Fast and efficient joint modelling of multivariate longitudinal data and time-to-event data with a pairwise-fitting approach

Peer-reviewed author version

De Witte , Dries; MOLENBERGHS, Geert; ALONSO ABAD, Ariel; NEYENS, Thomas & VERBEKE, Geert (2025) Fast and efficient joint modelling of multivariate longitudinal data and time-to-event data with a pairwise-fitting approach. In: Statistical modelling,.

DOI: 10.1177/1471082X251328452
Handle: http://hdl.handle.net/1942/45929

Fast and Efficient Joint modelling of Multivariate Longitudinal Data and Time-to-Event Data With a Pairwise-Fitting Approach

Dries De Witte [1], **Geert Molenberghs** [1,2], **Ariel Alonso Abad** [1,2], **Thomas Neyens** [1,2] and **Geert Verbeke** [1,2]

[1] L-BioStat, KU Leuven, Leuven, Belgium

[2] I-BioStat, Hasselt University, Diepenbeek, Belgium

Address for correspondence: Dries De Witte, L-BioStat, KU Leuven, Kapucijnenvoer 7, 3000 Leuven, Belgium.

E-mail: dries.dewitte@kuleuven.be.

Abstract: In empirical studies, multiple outcomes are often measured repeatedly over time, and interest frequently lies in studying the association between these longitudinal outcomes and a time-to-event outcome. Therefore, shared-parameter joint models for longitudinal and time-to-event outcomes have been developed.

However, while such joint models in theory also allow for multiple longitudinal outcomes, they are often restricted to a limited number of outcomes due to computational complexity when fitting the models. To address this problem, we propose a new joint model, which is based on correlated instead of shared random effects, and for which a pairwise-modelling strategy can be used. In this approach, the longitudinal outcomes are modelled with (generalized) linear mixed models and the survival outcome with a Weibull proportional hazards frailty model. Instead of fitting the full joint model, this approach involves fitting all possible bivariate models, and inference is based on pseudo-likelihood theory. The main advantage of our approach is that there is no restriction on the number of longitudinally measured outcomes that are jointly modelled with the time-to-event outcome.

Key words: High-dimensional Data; Linear Mixed Models; Pseudo-likelihood; Survival Data; Weibull Proportional Hazards Frailty Model

1 Introduction

In empirical research, subjects, such as patients in medical studies, are often followed-up over a certain time period, and several outcomes are measured repeatedly over time. A typical example are biomarker measurements that are collected for each subject. Alongside these longitudinal data, the time until a well-defined event can be of interest, for instance the time until death. Longitudinal data are

often analysed using (generalized) linear mixed models (Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005). For survival data, Cox' proportional hazards models or accelerated failure time models can be used (Collett, 2023). When there is a strong relationship between a longitudinal biomarker and the survival outcome, standard mixed models separately fit for the longitudinal component can result in biased estimation due to missingness (Wu and Carroll, 1988). To take into account this bias, or when interest lies in studying the association between the longitudinal outcomes and the survival outcome, a joint model for time-to-event data and multivariate longitudinal data is needed. Up until now, shared-parameter joint models, in which the survival outcome and longitudinal outcomes are linked by including (functions of) the random effects of the longitudinal outcomes as predictors in the survival model, have received attention.

Considerable developments have taken place in recent years in this field, e.g., by Rizopoulos (2012). An overview of such joint models can be found in Papageorgiou et al. (2019) and Tsiatis and Davidian (2004). In these joint models, the association between the longitudinal outcomes and the survival outcome is reflected by the regression coefficients of the functions of the random effects in the survival model. Several extensions of this model have been proposed, such as incorporating recurrent events and competing risks (Papageorgiou et al., 2019). Nevertheless, when a large number of longitudinal outcomes is modelled jointly with the survival outcome, the dimension of the random effects increases and computational problems (e.g., convergence problems) often arise because (numerical) integration

over the entire vector of all random effects is necessary. In an attempt to alleviate the computational burden, several algorithms have been proposed. For instance, Albert and Shih (2010) use the pairwise-fitting approach to model multivariate longitudinal data, but employ a two-stage approach to link the longitudinal outcomes with a discrete survival model. However, this method still relies on the shared-parameter model formulation. As a result, their approach remains restricted in the number of longitudinal outcomes it can handle.

Moreover, it has been shown that such a two-stage approach, where the estimation of the joint model is done in two steps, is biased. This bias arises since the estimates are not coming from the full joint distribution, and are therefore not corrected for missingness (Tsiatis and Davidian, 2004; Rizopoulos, 2012; Mauff et al., 2020). Mauff et al. (2020) proposed the use of a correction factor to eliminate this bias. However, most importantly, their model still only allows for a limited number of longitudinal outcomes, since a multivariate (generalized) linear mixed model has to be fitted in the first stage. Their simulation studies included up to six outcomes. Moreover, using the **JMbayes** package (Rizopoulos, 2016) in R, Nguyen et al. (2023) experienced convergence problems when trying to jointly model 32 longitudinal variables with the survival outcome. These authors reached convergence when only nine variables were selected.

Recently, Hickey et al. (2018) have developed the R package **joineRML** that uses the Monte Carlo expectation-maximization algorithm to maximize the full likelihood. While this package indeed allows for modelling multivariate longitudinal

data, it remains unclear how the algorithm performs when encountering high-dimensional multivariate longitudinal data. Models can also be fitted efficiently based on a Bayesian approach by using Markov Chain Monte Carlo (MCMC) methods (Lawrence Gould et al., 2015). However, when the random effects are of high dimension, MCMC methods can be computationally very demanding and it can be difficult to reach convergence.

We propose a new joint model for high-dimensional longitudinal data and survival data, which allows for the use of the pairwise-fitting approach. This approach was previously proposed by Fieuws and Verbeke (2006) to jointly model multivariate longitudinal data. Mauff et al. (2021) adapted this approach to the Bayesian setting, and tried to incorporate the multivariate joint model in a survival model by using a corrected two-stage approach (Mauff et al., 2020). However, they were unable to obtain a theoretically proper multivariate posterior distribution. We propose a different joint model, which is based on correlated random effects instead of shared random effects. Albert and Shih (2010) used the pairwise-fitting approach in a two-stage joint model with discrete time-to-event data, in which missingness is taken into account with a regression calibration approach. In our proposed joint model, a frailty model for univariate survival data is used for the survival outcome by employing a Weibull model with random effects (Duchateau and Janssen, 2008). This Weibull model is then linked to the mixed models for the longitudinal outcomes by assuming a joint multivariate distribution on the random effects. By doing so, the number of outcomes that can be modelled jointly

is no longer restricted. Jointly modelling a longitudinal outcome and a survival outcome by allowing for correlation between the random effects and frailties has been previously proposed in literature (Huang et al., 2011; Li et al., 2009). However, these studies were limited to models with a single longitudinal outcome. Our approach extends this framework by allowing for the inclusion of multiple longitudinal outcomes.

