Model Specification and Prediction in Joint Modelling

Jacob Cancino-Romero

School of Mathematics

University of Leeds

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

July 2020

Acknowledgements

I want to thank the Leeds Annual Research Scholarship (LARS) for three years of funding and School of Mathematics of the University of Leeds for the six months extension.

My thanks to my supervisors Prof. Jeanine Houwing-Duistermaat, Dr Stuart Barber and Dr Leonid V. Bogachev for their guidance and support in every stage of the PhD. I thank Dr Peter A. Thwaites and Dr Robert G. Aykroyd for their valuable feedback in the annual reviews and follow-up one-to-one meetings, Dr Andrew Clegg for providing data of scientific interest (Community Ageing Research study – CARE75+), and Dr Agnieszka Król, Prof. Dimitris Rizopoulos and Prof. Virginie Rondeau for kindly replying to my inquiries about their research.

I want to thank Gillian A. Hawker, David E. Matthews, and Roberto Salcedo-Aquino for their encouragement and for providing reference letters.

The CARE75+ study is funded by the NIHR CLAHRC Yorkshire and Humber - www.clahrc-yh.nihr.ac.uk. The views expressed are those of the author(s), and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Abstract

This thesis explores several methodological aspects of joint modelling of longitudinal outcomes and recurrent and terminal events, including variable selection, description, prediction, causal inference and model specification. The methods we discuss were motivated by the Community Ageing Research study (CARE75+) to investigate the relationships between frailty, falls and mortality. These outcomes have previously been analyzed with marginal models, but not as joint outcomes.

We propose a variable selection strategy to optimize prediction of joint models for longitudinal and time-to-event outcomes. This strategy combines penalized likelihood with the LASSO penalty and cross-validation methods to select the fixed effects that optimize simultaneously the mean-squared error (MSE) and the Integrated Brier Score (IBS). Our simulation studies suggest that it is not always possible to optimize simultaneously MSE and IBS, but there seems to be a region defined by the constraints close to an optimal solution. In such a case a small compromise between MSE and IBS is required, depending on which outcome is the priority.

Joint modelling has been an area of active research for description and prediction, but causal inference has received less attention. Using Direct Acyclic Graphs, we state our hypotheses about the paths between frailty, falls and mortality and confounders to formulate joint models adjusting for confounders. Via simulation studies we assessed the consequences of model misspecification, finding that even when link of the joint model and some features of the mean structure are not the correct ones, the fixed effects can still be correctly estimated.

Abbreviations

t	Time as a continuum
t_{ij}	The j^{th} observation time point specific to subject i
$m_i(t)$	True and unobservable longitudinal outcome of subject i at time t
$y_i(t)$	Observable measure of the longitudinal outcome of subject i at time t
y_{ij}	$y_i(t_{ij})$: Longitudinal outcome of subject i observed at time point t_{ij}
n_i	Number of repeated measures of the longitudinal outcome of subject i
n	Number of individuals in a sample
T_i^*	Terminal event time of subject i
δ_i	$\mathbb{1}(T_i^* \leq C_i)$ Event indicator of subject i
T_{ik}^*	Time of the kth event of subject i
δ_{ik}	k^{th} recurrent event indicator of subject i
$\mathbf{b}_i$	q-column vector of random effects of subject i in linear-mixed model
u_i	Random effect of subject i in survival analysis model
$N_i(t)$	Right-continuous counting process of the number of events
$r_0(t)$	Baseline hazard rate of the recurrent event
$r_i(t)$	Hazard rate of the recurrent event process of subject i
$h_0(t)$	Baseline hazard rate of the terminal event
$h_i(t)$	Hazard rate of the terminal event of subject i
$\mathbf{w}_i$	p-vector of baseline covariates of subject i
$\mathbf{x}_i(t)$	p-vector of fixed effects covariates (possibly time-varying) of subject i
$\mathbf{z}_i(t)$	q-vector of random effects covariates (possibly time-varying) of subject i
X_i	$n_i \times p$ design matrix of fixed effects covariates
Z_i	$n_i \times p$ design matrix of random effects covariates
$\boldsymbol{\beta}$	p-vector of fixed effects regression coefficients of a linear-mixed model
$\boldsymbol{\gamma}$	p-vector of fixed effects regression coefficients of a survival analysis model

Contents

List of Figures

Chapter 1

Introduction

In follow-up studies, usually different types of outcomes are collected for each sample unit, which may include multiple longitudinal repeated measures and the time until an event of particular interest occurs. The research questions of interest are typically formulated for separate analyses of the recorded outcomes. Nonetheless, sometimes it is more appropriate to model them jointly as separate models do not fully account for the structure in the data and may lead to incorrect conclusions due to biased estimates. This thesis is about joint modelling of longitudinal and time-to-event data.

The joint modelling methodology was developed to address problems in two different areas of statistics. On the one hand, in the area of longitudinal data analysis, joint models were originally developed to accommodate nonignorable missing data of a longitudinal response; this is when the probability of missingness is related to the missing, unobserved values (Sections 2.1.3 and 3.1.4). Disregarding the missing data from statistical analyses when missingness is nonignorable produces biased estimates. Jointly modelling the longitudinal response and the missing data process, introducing shared random effects between the two processes, is a strategy to accommodate the bias due to nonignorable missing data. On the other hand, in the area of survival analysis joint models have been studied in the Cox model with endogenous time-varying covariates. Endogenous covariates are directly measured from the individuals so they require individuals survival for their existence. When the terminal event is death, for instance, the observed history of endogenous covariates up to time t implies survival

up to this time, and the survival function given covariates does not have the usual interpretation. Hence with endogenous time-varying covariates the log-likelihood function ($\delta_i \log f(t_i;\boldsymbol{\theta}) + (1-\delta_i)\log S_i(t_i;\boldsymbol{\theta})$) is not valid (Sections 2.2.1 and 3.1). These challenges can be overcome by joint modelling the time-to-event outcome and the time-varying covariate trajectory.

A third situation is that joint modelling longitudinal and time-to-event data allows to explore the joint distribution of outcomes of different types and, in particular, to understand their association. For example, Ibrahim *et al.* (2010) discuss the importance of jointly modelling quality of life (QOL) and mortality in cancer patients. One might argue that, for a patient, improvement in QOL is often more important than any modest survival benefit in treatment decisions. Therefore, it is of great interest in cancer clinical trials to characterize the association between time-to-event and QOL through joint modelling and to understand the tradeoffs between QOL and survival. A specific treatment protocol with chemotherapy and radiotherapy may prolong survival or relapse, but the QOL in that prolonged period may be poor, and thus the clinician must decide whether such a benefit is worth it for the patient.

Different approaches have been proposed for the statistical analysis of joint models for longitudinal and time-to-event data, which can be grouped in likelihood maximization and Bayesian methods. The former can be grouped in Shared Random Effects Joint Models (SREJM) and Latent Class Joint Models (LCJM). The common features of these two alternatives are that for each outcome a regression submodel is specified (a linear mixed model for the longitudinal outcome and a survival analysis for the time-to-event) and a latent structure characterizes their association. And the difference between these two joint modelling alternatives is how they define the latent structure of their association. In a SREJM it is characterized by link functions of the random effects which are introduced as explanatory variables in the survival analysis submodel. In the LCJM the population is assumed to be heterogeneous but comprised of subpopulations with different patterns of change for the longitudinal outcome and different time-to-event risk profiles, so the association between outcomes is characterized by a third submodel to estimate the likelihood of the subjects for belonging to these subpopulations. In this thesis we explore with the Shared Random Effects Joint Model since it is the approach most widely used.

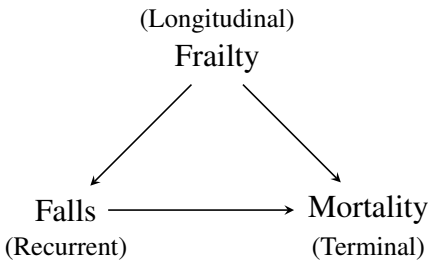


Figure 1.1: Possible relationship between frailty, falls and mortality, with frailty as common cause of falls and mortality, and falls as cause of mortality.

The joint modelling framework is not restricted to these specific combinations of outcomes and can be extended to more than two. In addition to exploring the relationship between longitudinal and terminal event outcomes, sometimes interest lies in studying their association with a recurrent event process. This three-outcome joint model is the cornerstone of the methodological aspects of joint modelling addressed in this thesis which are motivated by the Community Ageing Research (CARE75+) study, conducted in Northern England since 2015. The outcomes of interests are *frailty* (longitudinal outcome), *falls* (recurrent event) and *mortality* (terminal event). Figure 1.1 is an instance of the possible relationships between frailty, falls and mortality.

Frailty is a dynamic process of a reduction in the physical, psychological and social function associated with aging. It describes how the body gradually loses its built-in reserves, leaving it vulnerable to dramatic and sudden changes in health triggered by apparently minor illnesses, such as a chest infection, that otherwise the body could likely overcome. Frailty is associated with adverse outcomes such as frequent falls, disability, hospitalization, and mortality (Clegg *et al.*, 2013). The World Health Organization defines a *fall* as an event which results in a person coming to rest inadvertently on the ground, the floor, or other lower level. Fall-related injuries may be fatal or non-fatal though most are non-fatal. The CARE75+ study is a longitudinal cohort of older people with frailty for observational research. This study aims to understand why some people remain fit and resilient in older age while others develop health problems and frailty and to determine what (treatable) problems have a major impact on the quality of life in older age.

Frailty as the geriatric syndrome described above, like certain concepts in the social

and behavioural sciences, are not directly observable and their meaning and universally accepted definition remain subject of discussions, such as social class, intelligence, anxiety, depression, extrovert personality, utility, etc. Such concepts are referred to as *latent variables*. Latent variables are *hypothetical constructs* invented by scientists for the purpose of understanding some research area of interest and for which there are no operational method for direct measurement (Bollen, 2002; Everett, 2013).

Latent variables are given different names in different disciplines, such as random effects, common factors and latent classes. They are used to represent phenomena such as "True" variables measured with error, hypothetical constructs, unobserved heterogeneity, missing data, counterfactuals or "potential outcomes" (Rabe-Hesketh & Skrondal, 2004, Chapter 1).

Although latent variables are not observable, some of their effects on measurable *manifest* variables are observable, and hence subject to study. By analyzing the observable effects it is possible to learn about latent variables. The diagram in Figure 1.2, adapted from Lavrakas (2008), illustrates the process of using empirical evidence to learn about latent variables. A *construct* or hypothetical construct has an exclusively epistemological status (Rabe-Hesketh & Skrondal, 2004). It is an intellectual device by means of which one *construes* events, i.e they are simply concepts. Because hypothetical constructs do not correspond to real phenomena, it follows that they cannot be measured directly even in principle. Instead, the construct is operationally defined in terms of a number of items or indirect "indicators". Thus the *operational definition* is supposed to bridge theoretical arguments (construct and conceptual definition) and the empirical evidence by mechanisms of concrete language that allows to gather information about the construct. Finally, *measuring* involves both creating a rule (scale) and making assignments of cases into categories based on this rule.

Possible errors in the described process can be grouped in two types: *construct validity* and *measurement error*. Construct validity is the degree that the operational definition captures the theoretical concept, and measurement error is any deviation of the assigned symbol from the "true" value that should be designated to the object. A measure is said to be valid (to have a strong construct validity) if it measures what it claims to measure. Two types of error affect measures validity: systematic error (bias) and random error (variance) (Lavrakas, 2008).

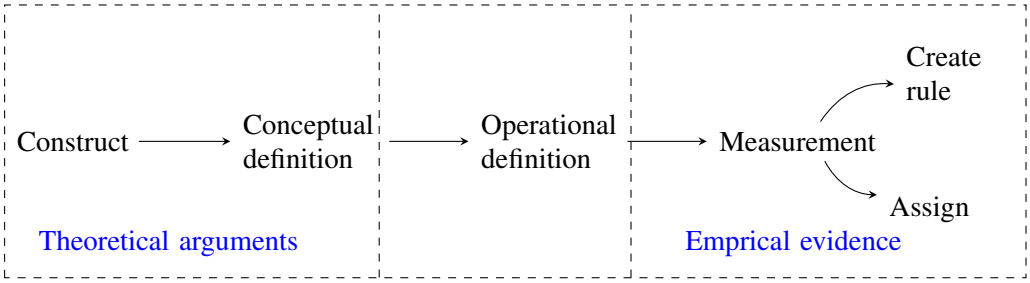


Figure 1.2: Process for creating a measuring instrument from theoretical arguments to empirical evidence.

Several measurements have been proposed to detect frailty in the elderly. Faller *et al.* (2019) provides a systematic review of the different instruments developed for the detection of frailty. With the CARE75+ study, researchers collect data to quantify frailty according to several instruments, for instance the Electronic Frailty Index (eFI) (Clegg *et al.*, 2016) and the Edmonton Frail Scale (EFS) (Rolfson *et al.*, 2006). In the joint modelling methods discussed in this thesis we used frailty in the EFS.

Frailty and falls have been subjects of research, and their relationship with mortality and other risk factors has been analyzed with marginal models as separate outcomes. In the CARE75+ study, the data to quantify frailty, summarized in the EFS, are collected intermittently on each participant at set times, hence frailty makes sense only as long as the participants are alive. In a time-to-event model for mortality, frailty acts as an endogenous time-varying covariate subject to measurement error. Joint modelling frailty and mortality accommodates endogenous time-varying covariates in the Cox model while accounting for the measurement error of frailty and the bias produced by ignoring the dependence between the two outcomes.

