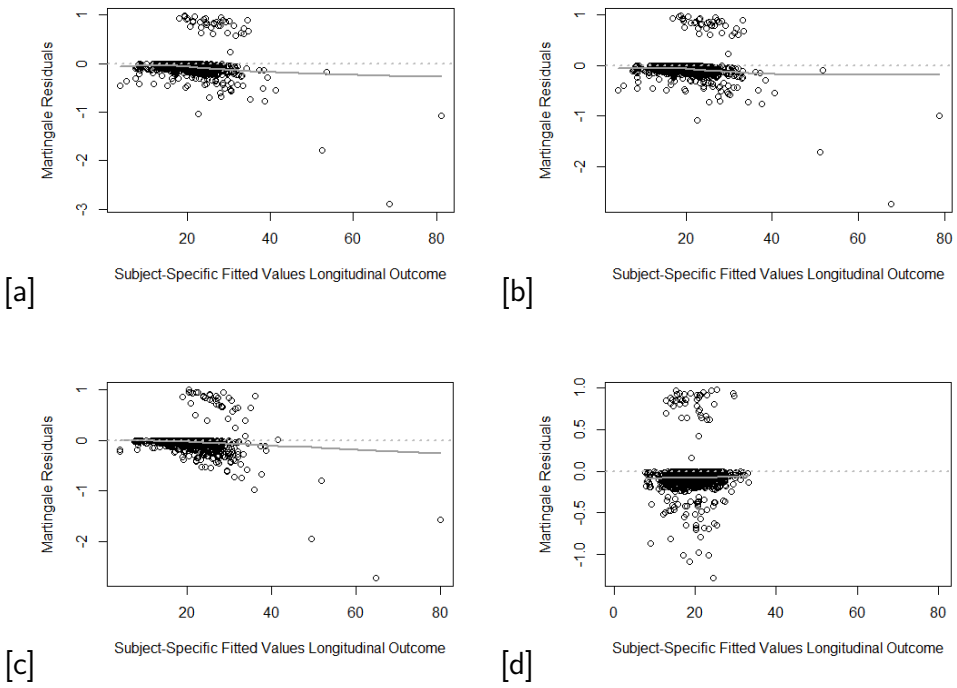
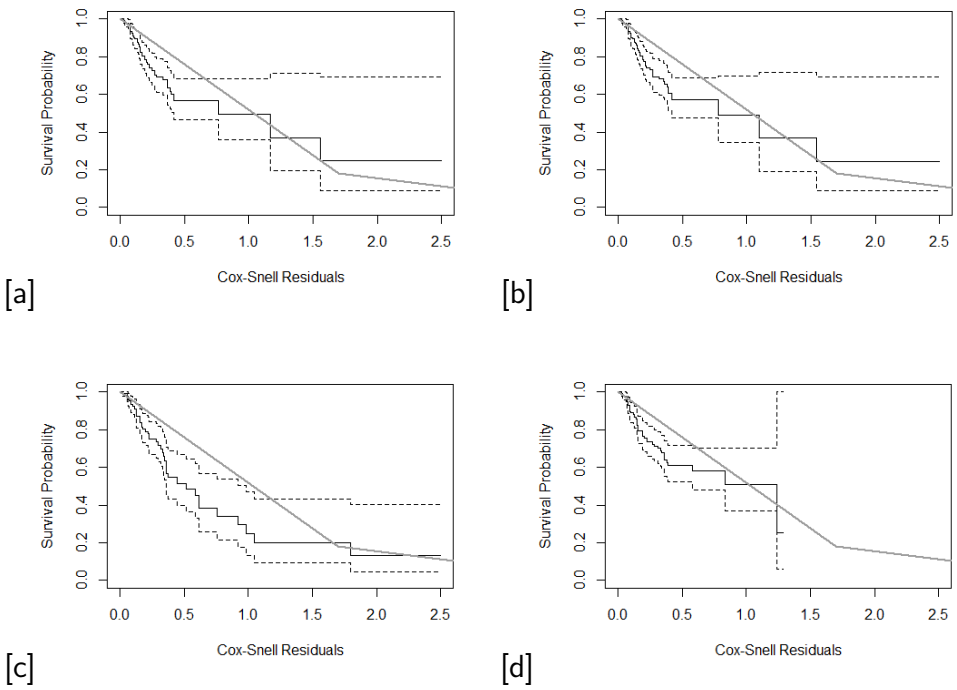


Figure 3.2.7 – Kaplan-Meier estimator of Cox-Snell residuals for (a) current value, (b) 6-months lagged, (c) time-dependent slope and (d) cumulative effects models. Dashed lines represent the estimator's 95% confidence interval, solid grey lines represent the unit exponential.