The rest of the paper is organized as follows. The case studies used to motivate and illustrate our methodology are described in Section 2. The methodology is laid out in Section 3, and the results of applying it to the example datasets is the subject of Section 4. In Section 5, simulation studies are reported to examine the performance of the proposed methodology.

2 Motivating Case Studies

2.1 Idiopathic Pulmonary Fibrosis

A motivating dataset for our methodology is the Idiopathic Pulmonary Fibrosis (IPF) dataset, in which a cohort of 57 patients with idiopathic pulmonary fibrosis (IPF) were followed longitudinally (Molyneaux et al., 2017), and which was previously analysed by Sun et al. (2019). Idiopathic pulmonary fibrosis (IPF) is a progressive and fatal lung disease characterized by the gradual scarring (fibrosis) of lung tissue. The median survival of patients with IPF is only 3–5 years (Schwartz et al., 1994). The dataset comprises data of 20 gene expression profiles that were

measured in the peripheral whole blood at each visit. Measurements were taken at fixed timepoints: at baseline (timepoint = 0), 1 month (timepoint = 1), 3 months (timepoint = 2), 6 months (timepoint = 3) and 1 year (timepoint = 4). The median number of measurements per patient is 3, with the number of measurements ranging from a minimum of 1 to a maximum of 5. In addition, time to death, age, gender and percent predicted forced vital capacity (FVC) were recorded. Out of the 57 patients, 23 patients (40%) died, while 34 (60%) patients were censored. Following Sun et al. (2019), age, FVC, and gene expression data, are standardized to have mean 0 and variance 1. Our aim is to study the association between the risk of death and the longitudinal gene expression profiles.

2.2 Primary Biliary Cirrhosis

In a second case study, we used the primary biliary cirrhosis dataset (pbc2) from the **JMBayes2** package. This dataset contains follow up data of 312 randomized patients, collected at the Mayo Clinic between 1974 to 1984 (Murtaugh et al., 1994). On average, there were 6.23 measurements for each individual, with a maximum of 16 measurements and a minimum of 1 measurement. During follow up, several biomarkers were recorded repeatedly for each patient. In our application, we focused on three biomarkers: serum bilirubin (mg/dl), serum albumin (mg/dl), and prothrombin time (seconds). Of interest was studying the relationship between these longitudinal biomarkers and time until death, while correcting for age at registration and sex. In total, 140 patients (44.87%) in the dataset died, mean-

ing that 55.13% of the patients were censored. Many researchers have extensively studied this dataset using joint modelling techniques (Crowther et al., 2013; Rizopoulos, 2012; Hickey et al., 2018), which makes it an ideal first case study to illustrate our proposed methodology.

3 Methodology

3.1 Joint Model for Multivariate Longitudinal Data and Time-to-Event Data

3.1.1 Submodel for Multivariate Longitudinal Data

To model two or more longitudinal outcomes simultaneously, a joint model for multivariate longitudinal data is needed. In the random-effects approach, the joint model is built by assuming a mixed model for each outcome (Laird and Ware, 1982; Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005). Let us consider L longitudinal outcomes. Then, for each outcome l $(l = 1, ..., L)$, we let Y_{ijl} denote the jth measurement $(j = 1, ..., n_{il})$ for subject $i = 1, ..., N$. The vector $\boldsymbol{Y_{il}'} = (Y_{i1l}, ..., Y_{in_{il}l})$ then represents the vector of all measurements for the lth outcome for subject i. For continuous outcomes, linear mixed models (LMMs) are often used. In a LMM, we assume that $\boldsymbol{Y_{il}}$ satisfies

$$\boldsymbol{Y_{il}}|\boldsymbol{b_{il}} \sim N(\boldsymbol{X_{il}}\boldsymbol{\beta_l} + \boldsymbol{Z_{il}}\boldsymbol{b_{il}}, \boldsymbol{\Sigma_{il}}),$$

in which $\boldsymbol{X_{il}}$ and $\boldsymbol{Z_{il}}$ are known design matrices, $\boldsymbol{\beta_l}$ is the vector of fixed-effects regression parameters, $\boldsymbol{b_{il}}$ is the vector of random-effects regression parameters, and $\boldsymbol{\Sigma_{il}}$ is an $(n_{il} \times n_{il})$ covariance matrix that depends on i only through its dimension n_{il}. Assuming conditional independence would imply $\boldsymbol{\Sigma_{il}} = \sigma_l^2 \boldsymbol{I_{n_{il}}}$, with $\boldsymbol{I_{n_{il}}}$ being the identity matrix. For discrete outcomes, a generalized linear mixed model (GLMM) can be used (Molenberghs and Verbeke, 2005). The GLMM assumes that the outcomes Y_{ijl} are independent, conditionally on the random effects $\boldsymbol{b_{il}}$, and have a density function that belongs to the exponential family of the form

$$f_i(y_{ijl}|\boldsymbol{b_{il}}) = \exp[(y_{ijl}\eta_{ijl} - a(\eta_{ijl}))/\phi_l + c(y_{ijl}, \phi_l)],$$

with mean $E(Y_{ijl}|\boldsymbol{b_{il}}) = a'(\eta_{ijl}) = \mu_{ijl}(\boldsymbol{b_{il}})$ and variance $Var(Y_{ijl}|\boldsymbol{b_{il}}) = \phi_l a''(\eta_{ijl})$, and with $h_l(\boldsymbol{\mu_{il}}(\boldsymbol{b_{il}})) = \boldsymbol{X_{il}\beta_l} + \boldsymbol{Z_{il}b_{il}}$ for a known link function h_l.

The L longitudinal outcomes can be modelled jointly by linking the corresponding random effects (Verbeke et al., 2014; Fieuws and Verbeke, 2004). Let $\boldsymbol{b_i}$ denote the vector encompassing all random effects from all mixed models, that is $\boldsymbol{b_i} = (\boldsymbol{b_{i1}}, \boldsymbol{b_{i2}}, ..., \boldsymbol{b_{iL}})$. To simultaneously model all L outcomes, a joint distribution for these random effects can be specified. Often, we assume that $\boldsymbol{b_i}$ is normally distributed with mean zero and variance-covariance $\boldsymbol{D}$. From the components in $\boldsymbol{D}$, the association between the outcomes can be derived. Further, we assume that, conditionally on $\boldsymbol{b_i}$, the outcomes are independent. Thus, the association between the outcomes is fully captured by the association between the random effects.