Throughout this thesis we explore and discuss different features of joint models for longitudinal outcomes and recurrent and terminal events, including variable selection, model specification, description, prediction and causal inference. The methods discussed are applied to the CARE75+ data set in different alternatives for joint modelling frailty, falls and mortality.

We consider that our work makes contributions in two areas: statistical methodology and applications. Our contributions to the statistical methodology are (1) proposing

a variable selection strategy for simultaneously optimizing prediction of the two outcomes in joint modelling of longitudinal and time-to-event data, (2) approaching the joint modelling framework from the causality perspective using DAGs, and (3) evaluating the consequences of misspecifying the mean and association structures of joint modelling a longitudinal outcome and recurrent and terminal events. In the applications context, geriatric frailty, recurrent falls and mortality have been analyzed before with marginal models, and we investigated their relationships with the joint modelling methodology.

This thesis has seven chapters. The present Introduction is Chapter 1. Chapter 2 is background material included for completeness and to establish the notation we use in the rest of the thesis. It explains the building blocks of joint modelling longitudinal and time-to-event data: (a) the linear mixed-effects model for the analysis of continuous longitudinal responses and (b) survival analysis models, emphasizing in the challenges that conveys data missing not at random and endogenous time-varying covariates that motivate the need for joint models. Additionally, we briefly describe other statistical topics relevant to the methods we address in subsequent chapters: prediction, penalized likelihood methods and causal inference.

Chapter 3 is an introduction to joint modelling, included to provide the methodological framework of our work of Chapters 4–6. It describes the two main joint models we explore in this thesis and some aspects of prediction in the context of joint modelling that we use in Chapter 5.

In Chapter 4, we analyze frailty, falls and mortality in the CARE75+ data set with the joint modelling methodology aiming to fit a joint model to describe the relationships among these three outcomes and covariates. The results of our statistical analyses of the CARE75+ study suggest that frailty is strongly associated with both falls and mortality. Specifically, the relative risks of falls and mortality are 1.551 and 1.537, respectively for subjects whose frailty is one unit above the population average, *ceteris paribus*. The effect of recurrent falls on mortality was not significant at the 5% level. It is possible for this relationship to be mediated by other adverse outcomes, like long-term institutional care, hospital admissions, injuries, fear of subsequent falls, reduced activities of daily living and lower quality of life (Masud & Morris, 2001), requiring a longer follow up period than is available in the analyzed data set to observe this effect.

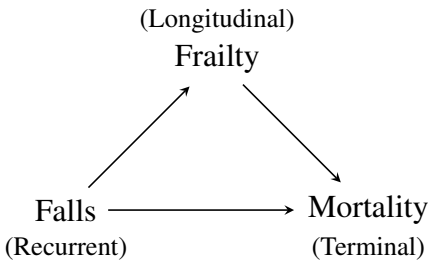


Figure 1.3: Possible relationship between frailty, falls and mortality, with falls as common cause of frailty and mortality, and frailty as cause of mortality.

We discuss the analyses and our conclusions in Chapter 4, and fitted an alternative plausible joint model for frailty, falls and mortality considering falls as a time-varying covariate for frailty and mortality, as illustrated in Figure 1.3.

The complicated structure of joint modelling conveys challenges for statistical modelling. In particular, variable selection is nontrivial and of paramount interest. In statistical modelling, variable selection is carried out in different ways depending on the intended use of the fitted model: *description*, *causal inference* or *prediction* (Shmueli *et al.*, 2010). In our statistical analyses to fit a joint model that describes the relationships between frailty, falls and mortality in the CARE75+ data set, variable selection was done by a stepwise procedure, first in the marginal model of each outcome and then in the joint model. This approach becomes more difficult with correlated covariates, in addition to the long processing times required to fit a joint model. Variable selection in joint modelling has been studied before using penalized likelihood methods aiming to optimize the goodness of fit (He *et al.*, 2015).

We propose a variable selection strategy that aims at optimizing prediction of a joint model for longitudinal and time-to-event outcomes. Prediction accuracy of this type of joint model has mainly focused on the time-to-event outcome in terms of the ability of the model to predict the subjects future event status, and the prospective accuracy of the longitudinal outcome to discriminate events from non-events using the ROC methodology (Rizopoulos, 2011; Zheng & Heagerty, 2007), or by computing the Brier score (BS) for a specific time-point (Król *et al.*, 2017). The focus of our variable selection strategy is on simultaneous optimization of prediction of both outcomes using analogous squared-error measures: mean-squared error (MSE) for the longitudinal

outcome and the Integrated Brier score (IBS) for the time-to-event outcome. By using the IBS we assess the accuracy of prediction of the time-to-event outcome integrating the squared-error over a relevant time interval rather than computing its value at a specific time-point, giving a more complete summary of the accuracy of prediction across time. This strategy combines penalized likelihood with the LASSO penalty and cross-validation methods to select the fixed effects that optimize simultaneously the MSE and the IBS in out-of-sample predictions. Our proposed strategy is discussed in Chapter 5 and applied to the CARE75+ data set for a joint model of frailty and mortality. Our simulation studies suggest that it is not always possible to optimize simultaneously the MSE and IBS, but there is a region defined by the constraints imposed to the log-likelihood close to an optimal solution. In order to explore these regions more closely and as a secondary criterion, we assessed the accuracy of variable selection of our proposed strategy relative to the true model.

Joint modelling of longitudinal and time-to-event data has been an area of active research for description and prediction, but causal inference has received less attention. In Chapter 6 we revisit the joint models fitted and discussed in Chapter 4 and interpret them in the light of the causal inference framework (Section 2.5), pointing out their limitations in this context. By using Directed Acyclic Graphs (DAGs) we switch the focus from description to causal inference. DAGs allow us to state our hypotheses about the paths between frailty, falls and mortality and confounders in order to reformulate the two joint models fitted in Chapter 4, but this time adjusting for confounders. Additionally, since the two alternative joint models represent two plausible underlying mechanisms of the data, we conducted a simulation study with these joint models to understand the consequences of model misspecification to help us decide on the best model for frailty, falls and mortality.

Finally, in Chapter 7 we summarize our conclusions, future work and possible extensions to the work we present in this document.

Chapter 2

Preliminaries

Joint modelling longitudinal and time-to-event outcomes was motivated by data missing not at random in longitudinal studies and by the need to accommodate endogenous time-varying covariates in survival analysis models. In this chapter we describe the linear-mixed model and survival analysis models since they are the building blocks of joint models.

An appealing characteristic of joint models is the possibility to predict how individual response trajectories change over time and to make dynamic predictions of both longitudinal and time-to-event outcomes as more data are being collected. In this chapter we introduce prediction with the linear mixed model and survival analysis models, describing how to assess the accuracy of predictions, with special emphasis on time-to-event outcomes. This topic is extended to joint models in Chapter 3.

An important part of this chapter is dedicated to introduce other Statistics topics related to methods and techniques we use in Chapters 4–6, mainly penalized likelihood methods and causal inference.

2.1 Linear Mixed Models

The generic term *correlated data* embodies several data structures: multivariate observations, clustered data, repeated measurements (under different experimental con-

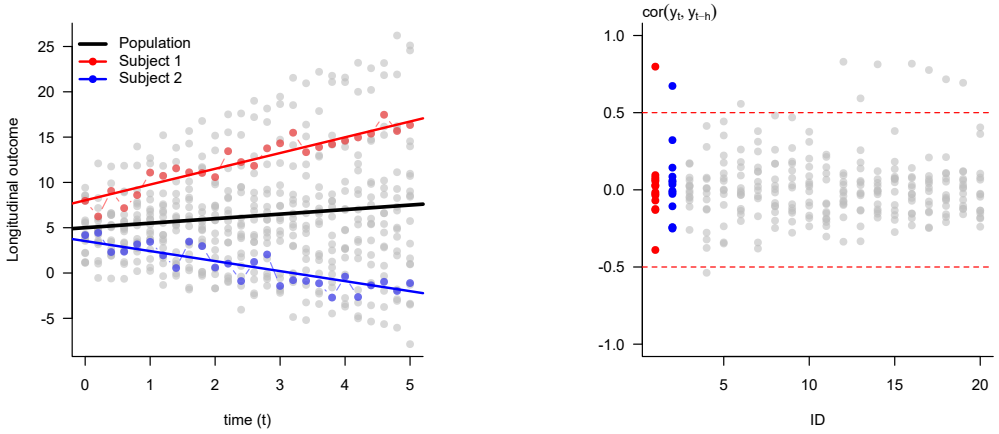
ditions), longitudinal data, and spatially correlated data (Verbeke, 1997). The focus of this section is on longitudinal data which can be broadly defined as the design where measurements of the same subject or individual (human beings, animals, laboratory samples, machines, etc.) are taken repeatedly through time (Rizopoulos, 2012). Longitudinal data are usually collected on a sample of subjects, with which it is possible to assess between-subject differences as in a cross-sectional design, but having repeated measures on the same subjects allows to investigate the within-subject change over time. The direct assessment of within-subject changes in the response over time can only be achieved with a longitudinal study design. Hence the primary goal of a longitudinal study is to characterize the change in response over time and the factors that influence change (Fitzmaurice *et al.*, 2012).

A distinctive feature of longitudinal data is that they are naturally *clustered*, where each cluster is comprised by the repeated measures of the same subjects at different occasions. From the description above we expect observations in the same cluster to be positively correlated (Fitzmaurice *et al.*, 2012; Molenberghs & Verbeke, 2000; Rizopoulos, 2012), a feature that implies that standard statistical methods that assume independent observations, such as ordinary linear regression, are not appropriate for the analysis of longitudinal data.

2.1.1 Model specification

The analysis of longitudinal data is based on the idea that each subject has their own subject-specific mean response profile in time, for which a functional form is assumed. We illustrate this idea in Figure 2.1a where the dots represent the longitudinal response data of a fictitious sample of 20 subjects, and the population average profile is represented by the solid black line. The data corresponding to two subjects (1 and 2) are the red and blue dots respectively, with overlaid solid lines representing their linear trend, illustrating the notion of subjects having their own response trajectories. Figure 2.1b contains the partial autocorrelation coefficients for lags $h > 0$ for each subject, illustrating the point of longitudinal data being positively correlated.

Longitudinal data can be analyzed with *Linear Mixed Models* (LMMs), statistical models for continuous (or quantitative) outcome variables in which the residuals are nor-



(a) (b)

Figure 2.1: (a) Longitudinal data of a fictitious sample of subjects (●) and the population profile (—). The longitudinal data of two subjects is highlighted (–●– –●–) with their subject-specific longitudinal profile (— —). (b) For each value on the x-axis (subjects in the sample) the points along the y-axis are the partial autocorrelation coefficient, $\text{cor}(y_{it}, y_{it-h}),\ h > 0$, of their longitudinal response.

mally distributed but may not be independent or have no constant variance. This name is due to the fact that LMMs are expressed as a linear combination of covariates (observed features of the data) and regression coefficients, and may involve a mix of fixed and random effects. Fixed effects are unknown constant parameters associated with covariates. Random effects are also associated with covariates, but in contrast to fixed effects, random effects are unobserved random variables assumed to be normally distributed (West *et al.*, 2014). The role of the random effects is to model the correlation due to repeated measures of the outcome variable across time.

To formally introduce the LMM, consider a sample of n subjects. Let $y_i(t)$ denote the value of the outcome variable measured on subject $i,\ i = 1, \ldots, n$ at time t. This implies that, in principle, the value of y_i is potentially known at any time t. However, in practice longitudinal data are collected at discrete time points, potentially different for each subject. Let $y_{ij} = y_i(t_{ij})$ denote the response of subject $i = 1, \ldots, n$ at time $t_{ij}, j = 1, \ldots, n_i$. In Figure 2.1a, $n = 20$ and $n_i = 26\ \forall i$, and it suggests that the

outcome data of each subject, y_{ij}, can be described by a linear function of time,

$$y_{ij} = \widetilde{\beta}_{i0} + \widetilde{\beta}_{i1} t_{ij} + \varepsilon_{ij}, \tag{2.1}$$

where the measurement error, ε_{ij}, represents the deviation of the observed response of subject i at time t_{ij} with respect to the subject-specific mean trajectory at the same time. Normality of the measurement error is a standard assumption, although some considerations about its variance are required since it is meant to model the within-subject variation. The data used to produce the plots of Figure 2.1, for instance, was simulated with $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

The LMMs methodology postulates that the intercept and slope of the individual profiles can be expressed in terms of the population mean trajectory and subject-specific deviations about the population mean. Reformulation of Equation (2.1) assuming a linear trajectory of the population mean gives a model in terms of fixed and random effects as described by Equation (2.2),

$$y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) t_{ij} + \varepsilon_{ij}, \tag{2.2}$$

where

$$\widetilde{\beta}_{i0} = \beta_0 + b_{i0},$$
$$\widetilde{\beta}_{i1} = \beta_1 + b_{i1},$$
$$\mathbf{b}_i = (b_{i0}, b_{i1})^\top \sim \mathcal{N}_2(\mathbf{0}, B),$$

with $B = \begin{bmatrix} \tau_0^2 & \\ \tau_{01} & \tau_1^2 \end{bmatrix}$ being the covariance matrix of the assumed normal distribution of $\mathbf{b}_i$.

The population mean trajectory is represented by $\beta_0 + \beta_1 t_{ij}$, where the intercept β_0 and the time slope β_1 are the fixed effects of the model. The random effects, $\mathbf{b}_i = (b_{i0}, b_{i1})^\top$, are subject-specific deviations with respect to the population intercept and slope, respectively. The random effects are assumed to follow a bivariate normal distribution with zero-mean vector and covariance matrix B.