As additional criteria for comparison between the four models, the AIC, BIC and log-likelihood values are presented on Table 13. Although the time-dependent slopes parameterization showed some lack of fit on the survival process, it still resulted in the lowest values for AIC and BIC and the highest log-likelihood. Its improvement in fit is also confirmed by the likelihood ratio test, which resulted in a test statistic of 18.3 and a p-value below 0.001. Again, the similarity between the current value and the 6-months lagged model are apparent.

Table 13 – Measures of model fit

Parameterization	AIC	BIC	Log-likelihood
Current value	6063.931	6108.788	-3016.966
6-months lagged	6065.167	6110.023	-3017.583
Time-dependent slopes	6047.669	6095.516	-3007.835
Cumulative effects	6073.913	6118.770	-3021.957

3.2.5 Discussion

When jointly modelling both the square root of the CD4 cell count and the time to event data, we observed that HCV coinfection becomes a significant variable in estimating the

biomarker values over time. While the main effect size remains similar throughout the separate and the joint models, the joint models provide smaller standard errors, increasing the statistical power of this sample and providing results similar to the ones found when separately modeling $\sqrt{CD4}$ via linear mixed effects regression using a larger sample of 3340 individuals (SILVA et al., 2018), a result that was not observed in the separate estimation of this longitudinal process.

The current value joint model parameterization had been used previously to analyze this same dataset in work submitted by Brum and Previdelli to *Communications in Statistics: Case Studies and Data Analysis*. Although the model utilized in their analysis did not account for interactions between coinfection group and time, similar results were found. Our application adds to the previou one by showing that the current value parameterization and the 6-month lagged parameterization yielded very similar results, not only in estimated coefficients, but in model residuals and overall fit measures, indicating that this data has the potential to provide useful insights on the disease progression and immunologic recovery of HIV patients up to six months in advance, allowing health professionals to make better informed decisions regarding patient treatment and prognosis.

In addition, the use of the time-dependent slopes parameterization provides an interesting biological interpretation, given that the current value association parameter is no longer significant when considering the overal longitudinal trajectory's slope as a covariate. Given this result, two patients who present with the same $\sqrt{CD4}$ count at a given time point, but whose overal trajectories differ in slopes, generate different estimates of time to immunologic recovery. This result also highlights the importance of longitudinal and long term follow-up of these patients, given that this interpretation depends on the availability of several time points of data. In addition, these findings are in accordance to the available literature, that mentions CD4/CD8 cell ratio may be insufficient even when CD4 cell count is at a healthy range (LU et al., 2015; SILVA et al., 2018).

3.2.6 Conclusion

Our study suggests an increasing trajectory of CD4 cell count over time may be more important than the absolute values of CD4 itself when predicting time to immunologic recovery. We emphasize that joint modeling can be a useful tool to increase the power of a relatively small sample of subjects being monitored longitudinally. This application of joint model parameterizations highlights the different biological interpretations that can be made with these models, from a simple present-time association, to predictions made ahead of time, to a more complex relationship between the longitudinal changes in the biomarker and the time to event. Not only do these applications provide valuable insight into the mechanism of

disease progression, they also have the potential to become dynamic prognostic tools aiding healthcare professional in personalized medicine.

Bibliography

ALIMI, R.; HAMI, M.; AFZALAGHAEE, M.; NAZEMIAN, F.; MAHMOODI, M.; YASERI, M.; ZERAATI, H. Multivariate Longitudinal Assessment of Kidney Function Outcomes on Graft Survival after Kidney Transplantation Using Multivariate Joint Modeling Approach: A Retrospective Cohort Study. *Iranian Journal of Medical Sciences*, n. Online First, nov. 2020. Disponível em: <https://doi.org/10.30476/ijms.2020.82857.1144>.