3.1.2 Submodel for Time-to-Event Data

Time-to-event data are often analysed with proportional hazards (PH) models (Duchateau and Janssen, 2008; Collett, 2023). Let t_i^* denote the true event time for subject i, $(i = 1, \ldots, N)$. Assuming that subjects are potentially right censored by censoring time c_i, the observed event time is $t_i = min(T_i^*, C_i)$. δ_i is the event indicator, which is equal to one if the true event time is observed, that is if $t_i^* < C_i$, and equal to zero otherwise. We further assume that the censoring is non-informative. The general form of a proportional hazards model can then be expressed as

$$h_i(t) = h_0(t) \exp(\boldsymbol{\alpha'} \boldsymbol{w_i}),$$

where $h_0(t)$ is the baseline hazard function, $\boldsymbol{w_i}$ the vector of known covariates and $\boldsymbol{\alpha}$ the vector of corresponding regression coefficients. Note that it is also possible to include exogeneous (external) time-dependent covariates in this model. Then, $\boldsymbol{w_i}$ can be expressed as $\boldsymbol{w_i(t)}$. In survival analysis, a Weibull distribution with shape parameter ρ and scale parameter λ is often assumed for the event times, such that $T_i \sim W(\lambda \exp(\boldsymbol{\alpha'} \boldsymbol{w_i}), \rho)$ (Collett, 2023). Under this assumption, the baseline hazard corresponds to $h_0(t) = \lambda \rho t^{\rho - 1}$.

To link the Weibull PH model to the mixed model for the longitudinal outcomes, we expand the model by including a random effect, such that all models can be linked together via a common distribution for all random effects. For this purpose, a univariate frailty model, previously proposed by Duchateau and Janssen (2008),

can be used:

$$h_i(t) = a_i \lambda \rho t^{\rho-1} \exp(\boldsymbol{\alpha'} \boldsymbol{w_i}) \tag{3.1}$$

In this model, each subject has its own random effect a_i, which is called a frailty. This frailty term can be used to describe heterogeneity or to determine which subjects are more frail; a high frailty value leads to an increased risk for the event, while a low frailty value leads to a decreased risk for the event.

Several distributions for the random effect a_i can be assumed. Elbers and Ridder (1982) have shown that a frailty model is identifiable with univariate data, as long as the frailty distribution has a finite mean and when covariates are included in the model. The two most popular distributions are the gamma distribution and the log-normal distribution, since the frailty cannot be negative (Duchateau and Janssen, 2008). When a gamma distribution is used to ensure identifiability of the model, we use a one-parameter gamma distribution with mean one: $a_i \sim \Gamma(\theta, 1/\theta)$. When a log-normal distribution is used, identifiability is ensured by setting the mean equal to one: $a_i \sim LN(1, \sigma_a^2)$.

To account for the association between the survival outcome and the multiple longitudinal outcomes, we allow the random effect of the Weibull PH frailty model, a_i, to be correlated with the random effects of the L linear mixed models $\boldsymbol{b_i}$. The correlations between the frailty and the random effects then represent the associations. Let us consider for instance the correlation between the survival outcome and longitudinal outcome l modelled with a random intercept b_{il} only. A positive correlation between a_i and b_{il} then indicates that a subject with a higher

value for the longitudinal outcome has an increased risk for the event. A negative correlation might indicate that a subject with a higher value for the longitudinal outcome has a decreased risk for the event. An important remark is that this approach allows for the inclusion of multiple survival outcomes as well. In our model, survival outcomes are simply treated as additional outcomes within the model, while in the shared-parameter joint model, only a single survival outcome is typically included, sharing random effects with the longitudinal outcomes.

Additionally, similar to the shared-parameter joint model, the visiting processes (Lipsitz et al., 2002) are assumed to be independent, given the observed history. This means that we assume that the decision of a subject to visit the clinic only depends on the observed past history but not on underlying, latent subject characteristics (Rizopoulos, 2012). In case this assumption is violated, the visiting process itself should be modelled. To this end, the visiting times can be seen as recurrent events, and the joint model can be extended with a time-to-event model for this type of (multivariate) survival data. The implementation of this extension in our proposed model is very straightforward, since the Weibull PH frailty model (3.1) is especially suitable for modelling multivariate survival data (Therneau and Grambsch, 2013).

3.2 Estimation

3.2.1 Pseudo-Likelihood

For many models, maximizing the full likelihood can be difficult. For instance, the normalizing constant can take a complicated form in multivariate exponential family models (Molenberghs and Verbeke, 2005). A possible solution might then be to replace the numerically challenging joint density by a simpler function that is a product of conditional or marginal densities, which is the idea behind pseudo-likelihood, or composite likelihood, estimation (Arnold and Strauss, 1991; Geys et al., 1999; Aerts et al., 2002; Varin et al., 2011). For instance, one can replace the joint density by all pairwise densities over the set of all possible pairs. Consider for example a three-way density

$$L_i = f_i(\mathbf{y_{1i}}, \mathbf{y_{2i}}, \mathbf{y_{3i}}|\theta_\mathbf{i}).$$

This can be replaced by the product

$$L_i^* = f_i(\mathbf{y_{1i}}, \mathbf{y_{2i}}|\theta_\mathbf{i}^*) \times f_i(\mathbf{y_{1i}}, \mathbf{y_{3i}}|\theta_\mathbf{i}^*) \times f_i(\mathbf{y_{2i}}, \mathbf{y_{3i}}|\theta_\mathbf{i}^*).$$

It is important to note that this does not affect the interpretation of the model and that the model parameters have the same meaning as with full likelihood. We refer to Geys et al. (1999) and Aerts et al. (2002) for a more formal introduction to pseudo-likelihood estimation. They also describe the required regularity conditions and the main asymptotic properties of the pseudo-likelihood estimators. The consistency and asymptotic normality of these estimators have been demonstrated by Arnold and Strauss (1991). They also show that a pseudo-likelihood estimator

is always less efficient than the maximum-likelihood estimator. Nevertheless, Aerts et al. (2002) state that the efficiency losses are minimal. Specifically for pairwise likelihood, simulation studies have shown high efficiency relative to a full likelihood approach (Renard et al., 2004; Fieuws et al., 2006; Feddag and Bacci, 2009; Varin et al., 2011). In addition, pseudo-likelihood is often still computationally feasible, when full likelihood is no longer possible. Geys et al. (1999) also argue that the efficiency loss is often negligible and can be justified by the computational advantages that it offers. Furthermore, important to note is that pseudo-likelihood is not a full likelihood method, meaning that, a priori, it is not guaranteed that the analysis is valid under missing at random (MAR) (Molenberghs et al., 2011). The process is said to be MAR, when the missingness it independent of the unobserved measurements, conditional on the observed ones (Little and Rubin, 2019).