In this model, the fixed effects are directly estimated from the data. In contrast, being

random variables b_{i0} and b_{i1} are not estimated directly, but rather the parameters of the covariance matrix of their joint distribution, $\tau_0^2, \tau_1^2, \tau_{01}$. The complete set of parameters to estimate in this model are $(\beta_0, \beta_1, \tau_0^2, \tau_1^2, \tau_{01}, \sigma^2)$.

Let $\mathbf{y}_i$ denote the vector of n_i repeated measures for subject i and $\boldsymbol{\varepsilon}_i$ the associated measurement error vector. The generalization of model (2.2) would include additional covariates and regression coefficients resulting in the linear mixed model proposed by Laird & Ware (1982), and is given by Equation (2.3),

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \tag{2.3}$$

where

X_i : $n_i \times p$ design matrix of covariates for the fixed effects,

Z_i : $n_i \times q$ design matrix (possibly same as X_i) associated to the random effects,

$\boldsymbol{\beta}$: p-vector of fixed effects regression coefficients,

$\mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}, B)$,

$\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_i)$,

B : covariance matrix of the joint distribution of the random effects,

Σ_i : $n_i \times n_i$ covariance matrix of measurement errors.

The p-vector $\boldsymbol{\beta}$ are the fixed effects of the model, and the subject-specific q-vectors $\mathbf{b}_i$ are the random effects. The columns in the design matrix X_i contain a unit vector for the intercept and possibly time-varying covariates measured at time points t_{ij}. The design matrix Z_i is formed by features of the data linking $\mathbf{b}_i$ with $\mathbf{y}_i$, usually some columns of X_i, and $Z_i = X_i$ is possible. The columns of X_i and Z_i correspond to covariates which can be time-varying or fixed baseline. An example of design matrices X_i and Z_i with a random intercept, random time-slope (i.e. $q = 2$) and additional $p-2$ fixed effects associated to time-varying covariates, $w_{i1}(t), \ldots, w_{ip-2}(t)$, measured at

time points $t_{ij}$is the following.

$$\underset{(n_i \times p)}{X_i} = \begin{pmatrix} 1 & t_{i1} & w_{i1}(t_{i1}) & \cdots & w_{ip-2}(t_{i1}) \\ 1 & t_{i2} & w_{i1}(t_{i2}) & \cdots & w_{ip-2}(t_{i2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{in_i} & w_{i1}(t_{in_i}) & \cdots & w_{ip-2}(t_{in_i}) \end{pmatrix}, \qquad \underset{(n_i \times 2)}{Z_i} = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix}$$

The fixed effects, $\boldsymbol{\beta}$, are interpreted exactly as the coefficients of ordinary linear regression, i.e. β_k represent the change in the mean response y_i per unit change in $w_{ik}, k = 1, \ldots, p$. Being subject-specific, the random effects are interpreted as the deviation of between the i^{th} subject and the population mean trend (intercept and slope).

The random effects vector, $\mathbf{b}_i$, is assumed to follow a multivariate normal distribution with mean zero and covariance matrix B. The measurement error vector of each subject, $\boldsymbol{\varepsilon_i}$, is assumed multivariate normal with mean zero and covariance matrix Σ_i. Here Σ_i depends on i only through its dimension n_i (i.e. the dimension of Σ_i is $n_i \times n_i$, the number of repeated measures of subject i), but the set of unknown parameters in Σ_i will not depend upon i. The random effects are assumed independent of measurement errors, so $\text{cov}(\mathbf{b}_i, \boldsymbol{\varepsilon}) = 0$, and the measurement error vectors are assumed independent between subjects, $\boldsymbol{\varepsilon_i} \perp \boldsymbol{\varepsilon_{i'}}$, $\forall i \neq i'$.

The contribution to the marginal likelihood of the i^{th} subject response vector is given by Equation (2.4).

$$f(\mathbf{y}_i) = \int_{\mathbb{R}^q} f(\mathbf{y}_i, \mathbf{b}_i) \mathrm{d}\mathbf{b}_i = \int_{\mathbb{R}^q} f(\mathbf{y}_i \mid \mathbf{b}_i) f(\mathbf{b}_i) \mathrm{d}\mathbf{b}_i, \tag{2.4}$$

where $\int_{\mathbb{R}^q} yf(\mathbf{b}_i)\mathrm{d}\mathbf{b}_i$ denotes the q-dimentional integral of $yf(\mathbf{b}_i)$ with respect to each element in the q-vector $\mathbf{b}_i$. A consequence of the normal distribution and independence assumptions of the random effects and the measurement error vectors is that, the n response vectors are conditionally independent and normally distributed given the random effects. This is, $\mathbf{y}_i|\mathbf{b}_i \sim \mathcal{N}_{n_i}(X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i, \Sigma_i)$, so the integral in Equation (2.4) has a closed-form solution leading to an n_i-dimensional normal distribution with

mean vector and covariance matrix given by

$$\begin{aligned}\mathbb{E}(\mathbf{y}_i) =& X_i\boldsymbol{\beta} \text{ and}\\ \text{var}(\mathbf{y}_i) =& V_i = Z_i B Z_i^\top + \Sigma_i.\end{aligned}$$

It is important to distinguish between $\mathbb{E}(\mathbf{y}_i) = X_i\boldsymbol{\beta}$ and $\mathbb{E}(\mathbf{y}_i|\mathbf{b}_i) = X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i$. The marginal expectation $\mathbb{E}(\mathbf{y}_i)$ is the mean value of $\mathbf{y}_i$ for the subset of the entire population sharing features X_i. The conditional expectation given the random effects, $\mathbb{E}(\mathbf{y}_i|\mathbf{b}_i)$, is the expectation of the subject-specific response $\mathbf{y}_i$ taking into account the random effects. A practical implication of making this distinction clear is that subject specific inferences and predictions of the outcome are possible once $\mathbf{b}_i$ is predicted. Neglecting prediction of the random effects (or using the expected value of $\mathbf{0}$) will produce exactly the same $\widehat{y}_i(t)$ for all subjects with the same covariates' values, X_i.

The parameters to estimate in this general LMM are $\boldsymbol{\theta}^\top = \left(\boldsymbol{\beta}^\top, \text{vech}(B), \boldsymbol{\theta}_\Sigma\right)$, where $\text{vech}(B)$ is the vector formed by stacking the columns of the lower triangular part of the symmetric matrix B and $\boldsymbol{\theta}_\Sigma$ is the vector of parameters chosen to model the covariance matrix Σ_i which, as stated previously, Σ_i depends on i only through its dimension.

Note that the model described by Equation (2.3) is general, especially in the assumptions about the measurement errors. Here no independence is assumed, and $\text{var}(\boldsymbol{\varepsilon}_i) = \Sigma_i$ allows for the possibility of modelling within-subject dependence of the measurement error when the chosen random effects structure is not sufficient to capture the correlation in the data. This topic is addressed, for instance, in Fitzmaurice *et al.* (2012); Hedeker & Gibbons (2006); Molenberghs & Verbeke (2000); Pinheiro & Bates (2006) , discussing various correlation structures for modelling within-subject dependence, including serial and spatial correlation structures and their implementation in R and SAS. For instance, common covariance structures to address serial correlation are compound symmetry, Toeplitz, autoregressive process of order 1, autoregressive-moving average of order (1,1), and banded; and exponential and Gaussian, to address spatial correlation.

A special case of the LMM arises assuming measurement errors with constant vari-

ance, i.e.

$$\Sigma_i = \sigma^2 I_{n_i}, \tag{2.5}$$

where I_{n_i} is the n_i-dimensional identity matrix and σ^2 the variance of the measurement errors. Laird & Ware (1982) named this model the "conditional-independence model" (CIM), since it implies that the n_i responses of subject i are conditionally independent given the random effects, $\mathbf{b}_i$,

$$f(\mathbf{y}_i \mid \mathbf{b}_i) = \prod_{i=1}^{n_i} f(y_{ij} \mid \mathbf{b}_i). \tag{2.6}$$

In this thesis we work with the CIM and, unless otherwise explicitly stated, this is the linear mixed model we always refer to. We introduced the general version to make clear that LMM can be used in modelling both the mean and covariance, pointing out the corresponding parts of the model.

Sometimes, it is convenient to express the LMM of Equation (2.3) for a single value of the response vector $\mathbf{y}_i$ at time t, by a row of matrices X_i and Z_i under the premise that longitudinal data can be potentially measured at any time t. This equivalent representation of the LMM is given by Equation (2.7), and will become useful in specifying the joint modelling framework.

$$y_i(t) = \mathbf{x}_i^\top(t)\boldsymbol{\beta} + \mathbf{z}_i^\top(t)\mathbf{b}_i + \varepsilon_i(t), \tag{2.7}$$

where $y_i(t)$ and $\varepsilon_i(t)$ are the i^{th} subject longitudinal outcome and measurement error at time t. The vectors $\mathbf{x}_i(t)$ and $\mathbf{z}_i(t)$ contain the covariates' values of fixed and random effects ($\boldsymbol{\beta}$ and $\mathbf{b}_i$), respectively, measured at time t.

2.1.2 Estimation

Parameter estimation of LMM is carried out under maximum likelihood principles. Assume the CIM, i.e. $\Sigma_i = \sigma^2 I_{n_i}$ and let $\boldsymbol{\theta}_b = \text{vech}(B)$. The parameters to estimate are $\boldsymbol{\theta}^\top = \left(\boldsymbol{\beta}^\top, \boldsymbol{\theta}_b, \sigma^2\right)$. Assuming normality of the random effects and measurement errors and $\text{cov}(\mathbf{b}_i, \boldsymbol{\varepsilon}_i) = 0$, the n response vectors $\mathbf{y}_i$ are conditionally independent

given the random effect, so by Equation (2.4) the log-likelihood function of the CIM is given by Equation (2.8),

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log\left(f(\mathbf{y}_i;\theta)\right) \\ &= -\frac{1}{2}\sum_{i=1}^{n}\Big(n_i\log(2\pi) + \log|V_i| + (\mathbf{y}_i - X_i\boldsymbol{\beta})^\top V_i^{-1}(\mathbf{y}_i - X_i\boldsymbol{\beta})\Big), \end{aligned} \tag{2.8}$$

where $\boldsymbol{\theta}^\top = \left(\boldsymbol{\beta}^\top, \boldsymbol{\theta}_b, \sigma^2\right)$ denotes the full parameter vector and $V_i = Z_i B Z_i^\top + \sigma^2 I_{n_i}$ is the covariance matrix of the n_i-dimensional normal distribution with $|V_i|$ denoting the determinant of the square matrix V_i.

Estimation of $\boldsymbol{\theta}$ is done in an iterative procedure by splitting $\boldsymbol{\theta}$ into the parameters of the fixed effects, $\boldsymbol{\beta}$, and the variance parameters, $(\boldsymbol{\theta}_b, \sigma^2)$.

The score equation for $\boldsymbol{\beta}$ is

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^\top} = \sum_{i=1}^{n} X_i^\top V_i^{-1}(\mathbf{y}_i - X_i\boldsymbol{\beta}) = 0.$$

If we assume V_i is known, the maximum likelihood estimator of the fixed effects vector $\boldsymbol{\beta}$, obtained by maximizing $\ell(\boldsymbol{\theta})$, conditional on the parameters in V_i, has a closed form and corresponds to the generalized least squares estimator,

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} X_i^\top V_i^{-1} X_i\right)^{-1} \sum_{i=1}^{n} X_i^\top V_i^{-1}\mathbf{y}_i, \tag{2.9}$$

with variance given by

$$\operatorname{var}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} X_i^\top V_i^{-1} X_i\right)^{-1}. \tag{2.10}$$

The estimator $\widehat{\boldsymbol{\beta}}$ of Equation (2.9) is unbiased for any choice of V_i, with the most efficient estimator of $\boldsymbol{\beta}$ being the one that uses the true value of V_i (Fitzmaurice *et al.*, 2012). An estimate of $\boldsymbol{\theta}_b$ and σ^2 can be obtained by replacing $\widehat{\boldsymbol{\beta}}$ in the log-likelihood of Equation 2.8 and maximizing $\ell(\boldsymbol{\theta}_b, \sigma^2 \mid \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}})$, where iterative procedures like

Newton–Raphson are commonly used.

Once $\widehat{V}_i$ is obtained, the the estimated variance of the fixed effects,$\widehat{\text{var}}(\widehat{\boldsymbol{\beta}})$, can be obtained by replacing $\widehat{V}_i$ in Equation (2.10).

The standard asymptotic maximum likelihood theory states that the maximum likelihood estimate of V_i will be asymptotically unbiased. However, in small samples, the maximum likelihood estimated of V_i will be biased because it does not take into account the fact that $\boldsymbol{\beta}$ is estimated from the data as well. This problem is similar in linear regression, where the variance estimate of the error term, $\widehat{\sigma}^2$, is known to be biased and the factor $n/(n-p)$ is applied for bias correction, where p is the length of $\boldsymbol{\beta}$.