ANDREONI, M.; GIACOMETTI, A.; MAIDA, I.; MERAVIGLIA, P.; RIPAMONTI, D.; SARMATI, L. HIV-HCV co-infection: epidemiology, pathogenesis and therapeutic implications. *European Review for Medical and Pharmacological Sciences*, v. 16, n. 11, p. 1473–1483, out. 2012. ISSN 1128-3602.

ATILA, C.; SAILER, C. O.; BASSETTI, S.; TSCHUDIN-SUTTER, S.; BINGISSER, R.; SIEGEMUND, M.; OSSWALD, S.; RENTSCH, K.; RUEEGG, M.; SCHAERLI, S.; KUSTER, G. M.; TWERENBOLD, R.; CHRIST-CRAIN, M. Prevalence and outcome of dysnatremia in patients with COVID-19 compared to controls. *European Journal of Endocrinology*, v. 184, n. 3, p. 409–418, mar. 2021. ISSN 1479-683X.

BADER, F.; MANLA, Y.; ATALLAH, B.; STARLING, R. C. Heart failure and COVID-19. *Heart Failure Reviews*, v. 26, n. 1, p. 1–10, jan. 2021. ISSN 1573-7322. Disponível em: <https://doi.org/10.1007/s10741-020-10008-2>.

BANDAK, G.; KASHANI, K. B. *Chloride in intensive care units: a key electrolyte*. [S.l.], 2017. Type: article. Disponível em: <https://f1000research.com/articles/6-1930>.

BONACINI, M.; LOUIE, S.; BZOWEJ, N.; WOHL, A. R. Survival in patients with HIV infection and viral hepatitis B or C: a cohort study. *AIDS*, v. 18, n. 15, p. 2039, out. 2004. ISSN 0269-9370. Disponível em: <https://journals.lww.com/aidsonline/Fulltext/2004/10210/Survival_in_patients_with_HIV_infection_and_viral.8.aspx>.

BROWN, E. R.; IBRAHIM, J. G.; DEGRUTTOLA, V. A Flexible B-Spline Model for Multiple Longitudinal Biomarkers and Survival. *Biometrics*, v. 61, n. 1, p. 64–73, 2005. ISSN 1541-0420. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0006-341X.2005.030929.x. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.2005.030929.x>.

BRUM, B.; PREVIDELLI, I. *Modelo conjunto para um estudo de coorte de HIV coinfectada pelos vírus HBV e HCV*. Dissertação (Mestrado) — Universidade Estadual de Maringá, 2018.

CEKIC, S.; AICHELE, S.; BRANDMAIER, A. M.; KöHNCKE, Y.; GHISLETTA, P. A Tutorial for Joint Modeling of Longitudinal and Time-to-Event Data in R. *Quantitative and Computational Methods in Behavioral Sciences*, p. 1–40, maio 2021. ISSN 2699-8432. Disponível em: <https://qcmb.psychopen.eu/index.php/qcmb/article/view/2979>.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. [S.l.]: Edgard Blücher, 2006. ISBN 978-85-212-0384-1.

COX, D. R.; HINKLEY, D. V. *Theoretical Statistics*. [S.l.]: CRC Press, 1979. ISBN 978-0-412-16160-5.

DIGGLE, P.; HEAGERTY, P.; LIANG, K.; ZEGER, S. *Analysis of Longitudinal Data*. OUP Oxford, 2013. (Oxford Statistical Science Series). ISBN 978-0-19-967675-0. Disponível em: <https://books.google.com.br/books?id=ur0BIXPuOukC>.

FLOR, J. C. de L.; GOMEZ-BERROCAL, A.; MARSCHALL, A.; VALGA, F.; LINARES, T.; ALBARRACIN, C.; RUIZ, E.; GALLEGOS, G.; GóMEZ, A.; SANTOS, A. de los; RODELES, M. Impacto de la corrección temprana de la hiponatremia en el pronóstico de la infección del síndrome respiratorio agudo grave del coronavirus 2 (SARS-CoV-2). *Medicina Clinica*, jul. 2021. ISSN 0025-7753. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8318697/>.