3.2.2 Pairwise-Fitting Approach

In principle, parameter estimates for the joint model described in Section 3.1 can be obtained by using standard software for fitting mixed models such as PROC NLMIXED in SAS. However, when a large number of outcomes is modelled jointly, the dimension of the random effects vector becomes large, such that maximizing the likelihood becomes infeasible and computational problems arise (Ivanova et al., 2016; Fieuws and Verbeke, 2006). To solve these computation problems, Fieuws and Verbeke (2006) propose to reduce the dimensionality of the problem by first fitting all bivariate models for all possible pairs $((S, Y_1), (S, Y_2),..., (S, Y_L))$ separately (Fieuws et al., 2006, 2007; Verbeke et al., 2014). In this so-called pairwise-

fitting approach, instead of maximizing the likelihood of the full joint model, the log-likelihood of each bivariate joint model for each pair is maximized separately:

$$\sum_{i=1}^{N} l_{rsi}(\boldsymbol{\Theta}_{\boldsymbol{r,s}}|\boldsymbol{y_{ri}}, \boldsymbol{y_{si}}), \quad r = 1, \ldots, m-1, \quad s = r+1, \ldots, m,$$

where $\boldsymbol{\Theta}_{\boldsymbol{r,s}}$ is the vector that contains all parameters in the bivariate joint model for pair (r, s). Further, we can combine all these pair-specific vectors into a stacked vector $\boldsymbol{\Theta}$. Note that vectors $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}^*$ are not identical. Some parameters in $\boldsymbol{\Theta}^*$, for instance the covariance between two random effects of two different outcomes, are estimated in only one bivariate joint model and are linked to one element in $\boldsymbol{\Theta}$. Other parameters, for instance the covariance between two random effects of the same outcome, are estimated in more than one bivariate joint model and are therefore linked to multiple elements in $\boldsymbol{\Theta}$. In the latter case, we can average the different maximum likelihood estimates from all bivariate models to obtain a single estimate for the corresponding parameter. These averages are a linear combination of maximum likelihood estimates and are therefore asymptotically normally distributed. It is possible to use a weighted approach instead of a simple average, for example, by defining weights based on precision estimates. However, such an approach introduces additional variability, as the weights must be estimated and are not fixed. Furthermore, as demonstrated in the simulation studies in Section 5, using a simple average already shows good statistical properties. Therefore, we argue that more complex weighting schemes do not provide substantial improvements over a simple average. This aligns with findings from related context, where simpler weighting schemes have been shown to outperform more complex

alternatives (Hermans et al., 2018; Tibaldi et al., 2003).

However, since we also need to take into account the variability between the pair-specific estimates, we cannot simply average the standard errors to obtain a standard error for the estimates. Moreover, two pairs with a common outcome share information and, hence, the pair-specific estimates are correlated. Therefore, standard maximum likelihood theory cannot be used for conducting inferences. Instead, it is proposed to use ideas from pseudo-likelihood theory for inference for $\boldsymbol{\Theta}$ (Arnold and Strauss, 1991; Geys et al., 1997). The asymptotic multivariate normal distribution for $\hat{\boldsymbol{\Theta}}$, which is the vector containing all pair-specific (maximum likelihood) estimates, can be expressed as

$$\sqrt{N}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \sim MVN(\mathbf{0}, \boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}) \tag{3.2}$$

where $\boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}$ is a robust variance estimator that can be seen as a 'sandwich-type' estimator. However, we are merely interested in the estimates for the parameters in $\boldsymbol{\Theta}^*$. As argued earlier, this can be done by taking averages: $\hat{\boldsymbol{\Theta}}^* = \boldsymbol{A}\hat{\boldsymbol{\Theta}}$, where $\boldsymbol{A}$ is an appropriate weight matrix. It can be seen that $\hat{\boldsymbol{\Theta}}^*$ then follows a multivariate normal distribution with mean $\boldsymbol{\Theta}^*$ and covariance matrix $\boldsymbol{A}\boldsymbol{\Sigma}(\hat{\boldsymbol{\Theta}})\boldsymbol{A}'$, with $\boldsymbol{\Sigma}(\hat{\boldsymbol{\Theta}})$ the covariance matrix for $\hat{\boldsymbol{\Theta}}$ as expressed in (3.2). An important distinction from more classical pseudo-likelihood methods is that in our pairwise approach, the likelihoods of the pairs are maximized separately, which makes our method computationally very feasible because each maximization only involves a (small) subvector of all parameters in the full model. In classical pseudo-likelihood methods, the joint density of the full model is replaced by all pairwise densities,

but is still maximized over the entire vector of all parameters. Parallel processing has the potential to further speed up computation.

3.3 Adjustment for a Non-Positive Definite Variance-Covariance Matrix

Although the estimates of the pairwise-fitting approach are consistent, there is a possibility that in small samples, the estimated variance-covariance matrix $\boldsymbol{D}$ of the random effects may not be positive definite. To address this issue, the nearest positive definite matrix can be obtained. To this aim, Rousseeuw and Molenberghs (1993) have introduced several techniques. One such approach is the eigenvalue method, which corrects non-positive definiteness of $\boldsymbol{D}$ as follows:

$$\boldsymbol{D}_+ = \boldsymbol{L}\tilde{\boldsymbol{E}}\boldsymbol{L}.$$

Here, $\boldsymbol{L}$ represents the collection of orthonormal eigenvectors of $\boldsymbol{D}$, and $\tilde{\boldsymbol{E}}$ denotes the diagonal matrix consisting of the eigenvalues of $\boldsymbol{D}$, but where any negative eigenvalues are substituted with a small positive value. This adjustment, however, impacts the statistical properties of $\boldsymbol{D}$, such as the unbiasedness and sampling variance. Nevertheless, unless in situations where the $\boldsymbol{D}$ matrix is non-positive definite even under a full likelihood approach, this is only an issue in small samples since in large samples, the nearest positive definite matrix will be the estimated $\boldsymbol{D}$ matrix itself, since the estimates of the pairwise-fitting approach are consistent (Fieuws and Verbeke, 2006).

3.4 Model Implementation

We used version 9.4 of the SAS System for Windows (SAS Institute, Cary NC). The bivariate models were fitted using the SAS-procedure NLMIXED. In this procedure, we only need to specify the conditional distributions of the responses given the random effects. Essentially, we can distinguish between two types of bivariate models.

The first type is a bivariate model with two longitudinal outcomes. In that case, the log-likelihood contribution of subject i to the bivariate joint model for outcomes l and k is:

$$l_i(\boldsymbol{\Theta}^*|\boldsymbol{y_{li}}, \boldsymbol{y_{ki}}) = \log\left(\int f_{li}(\boldsymbol{y_{li}}|\boldsymbol{b_i}) f_{ki}(\boldsymbol{y_{ki}}|\boldsymbol{b_i}) f(\boldsymbol{b_i}) d\boldsymbol{b_i}\right),$$

where $\boldsymbol{\Theta}^*$ represents the full parameter vector and $\boldsymbol{b_i}$ denotes the vector of all random effects.

The second type is a bivariate model with the survival outcome and one longitudinal outcome. By assuming that the correlation between the random effects accounts for the association between the two outcomes, i.e., assuming that the two outcomes are independent conditionally on the random effects, we can define the log-likelihood contribution of subject i to the bivariate joint model for the survival outcome and longitudinal outcome l as

$$l_i(\boldsymbol{\Theta}^*|t_i, \boldsymbol{y_{li}}) = \log\left(\int f(t_i|\boldsymbol{d_i})^{\delta_i} S(t_i|\boldsymbol{d_i})^{(1-\delta_i)} f_{li}(\boldsymbol{y_{li}}|\boldsymbol{d_i}) f(\boldsymbol{d_i}) d\boldsymbol{d_i}\right),$$

with $\boldsymbol{d_i}$ again denoting the vector of all random effects, now including the random effect for the survival outcome and the random effects of the longitudinal outcome.