The theory of restricted maximum likelihood (REML) estimation was developed (Harville, 1977; Patterson & Thompson, 1971) to address this problem. The main idea behind REML estimation is to separate the part of the data used in the estimation of V_i from the part used for the estimation of $\boldsymbol{\beta}$, i.e. the REML estimation of V_i eliminates $\boldsymbol{\beta}$ from the likelihood so that it is defined in terms of V_i. Rather than maximizing the log-likelihood of Equation (2.8), REML maximizes the modified log-likelihood function

$$\begin{aligned}\ell_{\text{REML}}(\boldsymbol{\theta}_b, \sigma^2) &= \ell(\boldsymbol{\theta}_b, \sigma^2 \mid \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}) - \frac{1}{2}\log\left|\sum_{i=1}^{n} X_i^\top V_i^{-1} X_i\right| \\ &\propto -\frac{1}{2}\log|V_i| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - X_i\widehat{\boldsymbol{\beta}})^\top V_i^{-1}(\mathbf{y}_i - X_i\widehat{\boldsymbol{\beta}}) \\ &\quad - \frac{1}{2}\log\left|\sum_{i=1}^{n} X_i^\top V_i^{-1} X_i\right|, \end{aligned} \tag{2.11}$$

where $\hat{\boldsymbol{\beta}}$ are the fixed effects estimates of Equation (2.9). The estimate $\widehat{V}_i$ obtained by maximizing $\ell_{\text{REML}}(\boldsymbol{\theta}_b, \sigma^2)$ corrects for the fact that $\boldsymbol{\beta}$ has also been estimated. This modified likelihood does not have a closed form, so numerical optimization methods are required. Most commonly used are Expectation-Maximization and Newton–Raphson algorithms.

Variance estimates of $(\boldsymbol{\theta}_b, \sigma^2)$ of REML can be obtained from the inverse of the corresponding block of the Fisher information matrix.

Prediction of the random effects is derived by an extension of the Gauss–Markov theorem over the random effects (Harville, 1977), which is equivalent to the expectation of the posterior distribution of the random effects given the observed data, $\mathbb{E}(\mathbf{b}_i|\mathbf{y}_i)$, given by (Commenges & Jacqmin-Gadda, 2015; Fitzmaurice *et al.*, 2012; Laird & Ware, 1982)

$$\begin{aligned}\mathbb{E}(\mathbf{b}_i \mid \mathbf{y}_i) &= \mathbb{E}(\mathbf{b}_i) + \operatorname{cov}(\mathbf{b}_i, \mathbf{y}_i) \left(\operatorname{var}(\mathbf{y}_i)\right)^{-1} [\mathbf{y}_i - \mathbb{E}(\mathbf{y}_i)] \\ &= BZ_i^\top V_i^{-1}(\mathbf{y}_i - X_i\boldsymbol{\beta}),\end{aligned} \tag{2.12}$$

and variance

$$\operatorname{var}(\mathbf{b}_i) = BZ_i^\top V_i^{-1} \left\{ V_i - X_i \left(\sum_{i=1}^{n} X_i^\top V_i^{-1} X_i \right)^{-1} X_i^\top \right\} V_i^{-1} Z_i B. \tag{2.13}$$

The parameters $\boldsymbol{\beta}$ and V_i are replaced by their estimators in (2.12) to obtain the empirical Bayes estimator, $\widehat{\mathbf{b}}_i = BZ_i^\top \widehat{V}_i^{-1}(\mathbf{y}_i - X_i\widehat{\boldsymbol{\beta}})$, and in (2.13) for a variance estimate.

An appealing characteristic of the mixed model is the possibility to predict how individual response trajectories change over time. This is the main reason for its use in the framework of joint modelling.

2.1.3 Missing data in longitudinal studies

Longitudinal studies are designed to collect data of a sample of subjects repeatedly over time, usually at prespecified follow up time points. Missing data occurs when the measurements for some subjects cannot be taken as planned and those measurements simply does not exist in the data set. Little & Rubin (2019) define missing data as unobserved values that would be meaningful for analyses if observed, i.e. a missing value hides a meaningful value.

Missingness can occur according to different patterns:

1. *Attrition* or *dropout* occur when a subject is withdrawn from the study before it is completed.
2. *Late entry* is when a subject does not provide some of the initial response measurements but until a later time point and stays until the end end of the study.
3. *Intermittent* patterns are those in which missing and observed measurements alternate along the follow up period.

Attrition results in an uninterrupted block of measurements followed by another block of missing data, similarly in a late entry process with the difference that the observed measurements follow the missing data block. Because of this feature of measurements being observed in uninterrupted blocks, attrition and late entry are considered monotone patterns and intermittent are also known as non-monotone.

Missing data impose important challenges for analyses with LMMs. The reduced sample size due to missigness causes a loss of efficiency of the estimates. Precision is directly related to the amount of data available, so missing data reduces precision of the estimates. Additionally, depending on the type of missing data mechanism, it can induce bias in the estimates.

In LMMs we assume that each subject in the study $i = 1, \ldots, n$ is designed to be measured at occasions $j = 1, \ldots, n_i$, so we expect to observe n vectors $\mathbf{y}_i^\top = (y_{i1}, \ldots y_{in_i})$. Let r_{ij} be an indicator variable that distinguishes observed from unobserved measurements. This is

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

The n_i-vector $\mathbf{r}_i^\top = (r_{ij}, \ldots, r_{in_i})$ is referred to as the *missing data process* and it induces the partition of vector $\mathbf{y}_i$ into two subvectors: $\mathbf{y}_i^{\text{o}} = \{y_{ij} : r_{ij} = 1\}$ and $\mathbf{y}_i^{\text{m}} = \{y_{ij} : r_{ij} = 0\}$ of sizes n_i^{o} and n_i^{m}, where $\mathbf{y}_i^{\text{o}}$ and $\mathbf{y}_i^{\text{m}}$ are the observed and missing measurements, and $n_i = n_i^{\text{o}} + n_i^{\text{m}}$. When the missing data process is restricted to dropout, the missing data process is of the form $(1, \ldots, 1, 0, \ldots, 0)$.

Missing data mechanisms can be thought of as a probability model describing the relation between the missing data process, $\mathbf{r}_i$, and the response data, $\mathbf{y}_i$. Let $\boldsymbol{\theta}_y$ and

$\boldsymbol{\theta}_r$ denote the parameter vectors that characterize the probability distribution of the response and the missing data processes, respectively, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_y, \boldsymbol{\theta}_r)$. Rubin (1976) proposed a taxonomy of the missing data mechanisms characterized by the conditional probability of the missing data process given the complete response vector, $\mathbf{y}_i = (\mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}})$:

$$f(\mathbf{r}_i \mid \mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}, \theta_r) \text{ or } f(\mathbf{r}_i \mid \mathbf{y}_i, \theta_r).$$

This taxonomy considers three types of missing data mechanisms: *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR) (Rubin, 1976).

Note that $\mathbf{y}_i^{\text{m}}$ are missing observations, so it should be understood as the longitudinal outcome measures that would have been obtained had they been observed.

Missing completely at random

The MCAR mechanism postulates that the probability of missing responses is independent of both observed and missing values, i.e.

$$f(\mathbf{r}_i \mid \mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}; \boldsymbol{\theta}_r) = f(\mathbf{r}_i; \boldsymbol{\theta}_r). \tag{2.14}$$

A useful way to think of MCAR is in terms of $\mathbf{y}_i^{\text{o}}$ and $\mathbf{y}_i^{\text{m}}$ as being random samples of the complete data $\mathbf{y}_i$, which means that the distribution of $\mathbf{y}_i$ is the same as the distribution of $\mathbf{y}_i^{\text{o}}$. Hence under MCAR inferences made by analyzing $\mathbf{y}_i^{\text{o}}$ are valid provided the statistical procedure used is valid. To illustrate MCAR, suppose a child in a drug prevention study withdraws because his/her parents move to take a job in a different city Graham (2012). In this case, the missing data pattern in the child's responses depends only on the parents taking a new job, but not on either the observed nor the potentially observed responses of the child.

Missing at random

MAR assumes that the probability of missingness depends on $\mathbf{y}_i^{\text{o}}$, but not on $\mathbf{y}_i^{\text{o}}$ given $\mathbf{y}_i^{\text{m}}$. The longitudinal data are MAR when $\mathbf{r}_i$ and $\mathbf{y}_i^{\text{m}}$ are conditionally independent given $\mathbf{y}_i^{\text{o}}$,

$$f(\mathbf{r}_i \mid \mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}; \boldsymbol{\theta}_r) = f(\mathbf{r}_i \mid \mathbf{y}_i^{\text{o}}; \boldsymbol{\theta}_r). \tag{2.15}$$

MAR arises, for instance, when a study protocol requires participants to be removed from the study when their response measurement exceeds a fixed medically relevant threshold. In this case, the missing data pattern is related to $\mathbf{y}_i^{\text{o}}$.

In contrast to MCAR, because in MAR the probability $\mathbf{r}_i$ depends on $\mathbf{y}_i^{\text{o}}$, the distribution of $\mathbf{y}_i^{\text{o}}$ is not the same as the distribution of $\mathbf{y}_i$, hence $\mathbf{y}_i^{\text{o}}$ cannot be considered a random sample of $\mathbf{y}_i$, only the distribution of $\mathbf{y}_i^{\text{m}}|\mathbf{y}_i^{\text{o}}$ is the same as the distribution of $\mathbf{y}_i^{\text{m}}$. Valid analyses can be obtained through a likelihood formulation that ignores the missing value mechanism provided the parameters describing the distribution of $\mathbf{y}_i^{\text{o}}$ are independent of the parameters describing the distribution of $\mathbf{y}_i^{\text{m}}$ (parameter distinctness condition (Molenberghs & Kenward, 2007)). Little & Rubin (2019) named this situation as *ignorability*.

A consequence of (2.15) is that the predictive distribution of the missing longitudinal responses $(\mathbf{y}_i^{\text{m}})$ given the observed data $(\mathbf{y}_i^{\text{o}})$ and the missing data process $(\mathbf{r}_i)$ depends only on $\mathbf{y}_i^{\text{o}}$ as follows

$$\begin{aligned} f(\mathbf{y}_i^{\text{m}} \mid \mathbf{y}_i^{\text{o}}, \mathbf{r}_i; \boldsymbol{\theta}) &= \frac{f(\mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}, \mathbf{r}_i; \boldsymbol{\theta})}{f(\mathbf{y}_i^{\text{o}}, \mathbf{r}_i; \boldsymbol{\theta})} = \frac{f(\mathbf{r}_i \mid \mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}; \boldsymbol{\theta}_r) f(\mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}; \boldsymbol{\theta}_y)}{f(\mathbf{r}_i \mid \mathbf{y}_i^{\text{o}}; \boldsymbol{\theta}_r) f(\mathbf{y}_i^{\text{o}}; \boldsymbol{\theta}_y)} \\ &= \frac{f(\mathbf{r}_i \mid \mathbf{y}_i^{\text{o}}; \boldsymbol{\theta}_r) f(\mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}; \boldsymbol{\theta}_y)}{f(\mathbf{r}_i \mid \mathbf{y}_i^{\text{o}}; \boldsymbol{\theta}_r) f(\mathbf{y}_i^{\text{o}}; \boldsymbol{\theta}_y)} = f(\mathbf{y}_i^{\text{m}} \mid \mathbf{y}_i^{\text{o}}; \boldsymbol{\theta}_y), \end{aligned}$$

and $\mathbf{y}_i^{\text{m}}$ can be validly predicted using only $\mathbf{y}_i^{\text{o}}$ under a model for the joint distribution $(\mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}})$.

Missing not at random

The MNAR mechanism states that the probability of missing longitudinal measures depends on both $\mathbf{y}_i^{\text{o}}$ and $\mathbf{y}_i^{\text{m}}$. In particular, the distribution of the missing data pattern, $\mathbf{r}_i$, depends on at least some elements of $\mathbf{y}_i^{\text{m}}$, even if conditioning on $\mathbf{y}_i^{\text{o}}$. An example of MNAR occurs in pain studies in which participants may ask for rescue medication when the intensity of pain they experience exceeds their own tolerance threshold (Rizopoulos, 2012).

As in MAR, in data sets with MNAR mechanisms $\mathbf{y}_i^{\text{o}}$ cannot be considered a random sample from $\mathbf{y}_i$. In contrast to MAR, in MNAR the predictive distribution of $\mathbf{y}_i^{\text{m}}|\mathbf{y}_i^{\text{o}}$ is not the same as $\mathbf{y}_i$, but rather depends on both $\mathbf{y}_i^{\text{o}}$ and $f(\mathbf{r}|\mathbf{y}_i)$.

Missing not at random model families

When longitudinal data are MNAR, we can only obtain valid inferences when analyses are based on the joint distribution of $\mathbf{y}_i$ and $\mathbf{r}_i$. The proposed models to accommodate MNAR can be grouped in three main families: *selection models*, *pattern mixture models* and *shared-parameter models* (Molenberghs & Kenward, 2007). These models are based on different factorizations of the joint distribution of the measurements and the missing data process.

1. Selection models (Heckman, 1976) are based on the factorization

$$f(\mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}, \mathbf{r}_i; \boldsymbol{\theta}) = f(\mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}; \boldsymbol{\theta}_y) f(\mathbf{r}_i \mid \mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}; \boldsymbol{\theta}_r).$$

2. Pattern mixture models, proposed by Little (1993), are based on

$$f(\mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}, \mathbf{r}_i; \boldsymbol{\theta}) = f(\mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}} \mid \mathbf{r}_i; \boldsymbol{\theta}_y) f(\mathbf{r}_i; \boldsymbol{\theta}_r).$$

   Here the probability of missing values weights the observed data to form a mixture model for each pattern of missing values.