GRODIN, J. L.; SIMON, J.; HACHAMOVITCH, R.; WU, Y.; JACKSON, G.; HALKAR, M.; STARLING, R. C.; TESTANI, J. M.; TANG, W. H. W. Prognostic Role of Serum Chloride Levels in Acute Decompensated Heart Failure. *Journal of the American College of Cardiology*, v. 66, n. 6, p. 659–666, ago. 2015. ISSN 1558-3597.

GRUTTOLA, V. D.; TU, X. M. Modelling Progression of CD4-Lymphocyte Count and Its Relationship to Survival Time. *Biometrics*, v. 50, n. 4, p. 1003–1014, 1994. ISSN 0006-341X. Publisher: [Wiley, International Biometric Society]. Disponível em: <https://www.jstor.org/stable/2533439>.

GUPTA, R.; KHOURY, J. C.; ALTAYE, M.; JANDAROV, R.; SZCZESNIAK, R. D. Assessing the Relationship between Gestational Glycemic Control and Risk of Preterm Birth in Women with Type 1 Diabetes: A Joint Modeling Approach. *Journal of Diabetes Research*, v. 2020, p. e3074532, jun. 2020. ISSN 2314-6745. Publisher: Hindawi. Disponível em: <https://www.hindawi.com/journals/jdr/2020/3074532/>.

HELLEBERG, M.; KRONBORG, G.; ULLUM, H.; RYDER, L. P.; OBEL, N.; GERSTOFT, J. Course and Clinical Significance of CD8 $^{+}$ T-Cell Counts in a Large Cohort of HIV-Infected Individuals. *Journal of Infectious Diseases*, v. 211, n. 11, p. 1726–1734, jun. 2015. ISSN 0022-1899, 1537-6613. Disponível em: <https://academic.oup.com/jid/article-lookup/doi/10.1093/infdis/jiu669>.

HENRY, B. M.; OLIVEIRA, M. H. S. de; BENOIT, J.; LIPPI, G. Gastrointestinal symptoms associated with severity of coronavirus disease 2019 (COVID-19): a pooled analysis. *Internal and Emergency Medicine*, abr. 2020. ISSN 1970-9366. Disponível em: <https://doi.org/10.1007/s11739-020-02329-9>.

HSIEH, F.; TSENG, Y.-K.; WANG, J.-L. Joint Modeling of Survival and Longitudinal Data: Likelihood Approach Revisited. *Biometrics*, v. 62, n. 4, p. 1037–1043, 2006. ISSN 1541-0420. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1541-0420.2006.00570.x. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2006.00570.x>.

IBRAHIM, J. G.; CHU, H.; CHEN, L. M. Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data. *Journal of Clinical Oncology*, v. 28, n. 16, p. 2796–2801, jun. 2010. ISSN 0732-183X. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4503792/>.

JI, Y.; LI, L. Lower serum chloride concentrations are associated with increased risk of mortality in critically ill cirrhotic patients: an analysis of the MIMIC-III database. *BMC Gastroenterology*, v. 21, n. 1, p. 200, dez. 2021. ISSN 1471-230X. Disponível em: <https://bmcgastroenterol.biomedcentral.com/articles/10.1186/s12876-021-01797-3>.

KAPLAN, E. L.; MEIER, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, v. 53, n. 282, p. 457–481, jun. 1958. ISSN 0162-1459. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1958.10501452. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501452>.

KIMURA, S.; HOZ, M. A. A. de la; RAINES, N. H.; CELI, L. A. Association of Chloride Ion and Sodium-Chloride Difference With Acute Kidney Injury and Mortality in Critically Ill Patients. *Critical Care Explorations*, v. 2, n. 12, p. e0247, nov. 2020. ISSN 2639-8028. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7688253/>.

KURLAND, B. F.; JOHNSON, L. L.; EGLESTON, B. L.; DIEHR, P. H. Longitudinal Data with Follow-up Truncated by Death: Match the Analysis Method to Research Aims. *Statistical science : a review journal of the Institute of Mathematical Statistics*, v. 24, n. 2, p. 211, 2009. ISSN 0883-4237. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2812934/>.