Integration and optimization was carried out using (adaptive) Gaussian quadrature with 21 quadrature points.

Although the NLMIXED procedure only supports normally distributed random effects, this restriction can be bypassed using the probability integral transformation (PIT) method proposed by Nelson et al. (2006). Specifically for a gamma distributed random effects, assuming that u_i is a random effect following a standard normal distribution, that is $u_i \sim N(0,1)$, a gamma distributed random variable a_i can be derived from u_i through the following transformation: $a_i = F_a^{-1}[(\Phi(u_i)]$, where F_a is the cumulative distribution function (CDF) of a_i, in this case the CDF of a gamma distribution, and $\Phi(.)$ is the standard normal CDF.

4 Applications

4.1 IPF Data

In this section, we discuss the results of applying the pairwise-fitting approach on the IPF dataset described in Section 2.1. Following Sun et al. (2019), a Weibull PH model is assumed:

$$h_i(t) = a_i \lambda \rho t^{\rho-1} \exp(\alpha_1 age_i + \alpha_2 gender_i + \alpha_3 FVC_i),$$

where we assume a log-normal distribution for the random effect a_i. For the longitudinal outcomes, we assume the following linear mixed models:

$$Y_{ijl} = \beta_{0l} + b_{il} + \beta_{1l} year_{ij} + \beta_{2l} year_{ij}^2 + \epsilon_{ijl},$$

where Y_{ijl} denotes the outcome for the lth gene ($l = 1, ..., 20$) measured for patient i at timepoint j. The random effect b_{il} can be interpreted as the average deviation of patient i from the population average for outcome l. To model the trajectories over time, a quadratic term was included in the model as suggested by Sun et al. (2019). The full correlation matrix derived from the variance-covariance matrix $\boldsymbol{D}$ can be found in Table 1 in the Supplemental Materials. No adjustment for a non-positive definite matrix was necessary, as the estimated variance-covariance matrix $\boldsymbol{D}$ was already positive definite.

Since we are mainly interested in the correlations between the survival random effect and the random intercepts of the longitudinal outcomes, these correlations are represented in Table 1. The highest correlation between survival time and the random intercepts was observed for MMP8 (0.61), followed by CEACAM8 (0.50) and OLFM4 (0.45). This suggests that a patient with an increased expression value of these genes in the blood might have an increased risk of death. The highest negative correlation could be observed for SNORA14A (-0.39), followed by GSTM1 (-0.31) and DEFA4 (-0.24), indicating that an increased expression value of these genes might be associated with a decreased risk of death. Standard errors using the Delta method were calculated for the correlations and ranged from 0.0181 to 0.9734. These standard errors along with the corresponding 95% confidence intervals are presented in Table 2 in the Supplemental Materials. Notably, the correlations between survival and DEFA4 (95% CI: [-0.5420, -0.0086]), HBZ_1 (95% CI: [0.0105, 0.4552]), OLFM4 (95% CI: [0.0874, 0.8028]), and CEACAM8

(95% CI: [0.2473, 0.7456]) were significant, as their confidence intervals did not include zero.

We also tried to fit a shared-parameter joint model by using the model implemented in the **joineRML** package (Hickey et al., 2018). However, the model did not converge because of the large number of longitudinal outcomes. This dataset has been previously analysed by Sun et al. (2019). In this study, latent classes are used with a primary focus on predicting the time to death rather than studying the association between the longitudinal outcomes and the survival outcome. As stated in their paper, it is not straightforward to assess the association between the longitudinal outcomes and the survival outcome in the latent class model. Moreover, when the number of longitudinal outcomes becomes large, they propose modelling the longitudinal outcomes independently from each other, while in a joint model, all outcomes should be modelled jointly.

Table 1: Correlations between the random intercepts and survival random effect for the IPF study

	EIF1AY	RPS4Y1	HBZ	TUBB2A	TUBB2A_1	SLC12A1	DDX3Y	DEFA4	UTY	HBZ_1
a_i	0.03	0.02	0.06	0.25	0.25	-0.11	-0.02	-0.28	-0.00	0.23
	DAAM2	TXLNGY	MMP8	OLFM4	TMEM176A	OLAH	SNORA14A	TUBB2A_2	GSTM1	CEACAM8
a_i	0.32	-0.02	0.61	0.45	-0.17	0.31	-0.39	0.03	-0.31	0.50

4.2 PBC Data

We have applied the pairwise-fitting approach on the pbc2 dataset introduced in

Section 2.2. For the longitudinal outcomes, we assume the following linear mixed models:

$$Y_{ij1} = (0.1 * prothrombin_{ij})^{-4} = \beta_{01} + b_{i1} + \beta_{11} year_{ij} + b_{i2} * year_{ij} + \epsilon_{ij1},$$

$$Y_{ij2} = albumin_{ij} = \beta_{02} + b_{i3} + \beta_{12} year_{ij} + b_{i4} * year_{ij} + \epsilon_{ij2},$$

$$Y_{ij3} = log(serBilir_{ij}) = \beta_{03} + b_{i5} + \beta_{13} year_{ij} + b_{i6} * year_{ij} + \epsilon_{ij3},$$

in which b_{i1}, b_{i3}, b_{i5} are the random intercepts, b_{i2}, b_{i4}, b_{i6} are the random slopes and ϵ_{ijl} are the error terms that are assumed to be independent and normally distributed with variance σ_l^2. The covariate $year_{ij}$ represents the number of years between the baseline visit (enrollment) and the date of visit j. For the survival outcome, the following Weibull PH frailty model is assumed:

$$h_i(t) = a_i \lambda \rho t^{\rho-1} \exp(\alpha_1 age_i),$$

where age_i is the age at enrollment. For the subject-specific random effect, a_i, either a log-normal or a gamma distribution is assumed. Since the estimated variance-covariance matrix $\boldsymbol{D}$ was non-positive definite, we applied the adjustment described in Section 3.3. The correlations between the random effects and survival random effect, derived from the adjusted $\boldsymbol{D}$, when assuming a log-normal distribution for a_i, can be found in Table 2. The full correlation matrix can be found in Table 3 in the Supplemental Materials. The correlation between the random effect for the survival outcome and the random intercept for $(0.1*\text{prothrombin})^{-4}$ is negative, which may indicate that a patient with a high prothrombin level at baseline has an increased risk of death. Similarly, there is a negative correlation between the random slope and the survival random effect, which may indicate that

a patient whose prothrombin level increases faster on average has an increased risk of death. We observe a high negative correlation between the random effect of the survival outcome and the random intercept for albumin. This suggests that a patient with higher values for albumin at baseline has a decreased risk of death. The negative correlation between the random slope of albumin and the survival random effect indicates that a patient with a steeper increase in albumin level over time has a decreased risk of death. Lastly, positive correlations were found between the random effect for the survival outcome and the random intercept and slope for log(serum bilirubin). This indicates that patients with a higher level of serum bilirubin at baseline or patients whose serum bilirubin level increases faster over time have an increased risk of death. All estimated parameters, together with the robust standard errors, can be found in Table 4 in the Supplemental Materials.