3. The shared-parameter models introduce, $\mathbf{b}_i$, a subject-specific latent variable or random effects, with probability distribution characterized by the parameter vector $\boldsymbol{\theta}_b$. Shared-parameter models are based on the factorization

$$f(\mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}, \mathbf{r}_i; \boldsymbol{\theta}) = \int_{\mathbf{b}_i} f(\mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}} \mid \mathbf{b}_i; \boldsymbol{\theta}_y) f(\mathbf{r}_i \mid \mathbf{b}_i; \boldsymbol{\theta}_r) f(\mathbf{b}_i; \boldsymbol{\theta}_b) \mathrm{d}\mathbf{b}_i.$$

   The random effects, $\mathbf{b}_i$, are shared between both factors of the joint distribution are meant to capture the association between $\mathbf{y}_i$ and $\mathbf{r}_i$. They can be thought of as referring to a latent trait driving both measurement and missing data process (Molenberghs & Kenward, 2007).

Joint modelling longitudinal and time-to-event data falls in the shared-parameter models family. The connection between joint modelling and the missing data framework is explained in Section 3.1.4.

2.2 Survival Analysis

In the simple case, time-to-event data (or survival analysis) is the study of the time it takes for a subject (or a group of subjects) to observe a particular event of interest, like death, first marriage, divorce, credit default, system failure, etc. In these settings, the main variable of analysis is the *time* until the event of interest occurs, often referred to as *survival time*, *time-to-event*, *failure time* or *event time*. More complex scenarios include the possibility of different types of events, called competing risks (different causes of death), the possibility to experience more than one type of event (car insurance claims for different eventualities), or the possibility transient events that can occur repeatedly to a subject over time (recurrent events). An even more complex setting involves consideration of multiple events that may occur simultaneously, either once or repeatedly, to individuals over time.

Three basic requirements define time-to-event measurements (Kalbfleisch & Prentice, 2002):

1. An unambiguous origin for the measurement of "time" (should be precisely defined for each subject. All subjects should be as comparable as possible at the origin.
2. An agreed scale of measurement. Usually this is calendar/clock time, but depending on the context of the data it can be operating time, accumulated load, etc., which is chosen to provide direct, operational meaning in the problem context. It is always nonnegative.
3. A precise definition of response, or occurrence of the event of interest. It must be precisely defined: death from a specific cause, time to relapse or death, time when performance first exceeds a specific threshold.

One of the distinguishing features of time-to-event data is that they are positive real-valued random variables and often positively skewed. Statistical methods that rely on normality cannot be applied directly to analyze time-to-event data. In survival analysis, the phenomenon of unobserved values of the response measurement is called *censoring*. Censoring is the second and most important characteristic of time-to-event data and its defining feature is that the time-to-event outcome might not be fully observed

for all subjects in the sample being analyzed. Ignoring censored data can produce severely biased estimates, so it requires special treatment, depending on the censoring mechanism that occurs in the data.

The first step to accommodate censoring is to determine the *censoring time* for each subject, $C_i,\ i = 1, \ldots, n$. If censoring is independent of the event process, then it is *non-informative*, otherwise it is *informative*. When the event time for a subject cannot be determined during the follow-up period because it has already occurred before such subject is enrolled, we are in the presence of *left censoring*. If the event cannot be determined because either the study ended and the event never occurred or such subject was lost to follow-up at any time during the study, then we are talking about a *right censored* observation. If the event occurs during the follow-up period, but its exact time cannot be determined and only it is known that the event occurred within a no-zero length time interval, then the observation is said to be *interval censored*. It is typically assumed non-informative and right censoring in survival analysis studies.

Throughout this document we assume non-informative right censoring.

Truncation is one more characteristic of survival analysis data, related to the sampling mechanism. Truncation is a condition that screens or excludes subjects from the study population. *Left truncation* occurs when subjects have been at risk before the beginning of the study, e.g. a prospective study where subjects are followed from a specific date until they die of a particular cause. *Right truncation* occurs when only individuals who have experienced the event of interest are observable, e.g. a retrospective study of mortality based on death records. A thorough explanation and consequences of truncation and censoring can be found, for instance, in Kalbfleisch & Prentice (2002) and Klein & Moeschberger (2006).

Let T_i^* denote the time elapsed for subject i to experience the event of interest among a sample of n subjects, and let C_i denote the censoring time for subject i. We define T_i as the minimum of the true event time, T_i^*, and the censoring time, C_i, so $T_i = \min(T_i^*, C_i)$. Define also the event indicator $\delta_i = \mathbb{1}(T_i^* < C_i)$ that takes the value of 1 if the observed event corresponds to a true event time and 0 otherwise. Thus in a sample of n individuals the time-to-event outcome is denoted by the pair $\{T_i, \delta_i\}$, with

T_i being the follow-up time for subject $i,\ i = 1, \ldots, n$, and δ_i indicating whether the follow-up is interrupted by the event or by censoring.

There are different ways to characterize the distribution of a continuous time-to-event random variable, T:

$$\begin{aligned}
\text{Density} \quad f(t) &= \lim_{\mathrm{d}t \to 0} \left\{ \frac{1}{\mathrm{d}t} \Pr(t \leq T < t + \mathrm{d}t) \right\} = \frac{\mathrm{d}}{\mathrm{d}t} F(t) \\
\text{Cumulative distribution function} \quad F(t) &= \Pr(T \leq t) = \int_0^t f(x)\mathrm{d}x \\
\text{Survival function} \quad S(t) &= \Pr(T > t) = 1 - F(t) \\
\text{Hazard rate} \quad h(t) &= \lim_{\mathrm{d}t \to 0} \left\{ \frac{1}{\mathrm{d}t} \Pr(t \leq T < t + \mathrm{d}t | T \geq t) \right\} = \frac{f(t)}{S(t)} \\
\text{Cumulative hazard} \quad H(t) &= \int_0^t h(x)\mathrm{d}x = -\log S(t).
\end{aligned}$$

The survival function, $S(t)$, is the probability for the event to occur after t. Its name comes from medical studies where the event of interest is mortality, so T quantifies the time-to-death and $S(t)$ is the probability of surviving at least to time t. The hazard function is the instantaneous rate at which the event occurs at t, and the cumulative hazard is the accumulation of the occurrence rate across time.

Let $\{t_i, \delta_i\} = \{T_i = t_i, \delta_i\}, i = 1, \ldots, n$ denote the survival information in a random sample from a distribution function F, parametrized by $\boldsymbol{\theta}$, with probability density f. The formulation of the likelihood function by considering that censored subjects were still alive by t_i i.e. $\delta_i = 0$, and therefore, such subjects contributes with $S_i(t_i; \boldsymbol{\theta})$ to the likelihood. Subjects for whom an event is observed at time t_i contribute with $f(t_i; \boldsymbol{\theta})$, i.e. $\delta_i = 1$. Combining this information, we obtain the log-likelihood function:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \log f(t_i; \boldsymbol{\theta}) + (1 - \delta_i) \log S_i(t_i; \boldsymbol{\theta}). \tag{2.16}$$

This log-likelihood can be expressed in many ways, and it is most commonly expressed

in terms of the hazard and cumulative hazard functions

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \log h(t_i; \boldsymbol{\theta}) - H_i(t_i; \boldsymbol{\theta}). \tag{2.17}$$

The analysis of survival data can be parametric, semiparametric or non-parametric. Our focus for this section is the Cox Proportional Hazards model since it is this particular survival analysis model that we use for joint modelling along with longitudinal data.

2.2.1 The Cox proportional hazards model

Two hazard rates, $h_1(t)$ and $h_2(t)$ are proportional if there is a constant $\psi > 0$ such that

$$h_2(t) = \psi h_1(t) \quad \forall t > 0, \tag{2.18}$$

where ψ is called the hazard ratio. The relationship defined by Equation (2.18) is called the *proportional hazards property* (PH).

Cox (1972) proposed a PH model where no assumptions are made about the actual form of the baseline hazard function $h_0(t)$. Let $\mathbf{w}_i^\top = (w_{i1}, \ldots, w_{ip})$ denote a vector of time-independent covariates related to subject i. Equation (2.19) defines the Cox PH model, also known as relative risk or relative hazards model.

$$h_i(t) = h_0(t) \exp\{\mathbf{w}_i^\top \boldsymbol{\gamma}\}. \tag{2.19}$$

where

$\mathbf{w}_i^\top = (w_{i1}, \ldots, w_{ip})$ is the vector of covariates associated to the hazard of subject i,
$\boldsymbol{\gamma}^\top = (\gamma_1, \ldots, \gamma_p)$ is the vector of regression coefficients,
$h_0(t)$ is the baseline hazard function.

The vector $\mathbf{w}_i^\top \boldsymbol{\gamma}$ is called risk score, prognostic index or linear predictor. The baseline hazard function (or baseline risk), $h_0(t)$, corresponds to the hazard of a subject that has

linear predictor $\mathbf{w}_i^\top \boldsymbol{\gamma} = 0$.

The interpretation of the regression coefficients, $\boldsymbol{\gamma}$, is the magnitude of the effect of the covariates $\mathbf{w}_i$ on the hazard. For instance, consider the following setting where $\mathbf{w}_i$ and b_{i0} are covariates, and $\boldsymbol{\gamma}$ and η their associated regression coefficients. Then

$$\begin{aligned} h_i(t) &= h_0(t)\exp(\mathbf{w}_i^\top \boldsymbol{\gamma} + \eta b_{i0}) \\ \iff \quad \log\left\{\frac{h_i(t)}{h_0(t)}\right\} &= \mathbf{w}_i^\top \boldsymbol{\gamma} + \eta b_{i0} \\ \iff \quad \frac{\mathrm{d}\log\psi_i}{\mathrm{d}b_{i0}} &= \frac{\psi_i^{'}}{\psi i_i} = \eta \end{aligned}$$

where $\psi_i = h_i(t)/h_0(t)$ is the hazard ratio which does not depend on time in PH models. That is, η represents the change in the relative hazard per unit change in b_{i0}. Also note that for $x, z \in \mathbb{R}$

$$\frac{\psi_i(b_{i0} = z + x)}{\psi_i(t \mid b_{i0} = z)} = \frac{h_i(b_{i0} = z + x)}{h_i(t \mid b_{i0} = z)} = \frac{\exp(\mathbf{w}_i^\top \boldsymbol{\gamma} + \eta(z + x))}{\exp(\mathbf{w}_i^\top \boldsymbol{\gamma} + \eta z)} = \exp(\eta x). \tag{2.20}$$

So, $\exp(\eta x)$ is the marginal relative hazard of the event between two individuals whose b_{i0} value differs by x units.

The estimation of the primary parameters of interest, $\boldsymbol{\gamma}$, can be obtained through maximization of the partial log-likelihood function (Cox, 1972) given by

$$\ell_p(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \delta_i \Big[\mathbf{w}_i^\top \boldsymbol{\gamma} - \log\Big(\sum_{t_j \geq t_i} \exp\{\mathbf{w}_i^\top \boldsymbol{\gamma}\}\Big)\Big]. \tag{2.21}$$

The partial likelihood does not depend on the baseline hazard, so it is left unspecified, which means that there is no need for making any assumptions about the distribution of the event times. However, it assumes that covariates act multiplicatively on the hazard rate. The partial log-likelihood Equation (2.21) can be interpreted as a measure of how well the model can order the patients with respect to their survival time.

The parameter estimates, $\widehat{\boldsymbol{\gamma}}$, of the Cox PH model are maximum likelihood and are

found by solving the score equations of the partial log-likelihood:

$$\frac{\partial \ell_p(\boldsymbol{\gamma})}{\partial \gamma_k} = \sum_{i=1}^{n} \delta_i \left\{ w_{ik} - \frac{\sum_{t_j \geq t_i} w_{jk} \exp(\mathbf{w}_j^\top \boldsymbol{\gamma})}{\sum_{t_j \geq t_i} \exp(\mathbf{w}_j^\top \boldsymbol{\gamma})} \right\} = 0.$$

The estimates $\widehat{\boldsymbol{\gamma}}$ are consistent and asymptotically normally distributed. Standard errors are estimated using the observed information matrix, $I^{-1}(\widehat{\boldsymbol{\gamma}})$, where

$$I(\widehat{\boldsymbol{\gamma}}) = -\sum_{i}^{n} \frac{\partial^2 \ell_{pi}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^\top \partial \boldsymbol{\gamma}} \bigg|_{\boldsymbol{\gamma}=\widehat{\boldsymbol{\gamma}}}.$$

Kalbfleisch & Prentice (2002) and Klein & Moeschberger (2006) give further details and discussions about the properties of the estimates based on the partial likelihood function.

In Chapters 5 and 6 we conduct simulation studies with different specifications of joint models of longitudinal and time-to-event data, which use as building blocks the linear mixed model and survival analysis models. In Sections 2.2.1 and 2.2.2 we describe data simulation algorithms for the survival analysis models relevant to our simulation studies.

Estimation of the Cox PH model is a standard routine in most statistical software. For instance, the R package `survival` fits a Cox PH model using the function `coxph()`.

Survival analysis with random effects (frailty model)

Inference with the Cox proportional hazards model assumes that the observations are statistically independent, at least conditionally independent given covariates. However, this assumption may be violated and a possible option to accommodate the dependence between observations is by introducing a random effect in the survival analysis model, as in the LMM.

In the survival analysis literature a random effect is referred to as *fraily*, and survival analysis models that incorporate frailties are named *frailty models*. Random effects in frailty models are usually assumed to have a multiplicative effect in the hazard rate, so

the subjects with larger random effects have a higher risk of the event, i.e. are more "frail". That is why survival analysis models with random effects are called frailty models. In this thesis we do not use this terminology in order to avoid confusion with the concept of frailty in the medical and public health context, which refers to a geriatric syndrome related to the gradual deterioration of the body as people age. The data we use to illustrate the various statistical methods discussed in this thesis is from a population study conducted to learn about frailty in the elderly, and frailty is precisely one of the outcome variables in our analyses.