LU, W.; MEHRAJ, V.; VYBOH, K.; CAO, W.; LI, T.; ROUTY, J.-P. CD4:CD8 ratio as a frontier marker for clinical outcome, immune dysfunction and viral reservoir size in virologically suppressed HIV-positive patients. *Journal of the International AIDS Society*, v. 18, n. 1, p. 20052, 2015. ISSN 1758-2652.

MAATEN, J. M. T.; DAMMAN, K.; HANBERG, J. S.; GIVERTZ, M. M.; METRA, M.; O'CONNOR, C. M.; TEERLINK, J. R.; PONIKOWSKI, P.; COTTER, G.; DAVISON, B.; CLELAND, J. G.; BLOOMFIELD, D. M.; HILLEGE, H. L.; VELDHUISEN, D. J. van; VOORS, A. A.; TESTANI, J. M. Hypochloremia, Diuretic Resistance, and Outcome in Patients With Acute Heart Failure. *Circulation. Heart Failure*, v. 9, n. 8, p. e003109, ago. 2016. ISSN 1941-3297.

MARTTINEN, M.; WILKMAN, E.; PETäJä, L.; SUOJARANTA-YLINEN, R.; PETTILä, V.; VAARA, S. T. Association of plasma chloride values with acute kidney injury in the critically ill - a prospective observational study. *Acta Anaesthesiologica Scandinavica*, v. 60, n. 6, p. 790–799, jul. 2016. ISSN 1399-6576.

MAUFF, K.; STEYERBERG, E. W.; NIJPELS, G.; HEIJDEN, A. A. van der; RIZOPOULOS, D. Extension of the association structure in joint models to include weighted cumulative effects. *Statistics in Medicine*, v. 36, n. 23, p. 3746–3759, 2017. ISSN 1097-0258. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7385. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7385>.

PINHEIRO, J.; BATES, D.; R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*. [S.l.], 2022. R package version 3.1-160. Disponível em: <https://CRAN.R-project.org/package=nlme>.

PIè, C. S.; RUé, M.; FABUEL, H. P.; FORTE, A.; ARMERO, C.; PIULACHS, X.; PáEZ, ; MELIS, G. G. A shared-parameter joint model for prostate cancer risk and psa longitudinal profiles. In: . Fundación Mapfre, 2015. p. 711–719. ISBN 978-84-9844-496-4. Accepted: 2016-02-08T12:54:41Z. Disponível em: <https://upcommons.upc.edu/handle/2117/82670>.

POURFRIDONI, M.; ABBASNIA, S. M.; SHAFAEI, F.; RAZAVIYAN, J.; HEIDARI-SOURESHJANI, R. Fluid and Electrolyte Disturbances in COVID-19 and Their Complications. *BioMed Research International*, v. 2021, p. 1–5, abr. 2021. ISSN 2314-6141, 2314-6133. Disponível em: <https://www.hindawi.com/journals/bmri/2021/6667047/>.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. Disponível em: <https://www.R-project.org/>.

REY, J. R.; CARO-CODóN, J.; ROSILLO, S. O.; INIESTA, M.; CASTREJóN-CASTREJóN, S.; MARCO-CLEMENT, I.; MARTíN-POLO, L.; MERINO-ARGOS, C.; RODRíGUEZ-SOTELO, L.; GARCíA-VEAS, J. M.; MARTíNEZ-MARíN, L. A.; MARTíNEZ-COSSIANI, M.; BUñO, A.; GONZALEZ-VALLE, L.; HERRERO, A.; LóPEZ-SENDóN, J. L.; MERINO, J. L.; INVESTIGATORS, f. t. C.-C. Heart failure in COVID-19 patients: prevalence, incidence and prognostic implications. *European Journal of Heart Failure*, v. 22, n. 12, p. 2205–2215, 2020. ISSN 1879-0844. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejhf.1990. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ejhf.1990>.

RIZOPOULOS, D. JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, v. 35, n. 9, p. 1–33, 2010. Disponível em: <https://doi.org/10.18637/jss.v035.i09>.

RIZOPOULOS, D. *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. [S.l.]: CRC Press, 2012. Google-Books-ID: xotIpb2duaMC. ISBN 978-1-4398-7286-4.