Table 2: Correlations between the random effects and survival random effect after adjustment for non-positive definiteness for the PBC study (log-normal distribution for a_i)

	a_i	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}
a_i	1.0000	-0.6652	-0.2461	-0.9009	-0.5927	0.3684	0.1282

We also fitted the model assuming a one-parameter gamma distribution for a_i. The results can be found in Tables 5 and 6 in the Supplemental Materials. The results are similar to the results of the model assuming a log-normal distribution for the survival random effect, and the conclusions are coherent between these versions. The results obtained using our approach are consistent with those reported in pre-

vious studies that have analysed the PBC dataset using standard joint modelling methods (e.g. Hickey et al. (2018).

5 Simulation Studies

5.1 High-dimensional longitudinal outcomes

The most important advantage of our proposed methodology is that it allows to jointly model as many outcomes as one would like, as long as the bivariate models for all possible pairs can be fitted. The number of pairs is, of course, a quadratic function of the number of outcomes. This means that computation time will essentially increase quadratically if the pairs are fitted sequentially. If a sufficient number of cores is available, all sets of pairwise models can be fitted in parallel, potentially further decreasing the computation time.

To illustrate this, a small simulation study was conducted. The performance of the estimators of the pairwise-fitting approach has been studied previously across different simulation settings by Fieuws and Verbeke (2005). In Section 5.2, a more extensive simulation study has been carried out to compare the pairwise-fitting approach with a full likelihood approach. The purpose of this simulation study was to demonstrate how the methodology can be used to model high-dimensional multivariate longitudinal outcomes jointly with a survival outcome. We simulated 500 datasets, each including $N = 1000$ subjects. Ten longitudinal outcomes were

simulated based on linear mixed models with a random intercept:

$$Y_{ijl} = \beta_{0l} + b_{il} + \beta_{1l} t_{ij} + \epsilon_{ijl}, \tag{5.1}$$

with Y_{ijl} denoting the lth ($l = 1, 2, \ldots, 9, 10$) longitudinal outcome measured for subject i at timepoint j ($t_{ij} = 1, 2, \ldots, 9, 10$) and with ϵ_{ijl} simulated from a normal distribution with zero mean and variance σ_l^2. The parameters were considered to be outcome-specific. For the survival outcome, we simulated survival times from the following Weibull PH frailty model:

$$h_i(t) = a_i \lambda \rho t^{\rho - 1} \exp(\alpha x_i), \tag{5.2}$$

with x_i a subject-specific covariate of which the values were sampled from a normal distribution with mean 40 and variance 2. When the simulated event time t_i exceeded 5, subject i was administratively censored at time 5, such that 10% of the subjects were censored. The frailties a_i were sampled from a log-normal distribution. All true parameter values can be found in Tables 7 and 8 in the Supplemental Materials for the longitudinal outcomes and for the Weibull PH frailty model, respectively. The matrix $\boldsymbol{D}$ that was used in the simulation study can be found in Table 9 in the Supplemental Materials. High positive correlations between the random effects were specified.

We applied the proposed methodology to each of the 500 datasets and summarized the results by calculating the mean of the estimates and standard errors obtained under this approach. The estimated correlation matrix between the random effects can be found in Table 10 in the Supplemental Materials. The estimated correla-

tions are very close to the true correlations, and no computational problems were encountered when analysing all of the 500 datasets. The variance-covariance matrix $\boldsymbol{D}$ can be found in Table 11 in the Supplemental Materials, and all estimated parameters together with the robust standard errors can be found in Table 12 and Table 13 of the linear mixed models and the Weibull PH frailty model respectively in the Supplemental Materials. The estimated values of these parameters closely align with the simulation parameters.

We also assessed the runtime required to fit the model. For this purpose, we used the first simulated dataset and ran the SAS program twenty times. We did this for an increasing number of longitudinal outcomes, modelled jointly with one survival outcome, starting with one longitudinal outcome up to ten longitudinal outcomes. We conducted the analyses on a HP ZBook laptop featuring an Intel Core i7 processor and 32.0 GB of RAM. The laptop was running SAS version 9.4 and Windows 11 Enterprise as the operating system. Pairs were fitted sequentially without using parallel processing. The average amount of runtime of the twenty times the SAS program was run for each scenario is represented in Figure 1. Additionally, we also fitted the models with a full likelihood approach, meaning that all longitudinal outcomes and the survival outcome are jointly fitted in one model instead of fitting all possible pairs. Note that when only one longitudinal outcome is modelled jointly with the survival outcome, the pairwise-fitting approach reduces to a full likelihood approach since only one pair needs to be fitted. With a full likelihood approach, the models failed to converge when more than two longitudinal

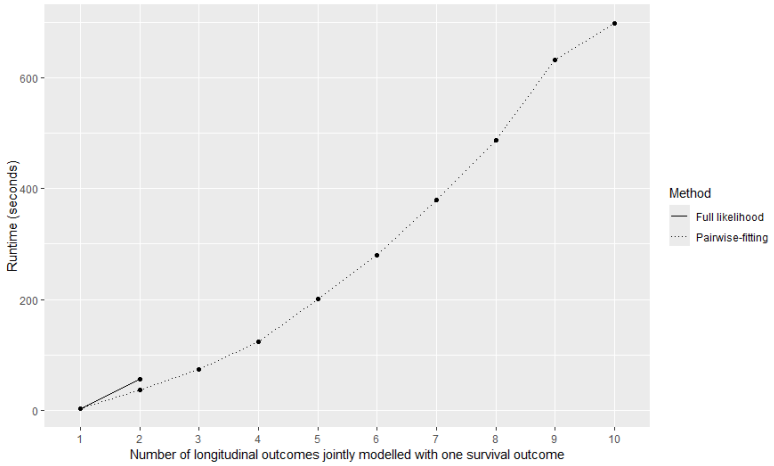

Figure 1: Amount of runtime that was required for fitting the model with a full likelihood approach (solid) or pairwise-fitting approach (dotted) for up to 10 longitudinal outcomes.

outcomes are considered. To further illustrate that the pairwise-fitting approach is not limited by the number of outcomes, we further increased the number of longitudinal outcomes to 20, 30, 40, and 50, and successfully fitted the model (Figure 2). The figure illustrates that computation time increases quadratically as the number of pairs increases. However, with parallel processing, the increase in computation time would be less.