The idea of random effects models in survival analysis is to consider the variability in the time-to-event outcome as coming from two separate sources: (1) the simple randomness described by the hazard function, and (2) randomness described by a random effect, a random variable that is either an individual variable (univariate), or a variable common to several individuals (multivariate) (Hanagal, 2011; Hougaard, 1995). A random effect is included in survival analysis models as a factor that acts miltiplicatively in the hazard rate and, as in the LMM framework, it accounts for variability in the data by modelling the correlation between observations (Hanagal, 2011).

The univariate case refers to the traditional individual time-to-event data, the random effect describes heterogeneity, this is, the influence of unobserved risk factors in a proportional hazards model (Hanagal, 2011; Hougaard, 1995). Let u_i be a random variable that denotes the random effect for subject i, then the proportional hazards model described in Equation 2.19 becomes

$$h_i(t) = u_i h_0(t) \exp\{\mathbf{w}_i^\top \boldsymbol{\gamma}\}. \tag{2.22}$$

In the multivariate case, the random effect, u_i, is common to a group of individuals, for instance twins or members of the same family, modelling the dependence between individuals of the same group (Hougaard, 1995). For example, in epidemiological studies failure times may be clustered into groups such as families or geographical units: some unmeasured characteristics shared by the members of that cluster, such as genetic information or common environmental exposures could influence time to the studied event. In a different context, correlated data may come from recurrent events, i.e., events which occur several times within the same subject during the period

of observation, so in some sense recurrent events are clustered within subjects. A random effects survival analysis model accommodates the correlation of events within clusters by including a random effect such that subjects (or recurrent events) within the same cluster share the same random effect (Rondeau *et al.*, 2003). For the k^{th} $(k = 1, \ldots, K_i)$ individual of the i^{th} group or cluster $(i = 1, \ldots, n)$, let T^*_{ik} denote the event times, and let C_{ik} be the right-censoring times. The observation $T_{ik} = \min(T^*_{ik}, C_{ik})$ and the censoring indicators are $\delta_{ik} = \mathbb{1}(T^*_{ik} < C_{ik})$. The hazard function for a frailty model is

$$h_{ik}(t \mid u_i) = u_i h_0(t) \exp\{\mathbf{w}_{ik}^\top \boldsymbol{\gamma}\} = u_i h_{ik}(t), \tag{2.23}$$

where $h_0(t)$ is the baseline hazard function, $\mathbf{w}_{ik}$ the covariate vector of the k^{th} subject in the i^{th} group with associated vector of regression parameters $\boldsymbol{\gamma}$, and u_i is the random effect of the i^{th} group.

In both univariate and multivariate cases, it is assumed that the u_i are independent and identically distributed from a positive-valued distribution with a single variance parameter ϕ. Note that a scale factor common to all subjects in the study population may be absorbed into the baseline function $h_0(t)$, so that frailty distributions are standardized to $\mathbb{E}(u_i) = 1$ (Wienke, 2010). In practice, two commonly assumed distributions of the random effects are

- $u_i \overset{\text{iid}}{\sim} \text{Gamma}(\phi^{-1}, \phi^{-1})$, such that $\mathbb{E}(u_i) = 1$ and $\text{var}(u_i) = \phi$, and
- $\log(u_i) \overset{\text{iid}}{\sim} \mathcal{N}(m, s^2)$, such that $\mu = \mathbb{E}(u_i) = \exp(m + s^2/2)$ and $\phi = \text{var}(u_i) = \exp(2m + s^2)(\exp(s^2 - 1))$. See Wienke (2010) for further details.

Assuming a parametric baseline hazard, the parameters to estimate in the univariate random effects survival analysis model of Equation (2.22) are $\boldsymbol{\theta}^\top = h_0, \boldsymbol{\gamma}, \phi$, where h_0 represent the parameters of the chosen functional form of the baseline hazard, $\boldsymbol{\gamma}$ the regression coefficients of the observed covariates, and ϕ the variance parameter of the assumed distribution of the random effect. The estimation is done by maximizing the marginal log-likellihood function (Duchateau & Janssen, 2007). Equation (2.24) shows the contribution of subject i to the marginal log-likelihood

$$\ell_i(\boldsymbol{\theta} \mid u_i) = \log \left\{ \int_0^\infty \prod_{i=1}^{n} \left[u_i h(t_i; \boldsymbol{\theta})\right]^{\delta_i} u_i S_i(t_i; \boldsymbol{\theta}) f(u_i) \mathrm{d}u_i \right\}, \tag{2.24}$$

where $f(u_i)$ is the density of the random effect u_i.

The marginal likelihood of Equation 2.24 has a closed form when the random effects follow a $\text{Gamma}(\phi^{-1}, \phi^{-1})$ distribution and the baseline hazard is governed by the $\text{Weibull}(\kappa, \rho)$ distribution. Otherwise iterative optimization routines can be used, for instance the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977). Duchateau & Janssen (2007) explains how to implement the EM algorithm for survival analysis models with random effects, and Hanagal (2011) includes the case where the baseline hazard is unspecified.

Rondeau *et al.* (2003) proposed a maximum penalized likelihood approach to estimate random effects models like those described by Equations (2.22) and (2.23) along with a spline-approximated baseline hazard.

Prediction of the random effects, $\widehat{u}_i$ can be carried out following an empirical Bayes approach, as the expectation of the posterior conditional distribution of the random effects given the other parameters estimates (Vaida & Xu, 2000)

Survival analysis models with random effects can be fitted using R for several distributions of random effects. For instance, (1) the function `coxph()` function of the `survival` package fits a PH model with unspecified baseline hazards; (2) the `parfm()` function of the `parfm` package (Munda *et al.*, 2012) has several options for specifying a parametric baseline hazard; and (3) the `frailtyPenal()` function of the `frailtypack` (Rondeau *et al.*, 2012) package has also multiple options to specify the distribution of the baseline hazard.

For further details and more thorough discussion of survival analysis models with random effects, refer to (Duchateau & Janssen, 2007; Wienke, 2010).

Data simulation from the proportional hazards model

Data simulation from the proportional hazards model relies on the inverse transform method.

According to the proportional hazards model, the survival probabilities are obtained by

$$S_i(t \mid \mathbf{w}_i) = \exp\{-H_i(t \mid \mathbf{w}_i)\} = \exp\{-H_0(t)e^{\mathbf{w}_i^\top \boldsymbol{\gamma}}\},$$

where $S_i(t|\mathbf{w}_i)$ is the survival probability given the vector of covariates $\mathbf{w}_i$, $H_0(t) = \int_0^t h_0(x)\mathrm{d}x$ the baseline cumulative hazard and $h_0(t|\mathbf{w})$ the baseline hazard function.

Let $\xi_i = S_i(t|\mathbf{w}_i)$ and $h_0(t) > 0$, where $\xi_i \sim \mathcal{U}[0,1]$. The distribution of the random variable T_i, obtained from

$$T_i = H_0^{-1}\left(-\log(\xi_i)e^{-\mathbf{w}_i\boldsymbol{\gamma}}\right), \tag{2.25}$$

will be determined from the distribution that governs $H_0(t)$.

Simulations from a frailty model can be based on

$$T_{ik} = H_0^{-1}\left(-\frac{\log(\xi_{ik})e^{-\mathbf{w}_{ik}\boldsymbol{\gamma}}}{u_i}\right), \tag{2.26}$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, K_i$, where u_i is a random variable from an appropriate distribution such that $\mathbb{E}(u_i) = 1$ and $\mathrm{var}(u_i) = \phi$.

The function `f.simFrail` in Appendix A.1 is a sample R code to simulate data from a proportional hazards (frailty) model with three covariates, with the following options:

- A baseline hazard governed by the $\mathrm{Weibull}(\kappa, \rho)$ distribution,
- Either a fixed censoring time or a censoring process from the $\mathrm{Weibull}(\nu, \lambda)$ distribution,
- Either no frailty or two options of frailty:
  - $u_i \sim \mathcal{LN}(\mu, \phi)$
  - $u_i \sim \mathrm{Gamma}(\phi^{-1}, \phi^{-1})$

The steps involved in the `f.simFrail` are summarized in Algorithm 2.1.

Algorithm 2.1 Data simulation of a proportional hazards model with a random effect.

Decide on the sample size (n), the functional form of the baseline hazard $(h_0(t))$, the distribution of the random effect and the value of the variance parameter (ϕ), the values of the regression coefficients $(\boldsymbol{\gamma})$, and a simulation scheme for the vector of baseline covariates $(\mathbf{w}_i)$. Choose a censoring scheme, C_i, which can be an administrative censoring time fixed for all i.

1: Sample n instances of the random effect distribution, $\mathbf{u}^\top = (u_1, \ldots, u_n)$.
   For $i = 1, \ldots, n$:
2: Sample $\xi_i \sim \mathcal{U}(0, 1)$.
3: Derive $H_0(t)$ and $H_0^{-1}(t)$.
4: Simulate the time to event, T_{ik}^*, as indicated by Equation (2.26) with $K_i = 1\ \forall i$ (or T_i^* if a random effect is not required, as indicated by Equation (2.25)).
5: Set $T_{ik} = \min(T_{ik}^*, C_i))$ and $\delta_{ik} = \mathbb{1}(T_{ik}^* < C_i)$.

When the chosen baseline hazard cannot be integrated analytically to obtain $H_0(t)$ or when $H_0^{-1}(t)$ does not have a closed form, numerical methods can be used. The Gauss–Kronrod method (Ziegel, 1987) is numerical integration option to obtain $H_0(t)$, and the crude bisection method (Monaco *et al.*, 2018) can be used to get inverse of the cumulative hazard function by finding the root of $H_i(t) + \log(\xi_i)$.

Algorithm 2.1 can also be used to simulate data from grouped data by noting that n now represents the number of groups. Thus the simulation will be carried out by repeating steps (2)–(4) as many times as individuals are required in each group, K_i, taking into account that subjects in the same group must share the same random effect, u_i.

Example As an example consider simulating assuming a baseline hazard from the Weibull(κ, ρ) distribution, i.e. $h_0(t) = \kappa\rho(\rho t)^{\kappa-1}$ where $\kappa > 0, \rho > 0$. The baseline cumulative hazard is $H_0(t) = (\rho t)^\kappa$, with inverse $H_0^{-1}(x) = \rho^{-1}x^{1/\kappa}$. Provided $\kappa, \rho, \boldsymbol{\gamma}, \mathbf{w}_i$ and the distribution of the random effect u_i with variance parameter ϕ, sampling survival times from a proportional hazards model with a baseline hazard gov-

erned by the $\text{Weibull}(\kappa, \rho)$ distribution would based on the following expression:

$$T_i = \frac{1}{\rho}\left(\frac{-\log(\xi_i)\exp\{-\mathbf{w}_i^\top \boldsymbol{\gamma}\}}{u_i}\right)^{1/\kappa}, \ \xi_i \sim \mathcal{U}[0,1].$$

Time-varying covariates

The PH model described in Equation (2.19) was initially developed assuming constant covariates, $\mathbf{w}_i^\top$, or least constant along the follow up period, for instance gender, ethnicity, age at baseline, etc. In practice, many studies that generate time-to-event data, record on a regular basis other variables whose value vary in time and it is often of interest to investigate the relationship between such variables and the time-to-event outcome. These are *time-varying* or *time-dependent* covariates.

The Cox PH model with baseline covariates vector $\mathbf{w}_i$ and time-varying covariate $y_i(t)$ takes the form described by Equation (2.27)

$$h_i(t) = h_0(t)\exp\{\mathbf{w}_i^\top \boldsymbol{\gamma} + \beta y_i(t)\}. \tag{2.27}$$

where

$\mathbf{w}_i^\top = (w_{i1}, \ldots, w_{ip})$ is the vector of covariates associated to the hazard of subject i

$\boldsymbol{\gamma}^\top = (\gamma_1, \ldots, \gamma_p)$ is the vector of regression coefficients,

$y_i(t)$: time-varying covariate,

β : regression coefficient associated to $y_i(t)$,

$h_i(t)$: baseline hazard function.

It is important to distinguish between two types of time-varying covariates because they are handled in different ways: *endogenous* or *internal* and *exogenous* or *external* time-varying covariates. Suppose the event of interest is death. An endogenous covariate, say $y_i(t)$, is intrinsic to subject i and can only be measured while the subject is alive Collett (2015), so observing a value for $y_i(t)$ at time t necessarily implies survival at least at time t. In clinical studies, endogenous covariates can be, for instance, biomarkers (white blood cell count, blood pressure, serum cholesterol level,

etc.), which are susceptible to be measured only as long subjects are still alive.

On the other hand, exogenous covariates do not necessarily require the subject's survival for their existence, they remain measurable and their distributions unchanged after the occurrence of the event (Commenges & Jacqmin-Gadda, 2015). There are two types of exogenous covariates. The first type correspond to covariates that change in such a way that their future values are exactly known, for instance subjects' age. The second type correspond to those covariates that are exist completely independent of the subjects, like environmental factors Collett (2015).