SELF, S.; PAWITAN, Y. Modeling a Marker of Disease Progression and Onset of Disease. In: JEWELL, N. P.; DIETZ, K.; FAREWELL, V. T. (Ed.). *AIDS Epidemiology: Methodological Issues*. Boston, MA: Birkhäuser, 1992. p. 231–255. ISBN 978-1-4757-1229-2. Disponível em: <https://doi.org/10.1007/978-1-4757-1229-2_11>.

SHRIMANKER, I.; BHATTARAI, S. Electrolytes. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2022. Disponível em: <http://www.ncbi.nlm.nih.gov/books/NBK541123/>.

SILVA, C. M. d.; PEDER, L. D. d.; SILVA, E. S.; PREVIDELLI, I.; PEREIRA, O. C. N.; TEIXEIRA, J. J. V.; BERTOLINI, D. A. Impact of HBV and HCV coinfection on CD4 cells

among HIV-infected patients: a longitudinal retrospective study. *Journal of Infection in Developing Countries*, v. 12, n. 11, p. 1009–1018, nov. 2018. ISSN 1972-2680.

SULTANA, R.; AHSAN, A. A.; FATEMA, K.; AHMED, F.; SAHA, D. K.; SAHA, M.; NAZNEEN, S.; MAHBUB, N.; ASHRAF, E. Pattern of electrolytes in a cohort of critically ill COVID-19 patients. *BIRDEM Medical Journal*, p. 46–50, dez. 2020. ISSN 2305-3720. Disponível em: <https://www.banglajol.info/index.php/BIRDEM/article/view/50980>.

TAN, C. W.; HO, L. P.; KALIMUDDIN, S.; CHERNG, B. P. Z.; TEH, Y. E.; THIEN, S. Y.; WONG, H. M.; TERN, P. J. W.; CHANDRAN, M.; CHAY, J. W. M.; NAGARAJAN, C.; SULTANA, R.; LOW, J. G. H.; NG, H. J. Cohort study to evaluate the effect of vitamin D, magnesium, and vitamin B12 in combination on progression to severe outcomes in older patients with coronavirus (COVID-19). *Nutrition (Burbank, Los Angeles County, Calif.)*, v. 79-80, p. 111017, dez. 2020. ISSN 1873-1244.

TEZCAN, M.; GOKCE, G. D.; SEN, N.; KAYMAK, N. Z.; OZER, R. Baseline electrolyte abnormalities would be related to poor prognosis in hospitalized coronavirus disease 2019 patients. *New Microbes and New Infections*, v. 37, p. 100753, set. 2020. ISSN 2052-2975. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7462442/>.

THERNEAU, T. M. *A Package for Survival Analysis in R*. [S.l.], 2022. R package version 3.4-0. Disponível em: <https://CRAN.R-project.org/package=survival>.

TSIATIS, A. A.; DEGRUTTOLA, V.; WULFSOHN, M. S. 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*, v. 90, n. 429, p. 27–37, 1995. ISSN 0162-1459. Publisher: [American Statistical Association, Taylor & Francis, Ltd.]. Disponível em: <https://www.jstor.org/stable/2291126>.

VRIEZE, S. I. Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, v. 17, n. 2, p. 228, 2012. ISSN 1939-1463. Publisher: US: American Psychological Association. Disponível em: <https://psycnet.apa.org/fulltext/2012-03019-001.pdf>.

WU, L.; LIU, W.; YI, G. Y.; HUANG, Y. Analysis of Longitudinal and Survival Data: Joint Modeling, Inference Methods, and Issues. *Journal of Probability and Statistics*, v. 2012, p. e640153, dez. 2011. ISSN 1687-952X. Publisher: Hindawi. Disponível em: <https://www.hindawi.com/journals/jps/2012/640153/>.

WULFSOHN, M. S.; TSIATIS, A. A. A Joint Model for Survival and Longitudinal Data Measured with Error. *Biometrics*, v. 53, n. 1, p. 330–339, 1997. ISSN 0006-341X. Publisher: [Wiley, International Biometric Society]. Disponível em: <https://www.jstor.org/stable/2533118>.