5.2 Comparison with Full Likelihood Approach

In a second simulation study, the goal was to compare the estimates of the full likelihood approach with the estimates obtained from using the pairwise-fitting approach. To this end, we have simulated 1000 datasets, each containing two

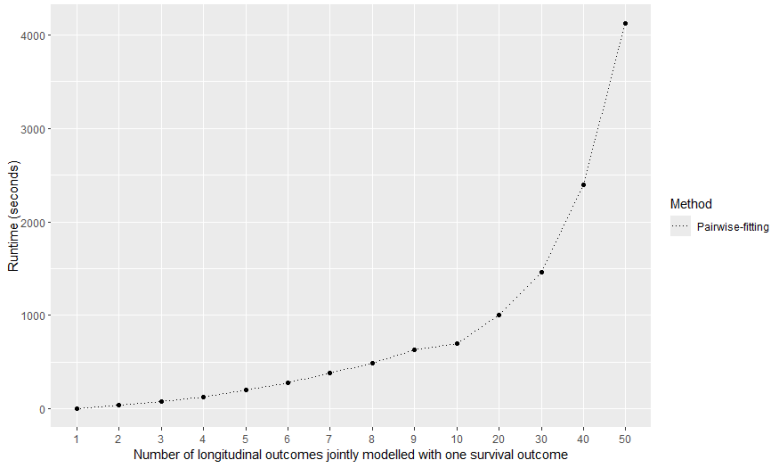

Figure 2: Amount of runtime that was required for fitting the model with a pairwise-fitting approach for up to 50 longitudinal outcomes.

longitudinal outcomes and one survival outcome. We have only considered two longitudinal outcomes since this setting can still be analysed with a full likelihood approach. Each dataset consisted out of $N = 1000$ subjects. The time-to-event data was simulated based on the same model as in (5.2), with the difference that no subjects were censored. For the longitudinal outcomes, we used the same random-intercepts model as in (5.1). We analysed these datasets using both the full likelihood method and our proposed pairwise-fitting approach. For each approach, we then calculated the average of all 1000 estimates. The bias was obtained by subtracting this average from the true mean. Additionally, we have also calculated the root mean squared error (RMSE), which can be used to assess the accuracy of the estimates, and the coverage of 95% confidence intervals (Table 14 in the Supplemental Materials). In general, the two approaches appear to be similar in terms of bias, RMSE and coverage.

After having conducted the simulation study without any censoring in the survival outcome, we conducted a subsequent simulation study where censoring was induced. When the simulated event time t_i exceeded 4, subject i was administratively censored at time 4, such that 16% of the subjects were censored. When a subject was censored, we simulated dropout in their two longitudinal outcomes by marking the values after the 5th timepoint as missing. Consequently, these subjects were only observed at timepoints $t = 1, 2, 3, 4,$ and 5.

The results are presented in Table 15 in the Supplemental Materials. We noticed that the bias in some parameters using the full likelihood approach is smaller compared to using the pairwise-fitting approach. Comparing the biases of the pairwise-fitting approach with censoring to the setting without censoring, it appears that there is a little more bias when censoring is present. However, the difference does not seem to be substantial, certainly in light of the significant computational advantage the pairwise-fitting approach offers, namely enabling efficient estimation of joint models without any limitation on the number of longitudinal outcomes. As stated before in Section 3.2.1, the computational advantages were also the argument of Geys et al. (1999) to justify the minor efficiency loss when using pseudo-likelihood methods. Indeed, while the pairwise-fitting approach may result in some efficiency loss compared to a full likelihood estimation, this compromise is justified. In situations where the number of outcomes becomes very large, full likelihood methods can no longer be used. Our approach provides a scalable solution that remains feasible in these cases where full likelihood methods cannot

be applied. Additionally, none of the estimated variance-covariance matrices $\boldsymbol{D}$ in both simulation studies were non-positive definite, which also demonstrates that this is a small sample issue and that the estimates of the pairwise-fitting approach are consistent.

6 Discussion

In this paper, we have proposed a new joint model for high-dimensional multivariate longitudinal data and survival data, which is based on correlated random effects. The model is fitted using a pairwise modelling approach. The correlation between the random effects for the longitudinal outcomes and the random effect for the survival outcome provides us insight into the association structure. While the standard shared parameter joint models suffer from computational problems when many longitudinal outcomes are considered, our model does not and it allows researchers to fit joint models without suffering from a restriction on the number of outcomes, as long as the bivariate models can be fitted. Moreover, our approach not only accommodates multiple longitudinal outcomes but can be naturally extended to incorporate multiple survival outcomes as well.

We also used our model to analyse data of patients with idiopathic pulmonary fibrosis. In this case study, we studied the association between the expression profiles of 20 genes and risk of death. While a shared-parameter joint model was no longer applicable, our model could still be used. To illustrate further that our

model can be fitted irrespective of the number of outcomes, we simulated datasets containing up to 50 longitudinal outcomes and a survival outcome, and analysed it with the proposed joint model. Moreover, by means of two simulation studies — one without censoring and one with censoring — we compared the performance of our approach to that of a full likelihood approach. While the likelihood-based method is, as expected, generally more efficient, we demonstrate that the bias in our model is relatively small. Importantly, in cases with many outcomes, where a full likelihood approach is no longer feasible, our approach remains very applicable.

The proposed model can be extended with ease. For instance, outcomes of a different type can be considered, such as count or binary longitudinal data. These outcomes can be modelled with generalized linear mixed models and can be incorporated in the model by allowing for correlation between the random effects. For the survival model, a model for recurrent events or for competing risks could also be incorporated. Previous work by Molenberghs et al. (2010) and Molenberghs et al. (2015) has explored the use of Weibull models with gamma and normal random effects, referred to as the combined model. In this framework, gamma random effects account for overdispersion, while normal random effects represent the frailty term. This extension can also be incorporated into our approach.

Nevertheless, a possible limitation of our model is that it can only be used if all possible pairs can be fitted. This might be challenging with standard software when more complex random effects structures are assumed, for instance when polynomials or splines are used as random effects, and when one would like to

allow for correlation between all these random effects of one outcome and all these random effects of another outcome. However, we argue that, in practice, most of the correlation between the longitudinal outcomes and the survival outcome can be captured by allowing the random intercepts and slopes to be correlated with the survival random effect. Then, one can still use polynomials or splines in the random effects structure, provided that these random effects remain uncorrelated with those of the other outcomes. It should also be noted that this does not prohibit the use of polynomials and splines in the fixed effects structure.

Supplementary materials

The supplementary materials include

- File with additional Tables (Additional Tables.pdf).

They are available from the journals page at https://journals.sagepub.com/home/smj.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

References

Aerts, M., Molenberghs, G., Ryan, L. M., and Geys, H. (2002). *Topics in modelling of clustered data*. Chapman and Hall/CRC.

Albert, P. S. and Shih, J. H. (2010). An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. *The annals of applied statistics*, **4**(3), 1517.

Arnold, B. C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 233–243.

Collett, D. (2023). *Modelling survival data in medical research*. CRC press.

Crowther, M. J., Abrams, K. R., and Lambert, P. C. (2013). Joint modeling of longitudinal and survival data. *The Stata Journal*, **13**(1), 165–184.

Duchateau, L. and Janssen, P. (2008). *The frailty model*. Springer.

Elbers, C. and Ridder, G. (1982). True and spurious duration dependence: The identifiability of the proportional hazard model. *The Review of Economic Studies*, **49**(3), 403–409.