Kalbfleisch & Prentice (2002) formalizes the definition of exogenous time-varying covariate. An exogenous time-varying covariate is a *predictable process*, i.e. its value at any time t is known infinitesimally before t. Let $y_i(t)$ denote the value of a time-varying covariate of subject i at time t, and let $\mathcal{F}_i^y(t) = \{y_i(s); 0 \leq s < t\}$ give the covariate history up to time t. An exogenous covariate is a time-varying variable satisfying the following condition:

$$\underbrace{\Pr\left\{T_i \in [s, s+\mathrm{d}s) \mid T_i \geq s, \mathcal{F}_i^y(s)\right\}}_{h\left\{s|\mathcal{F}_i^y(s)\right\}} = \underbrace{\Pr\left\{T_i \in [s, s+\mathrm{d}s) \mid T_i \geq s, \mathcal{F}_i^y(t)\right\}}_{h\{s|\mathcal{F}^y(t)\}}, \tag{2.28}$$

for all $s,\ t$ such that $0 < s \leq t$, and $\mathrm{d}s \to 0$.

An equivalent condition is,

$$\Pr\left\{\mathcal{F}_i^y(t) \mid \mathcal{F}_i^y(s), T_i \geq s\right\} = \Pr\left\{\mathcal{F}_i^y(t) \mid \mathcal{F}_i^y(s), T_i = s\right\}, \quad s \leq t \tag{2.29}$$

which formalizes the idea that $y(t)$ is associated with $h(t)$, but its future path up to time $t > s$ is not affected by the event occurring at time s.

Endogenous covariates do not satisfy the exogeneity condition and they are not a predictable process.

The partial likelihood of the proportional hazards model (Equation (2.21)) requires the knowledge of the exact values of the covariates at each time of event for all individuals at risk. The standard proportional hazards model is extended to incorporate exogenous

time-varying covariates, as shown in Equation (2.27) and parameter estimation is based on a more general version of the partial log-likelihood function, described in Equation (2.30).

$$\ell_p(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \delta_i \Big[\mathbf{w}_i^\top \boldsymbol{\gamma} + \beta y_i(t_i) - \log \Big(\sum_{k \in \mathcal{R}(t_i)} \exp\{\mathbf{w}_k^\top \boldsymbol{\gamma} + \beta y_i(t_i)\} \Big) \Big] \tag{2.30}$$

where $\mathcal{R}(t)$ is the *risk set* defined by the subjects still at risk of the event at time t, $y_i(t)$ is the value of the time-varying covariate of subject i measured at time t and β its regression coefficient.

Estimation of the extended Cox model based on the partial log-likelihood of Equation (2.30) requires the imputation of the time-varying covariates at each time of event for all subjects. It is typically assumed that time-varying covariates remain constant between consecutive measurement time points, i.e. imputation is done by carrying the last value forward. Refer to Lawless (2011) or Collett (2015) for further discussion of the extended Cox model with time-varying covariates.

Dealing with endogenous covariates requires a special treatment because the log-likelihood of Equation (2.16) is not valid. As we mentioned, endongenous covariates require survival of the subject for their existence. Therefore, when the terminal event is death, the path $\mathcal{F}_i^y(t)$ carries direct information about the time-to-death, i.e. provided that $y_i(t - ds)$ with $ds \to 0$ exists, the survival function satisfies (Kalbfleisch & Prentice, 2002; Rizopoulos, 2012)

$$S_i(t \mid \mathcal{F}_i^y(t)) = \Pr(T_i > t \mid \mathcal{F}_i^y(t)) = 1. \tag{2.31}$$

Similarly, subject i dying at time s necessarily implies the nonexistence of $y_i(t)$ at time $t > s$, which is a violation of the exogeneity condition of Equations (2.28) and (2.29).

$$S_i(t \mid \mathcal{F}_i^y(t)) = \lim_{\mathrm{d}s \to 0} \left\{ \frac{1}{\mathrm{d}s} \Pr(t \leq T < t + \mathrm{d}s \mid T \geq t) \right\}$$

is not directly related to a survival function. This is, the functions

$$h_i(t \mid \mathcal{F}_i^y(t)) = \exp\left\{-\int_0^t h_i(t \mid \mathcal{F}_i^y(t))\mathrm{d}s\right\} \text{ and}$$
$$f_i(t \mid \mathcal{F}_i^y(t)) = h_i(t \mid \mathcal{F}_i^y(t)) \times S_i(t \mid \mathcal{F}_i^y(t)),$$

do not have the usual survival and density function interpretations. Due to these feature, the log-likelihood (2.16) based on $f(\cdot)$ and $S(\cdot)$ is not meaningful for endogenous covariates.

Additionally, the "true" covariate $m(t)$ is unobservable and instead a measurement with error $y(t) = m(t)+\varepsilon(t)$ is observed and available (as explained for latent variables in Section 1), where $\varepsilon(t)$ is a random error. This measurement error $\varepsilon(t)$ mainly refers to within subject variation, just as the repeated measures correlation behavior discussed in Section 2.1.

One possible way to circumvent the challenges that endogenous time-varying covariates bring to analyze time-to-event data is by joint modelling the trajectory of $y_i(t)$ and the hazard rate of the event. Joint modelling longitudinal and time-to-event data provides a framework suitable for dealing with this challenge. In Section 2.1.3 we discussed the problem of missing data in longitudinal studies under MNAR mechanism, and pointed out to joint modelling the missing data process and the longitudinal outcome as a solution for this problem. Interestingly, by joint modelling longitudinal and time-to-event data is the same solution to both problems arising from MNAR in longitudinal studies and endogenous time-varying covariates in survival analysis models.

We introduce the joint modelling framework in Chapter 3.

Data simulation from survival analysis models with endogenous time-varying covariates

We discuss data simulation from a survival analysis model with endogenous time-varying covariates because it is the basis of simulation from a joint model of longitudinal and time-to-event data, discussed in Chapter 3.

Suppose we are interested in simulating data from the following survival analysis model

$$h_i\left(t \mid \mathcal{F}_i^y(t)\right) = h_0(t) \exp\left\{\mathbf{w}^\top \boldsymbol{\gamma} + \beta y_i(t)\right\}. \tag{2.32}$$

where $y_i(t)$ is an endogenous time-varying covariate.

The survival function is given by

$$\begin{aligned} S_i\left(t \mid \mathcal{F}_i^y(t)\right) &= \Pr\left\{T_i > t \mid \mathcal{F}_i^y(t)\right\} \\ &= \exp\left\{-\int_0^t h_i\left(s \mid \mathcal{F}_i^y(s)\right) \mathrm{d}s\right\} \\ &= \exp\left\{-\exp\{\mathbf{w}_i^\top \boldsymbol{\gamma}\} \int_0^t h_0(s) \exp\{\beta y_i(s)\} \mathrm{d}s\right\} \\ &= \exp\left\{-H_i\left(t \mid \mathcal{F}_i^y(t)\right)\right\}, \end{aligned} \tag{2.33}$$

where

$$H_i\left(t \mid \mathcal{F}_i^y(t)\right) = \exp\{\mathbf{w}_i^\top \boldsymbol{\gamma}\} \int_0^t h_0(s) \exp\{\beta y_i(s)\} \mathrm{d}s \tag{2.34}$$

is the cumulative hazard function.

Note that in this case, $H_i\left(t \mid \mathcal{F}_i^y(t)\right)$ and $S_i\left(t \mid \mathcal{F}_i^y(t)\right)$ depend on the whole trajectory of the time-varying covariate $y_i(t)$ up to time t.

We can use the inverse transform method. Let $\xi_i = S_i(t)$, where $\xi_i \sim \mathcal{U}[0, 1]$. From the relationship between the survival and the cumulative hazard functions, we have $H_i(t) = -\log S_i(t) \iff H_i(t) = -\log(\xi_i)$. Thus the time to event for subject i can be simulated from

$$T_i = H_i^{-1}\left(-\log(\xi_i) \mid \mathcal{F}_i^y(t)\right). \tag{2.35}$$

To simulate data with time-varying covariates we need to treat $y_i(t)$ as a function of time in the data simulation algorithm, taking into account that is a subject-specific process. One option is to set a general time function with subject-specific parameters, for instance, $y_i(t) = b_{i0} + b_{i1}t$, where $b_{i0}, b_{i1} \in \mathbb{R}$. Algorithm 2.2 describes the steps for simulating data from model (2.2).

Note that because the integrand in Equation (2.34) might involve complicated functions of time depending on the chosen baseline hazard and time function $y_i(t)$ this integral might not have a closed form and the inverse of the cumulative hazard might not be

Algorithm 2.2 Data simulation of a proportional hazards model with endogenous time-varying covariate $y_i(t)$.

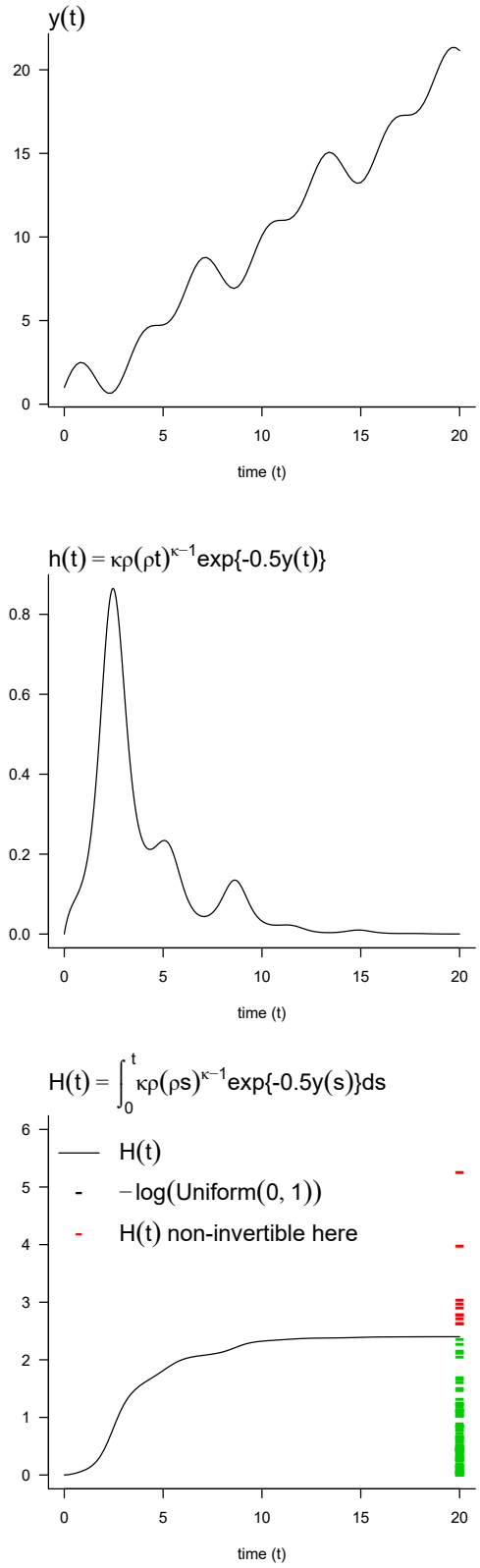
Decide on the sample size (n), the functional form of the baseline hazard $(h_0(t))$, the functional form of the time-varying covariate $(y_i(t))$, the values of the regression coefficients $(\boldsymbol{\gamma})$ and β, a simulation scheme for the vector of baseline covariates $(\mathbf{w}_i)$, and a censoring scheme, C_i, possibly a fixed administrative censoring $C_i = c$ for all i. Additionally, decide on the time points $(t_{ij}, i = 1, \ldots, n\ j = 1, \ldots, n_i)$ at which $y_i(t)$ will be simulated.
For $i = 1, \ldots, n$:

1: Sample $\xi_i \sim \mathcal{U}(0, 1)$.
2: Simulate T_i^*, the time-to-event as in Equation (2.35).
3: Set $T_i = \min(T_i^*, C_i)$ and $\delta_i = \mathbb{1}(T_i < C_i)$.
4: Simulate $y_i(t_{ij})$ from the chosen function $y_i(t)$ at time points t_{ij}, such that $t_{ij} \leq T_i$.

invertible analytically. Thus numerical integration and a numerical inverse for $H^{-1}(\cdot)$ will be required as indicated in Algorithm 2.1.

Additionally, while choosing the parameters of the baseline hazard $h_0(t)$, the time function of time-varying covariate $y_i(t)$ and the value of its regression coefficient β, it is a good practice to plot the cumulative hazard within the time interval of interest before simulating the data to make sure the cumulative hazard is invertible in the whole time interval of interest. Figure 2.2 illustrates a problem that might occur in the data simulation process. In this example for a unique subject the time varying covariate is defined as $y(t) = \beta_0 + \beta_1 f(t)$, where $\beta_0 = 2$, $\beta_1 = 1$, and $f(t) = t + \cos(2t) + \sin(t)$ is a function of time. The baseline hazard h_0 is governed by the $\text{Weibull}(\kappa, \rho)$ distribution. Notice that the cumulative hazard in the bottom panel has an asymptote at 2.5, so it is not invertible at the values with a marker (–), which correspond to $-\log(\xi_i) > 2.5$.

Appendix A.2 contains sample code to simulate data from the model of Equation (2.32), includes different options for the baseline hazard (Weibull, log-logistic, Gompretz, Makeham and bathtub) and random effects from the Gamma and Log-normal distributions. The integration of $h_i(t|\mathcal{F}_i^y(t))$ and inversion of $H(t|\mathcal{F}_i^y(t))$ are done numerically, so the code customized to any baseline hazard. The sample code is programmed with some time functions to specify the time-varying covariate $y_i(t)$.