Feddag, M.-L. and Bacci, S. (2009). Pairwise likelihood for the longitudinal mixed rasch model. *Computational statistics & data analysis*, **53**(4), 1027–1037.

Fieuws, S. and Verbeke, G. (2004). Joint modelling of multivariate longitudinal profiles: pitfalls of the random-effects approach. *Statistics in medicine*, **23**(20), 3093–3104.

Fieuws, S. and Verbeke, G. (2005). Evaluation of the pairwise approach for fitting joint linear mixed models: A simulation study. Technical report, Technical Report TR0527, Biostatistical Centre, Katholieke Universiteit ....

Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, **62**(2), 424–431.

Fieuws, S., Verbeke, G., Boen, F., and Delecluse, C. (2006). High dimensional multivariate mixed models for binary questionnaire data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, **55**(4), 449–460.

Fieuws, S., Verbeke, G., and Molenberghs, G. (2007). Random-effects models for multivariate repeated measures. *Statistical methods in medical research*, **16**(5), 387–397.

Geys, H., Molenberghs, G., and Ryan, L. M. (1997). Pseudo-likelihood inference for clustered binary data. *Communications in Statistics-Theory and Methods*, **26**(11), 2743–2767.

Geys, H., Molenberghs, G., and Ryan, L. M. (1999). Pseudolikelihood modeling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, **94**(447), 734–745.

Hermans, L., Nassiri, V., Molenberghs, G., Kenward, M. G., Van der Elst, W., Aerts, M., and Verbeke, G. (2018). Clusters with unequal size: Maximum likelihood versus weighted estimation in large samples. *Statistica Sinica*, **28**(3), 1107–1132.

Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2018). joinerml: a joint model and software package for time-to-event and multivariate longitudinal outcomes. *BMC medical research methodology*, **18**, 1–14.

Huang, X., Li, G., Elashoff, R. M., and Pan, J. (2011). A general joint model for longitudinal measurements and competing risks survival data with heterogeneous random effects. *Lifetime data analysis*, **17**, 80–100.

Ivanova, A., Molenberghs, G., and Verbeke, G. (2016). Mixed models approaches for joint modeling of different types of responses. *Journal of Biopharmaceutical Statistics*, **26**(4), 601–618.

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

Lawrence Gould, A., Boye, M. E., Crowther, M. J., Ibrahim, J. G., Quartey, G., Micallef, S., and Bois, F. Y. (2015). Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the dia bayesian joint modeling working group. *Statistics in medicine*, **34**(14), 2181–2195.

Li, N., Elashoff, R. M., and Li, G. (2009). Robust joint modeling of longitudinal measurements and competing risks failure time data. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, **51**(1), 19–30.

Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Gelber, R., and Lipshultz, S. (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics*, **58**(3), 621–630.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Mauff, K., Steyerberg, E., Kardys, I., Boersma, E., and Rizopoulos, D. (2020). Joint models with multiple longitudinal outcomes and a time-to-event outcome: a corrected two-stage approach. *Statistics and Computing*, **30**, 999–1014.

Mauff, K., Erler, N. S., Kardys, I., and Rizopoulos, D. (2021). Pairwise estimation of multivariate longitudinal outcomes in a bayesian setting with extensions to the joint model. *Statistical Modelling*, **21**(1-2), 115–136.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. ISBN 978-0-387-25144-8. doi:10.1007/0-387-28980-1.

Molenberghs, G., Verbeke, G., Demétrio, C. G., and Vieira, A. M. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, **25**(3), 325–347.

Molenberghs, G., Kenward, M. G., Verbeke, G., and Birhanu, T. (2011). Pseudo-likelihood estimation for incomplete data. *Statistica Sinica*, pages 187–206.

Molenberghs, G., Verbeke, G., Efendi, A., Braekers, R., and Demétrio, C. G. (2015). A combined gamma frailty and normal random-effects model for repeated, overdispersed time-to-event data. *Statistical Methods in Medical Research*, **24**(4), 434–452.

Molyneaux, P. L., Willis-Owen, S. A., Cox, M. J., James, P., Cowman, S., Loebinger, M., Blanchard, A., Edwards, L. M., Stock, C., Daccord, C., et al. (2017).

Host–microbial interactions in idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine*, **195**(12), 1640–1650.

Murtaugh, P. A., Dickson, E. R., Van Dam, G. M., Malinchoc, M., Grambsch, P. M., Langworthy, A. L., and Gips, C. H. (1994). Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits. *Hepatology*, **20**(1), 126–134.

Nelson, K. P., Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J., Parzen, M., and Strawderman, R. (2006). Use of the probability integral transformation to fit nonlinear mixed-effects models with nonnormal random effects. *Journal of Computational and Graphical Statistics*, **15**(1), 39–57.

Nguyen, H. T., Vasconcellos, H. D., Keck, K., Reis, J. P., Lewis, C. E., Sidney, S., Lloyd-Jones, D. M., Schreiner, P. J., Guallar, E., Wu, C. O., et al. (2023). Multivariate longitudinal data for survival analysis of cardiovascular event prediction in young adults: insights from a comparative explainable study. *BMC medical research methodology*, **23**(1), 23.

Papageorgiou, G., Mauff, K., Tomer, A., and Rizopoulos, D. (2019). An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual review of statistics and its application*, **6**, 223–240.

Renard, D., Molenberghs, G., and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics & Data Analysis*, **44**(4), 649–667.

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

Rizopoulos, D. (2016). The R package JMbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *Journal of Statistical Software*, **72** (7), 1–45. doi:10.18637/jss.v072.i07.

Rousseeuw, P. J. and Molenberghs, G. (1993). Transformation of non positive semidefinite correlation matrices. *Communications in Statistics–Theory and Methods*, **22**(4), 965–984.

Schwartz, D., Helmers, R., Galvin, J., Van Fossen, D., Frees, K., Dayton, C., Burmeister, L., and Hunninghake, G. (1994). Determinants of survival in idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine*, **149**(2), 450–454.

Sun, J., Herazo-Maya, J. D., Molyneaux, P. L., Maher, T. M., Kaminski, N., and Zhao, H. (2019). Regularized latent class model for joint analysis of high-dimensional longitudinal biomarkers and a time-to-event outcome. *Biometrics*, **75**(1), 69–77.

Therneau, T. and Grambsch, P. (2013). *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer New York. ISBN 9781475732948. URL https://books.google.be/books?id=ojOmBQAAQBAJ.

Tibaldi, F., Abrahantes, J. C., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T., and Wolfinger, R. (2003). Simplified hi-

erarchical linear models for the evaluation of surrogate endpoints. *Journal of Statistical Computation and Simulation*, **73**(9), 643–658.

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

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42.

Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer.

Verbeke, G., Fieuws, S., Molenberghs, G., and Davidian, M. (2014). The analysis of multivariate longitudinal data: a review. *Statistical methods in medical research*, **23**(1), 42–59.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, pages 175–188.