In Figure 2.2 $y(t)$ is a time-varying covariate (top) of a survival analysis model (middle). Being a function of time, $y(t)$ affects the form of the hazard function.
The cumulative hazard is non-decreasing (bottom). However, with time-depending covariates $H(t)$ might exhibit an asymptotic behavior, making this function not invertible at values greater that the asymptote.

Figure 2.2: Top: Trajectory of the time-varying covariate $y(t)$; middle: Proportional hazards with baseline hazard $h_0(t) = \kappa\rho(\rho t)^{\kappa-1}$ and $y(t)$ as time-varying covariate; bottom: Cumulative hazard with - and - indicating where $H(t)$ should be inverted to simulate the time-to-event, $T = H^{-1}(-\log(\xi))$.

2.2.2 Recurrent events

The survival analysis models we have addressed so far are for situations in which the event occurs only once. There are other situations where the event of interest is recurrent thus individuals can experience it multiple times in the follow up period. These events are know in the literature as *recurrent events*. A recurrent event is not to be confused with *competing risks*. The former refers to sequential occurrences of events of the same type, while the latter is a situation where mutually exclusive events of different types are all susceptible to occur, but the occurrence of one precludes the occurrence of the others. Examples of recurrent events include sequences of tumor recurrences, infection episodes, recurrent falls, epileptic seizures, bleeding incidents, etc., and an example of competing risks are cancer and heart disease as possible causes of death.

Recurrent events models constitute the final building block of the joint models we explore in this thesis, so we dedicate this section to give an overview of the statistical analysis of recurrent events. Competing risks are out of the scope of this thesis.

The analysis of recurrent events impose additional challenges compared to single event situations. First, as in the LMM context for quantitative outcomes, there might be variation between individuals in their susceptibility to recurrent events, and the recurrence times within subjects may not be independent. Second, following the occurrence of an event, e.g. heavy bleed, may imply that subjects are not at risk of a subsequent event for a short period of time, so it is necessary to keep track of subjects who are at risk all along the follow up period to properly define the risk set (Collett, 2015).

The standard Cox model was addapted to handle exogenous time-varying covariates, similar variations of the Cox model based on the counting process formulation (Andersen *et al.*, 2012) provide a modelling framework for the analysis of recurrent event data. Three main models are mentioned in the literature for the analysis of recurrent events: The Andersen–Gill (AG) model (Andersen & Gill, 1982), the Prentice–Williams–Peterson model (Prentice *et al.*, 1981), and the Wei–Lin–Weissfeld model (Wei *et al.*, 1989). We will describe the AG model since we followed this approach for the analysis of recurrent falls within the joint modelling framework. For a comparison

of the three approaches refer to Ozga *et al.* (2018), and Cook & Lawless (2007) for a thorough discussion about the analysis of recurrent events.

Consider a sample of n individuals. The recurrent event process for subject i starting at $t = 0$, is $0 < T_{i1} < \ldots < T_{iK_i} < \infty$, where T_{ik} is the time for the k^{th} event. There is a counting process $\{N(t), t \geq 0\}$ associated to the recurrent event process that represent the cumulative number of events generated by the process, i.e. $N_i(t) = \sum_{k=1}^{\infty} \mathbb{1}(T_{ik} \leq t)$ is the number of events occurring over the interval $(0, t]$. More generally, $N(s, t) = N(t) - N(s)$ represents the number of events occurring over the interval $(s, t]$. The "proneness" to new events for an individual is described by the intensity process, $r_i(t)$, defined by

$$r_i(t) = \lim_{\mathrm{d}t \to 0} \left\{ \frac{1}{\mathrm{d}t} \Pr \left\{ N_i(t + \mathrm{d}t) - N_i(t) \geq 1 \mid \mathcal{F}_i^N(t) \right\} \right\}, \tag{2.36}$$

where $\mathcal{F}_i^N(t)$ is the history of the counting process up to time t. The intensity rate, $r_i(t)$, is a nonnegative function and depends only on the past history of N_i, i.e. the jumping times of $N_i(t)$ up to time t. Since the value of $r_i(t)$ determines the immediate probabilistic future of $N_i(t)$, it is natural to base statistical models on $r_i(t)$ by letting it depend also on covariates, as in the Cox model. The AG model is defined by Equation (2.37)

$$r_i(t) = \Delta_i(t) r_0(t) \exp \left\{ \mathbf{w}_i^\top \boldsymbol{\gamma} \right\}, \tag{2.37}$$

where

$r_0(t) =$ is the baseline intensity function,

$\mathbf{w}_i^\top = (w_{i1}, \ldots, w_{ip})$ is a p-vector of baseline covariates,

$\boldsymbol{\gamma}^\top = (\gamma_1, \ldots, \gamma_p)$ is a p-vector or regresson coefficients,

$\Delta_i(t)$: is a left continuous at risk process.

The at risk process, $\Delta_i(t)$, defined by

$$\Delta_i(t) = \begin{cases} 1 & \text{if subject } i \text{ is at risk at time } t \\ 0 & \text{otherwise} \end{cases}$$

remains at unity unless a subject temporarily ceases to be at risk in some time period,

or until the follow-up time is censored. A difference between the AG model and the Cox model is that in the latter only one event could be observed, so that $\Delta_i(t) = 0$ after the event has occurred. In the AG setting, events may be recurrent so $\Delta_i(t)$ is "reset" every time the event occurs and may take the value of either 0 or 1 whenever subject i is at risk for another recurrence of the event. In AG model, $\Delta_i(t)$ could work as a censoring indicator function.

Let T^*_{ik} denote the time for the k^{th} recurrent event of subject i, where $k = 1, \ldots, K_i$ and $i = 1, \ldots, n$. Assume a noninformative right censoring process, C_i, so we observe $T_{ik} = \min(T^*_{ij}, C_i)$ and the indicator $\delta_{ik} = 1$ if $T_{ik} = T^*_{ik}$ and 0 otherwise. Let $(t_{ik}, \delta_{ik}) = (T_{ik} = t_{ik}, \delta_{ik})$. The estimation of the model in Equation (2.37) is based on maximization of the partial likelihood function given by Equation (2.38).

$$\ell_p(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \sum_{k=1}^{K_i} \delta_{ik} \Big[\mathbf{w}_i^\top \boldsymbol{\gamma} - \log \Big(\sum_{l \in \mathcal{R}(t_{ik})} \exp\{\mathbf{w}_l^\top \boldsymbol{\gamma}\} \Big) \Big], \tag{2.38}$$

where the risk set, $\mathcal{R}(t) := \{l : t_{il} \geq t; l = 1, \ldots, K_i \text{ and } i = 1, \ldots, n\}$, is comprised by all those distinct event times that have not occurred by time t.

The assumption of within subject independent recurrence times can be relaxed and by adding a subject specific random effect, u_i, to accommodate the within subject event dependence, in a similar fashion as explained in Section 2.2.1. The recurrent event model with a random effect is described by Equation (2.39)

$$r_i(t \mid u_i) = u_i \Delta_i(t) r_0(t) \exp\left\{\mathbf{w}_i^\top \boldsymbol{\gamma}\right\}, \tag{2.39}$$

where u_i are independent random variables from a positive-valued distribution characterized by a unique parameter ϕ such that $\mathbb{E}(u_i) = 1$ and $\text{var}(u_i) = \phi$.

Assuming a parametric baseline intensity, the parameters to estimate of the model in Equation (2.39) are $\boldsymbol{\theta}^\top = (r_0, \boldsymbol{\gamma}, \phi)$, where r_0 are the parameters of a chosen functional form of the baseline intensity, $\boldsymbol{\gamma}$ are the regression coefficients and ϕ the variance parameter of the assumed distribution of the random effects. Estimation is based on

maximization of the likelihood function given by Equation (2.40).

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K_i} \int_0^{\infty} [r_i(t_{ik} \mid u_i)]^{\delta_{ik}} \exp\left\{ - \int_{t_{i(k-1)}}^{t_{ik}} r_i(t \mid u_i) \mathrm{d}t \right\} \mathrm{d}u_i \tag{2.40}$$

where $r_i(\cdot|u_i)$ is as defined in Equation (2.39).

Exogenous time-varying covariates, say $y_i(t)$, can be incorporated to the recurrent event models described in Equations (2.37) and (2.39) by adapting their respective partial log-likelihood and likelihood functions. The log-likelihood function $\ell_p(\boldsymbol{\gamma})$ of Equation (2.38) requires to redefine the risk set, $\mathcal{R}(t)$, by taking into account the times at which $y_i(t)$ is measured. The modification of the likelihood function $\mathcal{L}(\boldsymbol{\theta})$ of Equation (2.40) is in the same sense, by specifying appropriate integration limits taking into account the measurement times of $y_i(t)$ between consecutive events, i.e. $t \in (T_{i(k-1)}, T_{ik})$ and adding up these integrals.

The parameters of this recurrent event model are estimated by maximizing the likelihood function of Equation (2.40). Munda *et al.* (2012) proposed direct optimization of the likelihood function by transforming the random effect distribution with the Laplace transform and specifying a parametric baseline intensity. This is implemented in the `parfm()` function of the `parfm` package in R. Rondeau *et al.* (2003) maximize the likelihood by applying the Marquardt algorithm (Marquardt, 1963), which combines the Newton–Raphson and steepest descent algorithms. This estimation method is implemented in the `frailtyPenal()` function of the R package `frailtypack` with the option for either parametric Weibull or spline-approximated baseline intensity, following the penalized likelihood approach proposed by Joly *et al.* (1998). The R function `coxph()` function of the `survival` package can also be used to fit this recurrent event model for unspecified baseline intensity.

A final remark on recurrent events is that the analysis can be carried out in two timescales: (1) *gaps* or *waiting* times, and (b) *calendar* or *total* time scale. Gaps, also known as *inter-event times*, are the duration or time interval between two consecutive events, i.e. $I_{ik} = T_{ik} - T_{ik-1}$, $k = 1, \ldots, K_i$ with $t_0 = 0$. In the gap time representation, the time at risk for the k^{th} event is the time from the end of the $(k-1)^{\text{th}}$ event to the k^{th} event with t_0 denoting the start of the study. A common assumption is that gaps between

successive events are independent, i.e. individuals are "renewed" after each event. In the calendar time representation, the start of the at-risk period is not reset to zero but to the actual time since entry to the study.

The description of the AG model of Equations (2.37) and (2.39) is the calendar timescale. They can be expressed in the gap timescale by replacing the calendar times T_{ik} with the gaps I_{ik}. In the gaps timescale we model the time since the last event. In the calendar timescale we model the time since origin at which events occur.

Simulation of recurrent events

Consider the situation where individuals are followed for the times of occurrence of some recurrent event and a total time scale shall be used. We intentionally omit the subject subindex, i, to simplify the notation, but it must be clear that we refer to a single event process. We define T_k as the time from starting point 0 to occurrence of the k-th event. Let $N(t) = \#\{k, T_k \leq t\}$ denote the counting process representing the number of events experienced before time t. Assuming that prior events do not affect the risk for future events, the hazard process of $N(t)$ is given by Equation

$$r(t)\mathrm{d}t = \mathbb{E}[N(t + \mathrm{d}t) - N(t) \mid \mathcal{F}^N] \tag{2.41}$$

with $\mathcal{F}^N$ being the history up to time t. The cumulative intensity function is defined by

$$R(t) = \int_0^t r(x)\mathrm{d}x.$$

As recurrent events are naturally ordered, event times T_k can be derived from the inter-event times, $I_k := T_k - T_{k-1}$ with $T_0 = 0$ by $T_k = \sum_{j-1}^{k} I_j$. As the risk for events depends on total time, the distribution of an inter-event time depends on the time of the preceeding event unless we deal with the simple situation of constant hazards, i.e. when $T_k \sim \mathrm{Exp}(\rho)$. Therefore, we consider the conditional distributions of inter-event times given the time of the immediate preceeding event. Let $I_k \mid T_{k-1}$ denote these conditional inter-event times. The conditional hazard function, $\tilde{r}^k$ of $I_k \mid T_k = t$ can

be derived for $k > 1$ by

$$\begin{aligned}
\tilde{r}^k \mathrm{d}s &= P(s \leq I_k \leq s + \mathrm{d}t \mid I_k \geq s, T_{k-1} = t) \\
&= P(s \leq T_k - t \leq s + \mathrm{d}t \mid T_k - t \geq s, T_{k-1} = t) \\
&= P(s + t \leq T_k \leq s + t + \mathrm{d}t \mid T_k \geq s + t, T_{k-1} = t) \\
&= \mathbb{E}[\mathrm{d}N(s + t) \mid T_k \geq s + t, T_{k-1} = t] \\
&= r(s + t)\mathrm{d}(s),
\end{aligned}$$

and $\tilde{r}^1 = r(s)$. This means that the **conditional** hazard for the k^{th} recurrent event to occur at time s given that the $(k-1)^{\text{th}}$ event occurred at time t is the same as the **unconditional** hazard from the time elapsed between t and s, $s > t$.

Accordingly we can derive the cumulative hazard of $I_k \mid T_{k-1} = t$ for $k > 1$ by

$$\begin{aligned}
\tilde{R}^k(s \mid T_{k-1} = t) &= \int_0^s \tilde{r}^k(x \mid T_{k-1} = t)\mathrm{d}x \\
&= \int_0^s r^k(x + t)\mathrm{d}x \\
&= R(s + t) - R(t),
\end{aligned}$$

and $\tilde{R}^1(s) = R(s)$. This is, the **conditional** hazard for the k^{th} recurrent event to occur at time w given that the $(k-1)^{\text{th}}$ event occurred at time t is the difference between the **unconditional** cumulative hazard at t and $s + t$.

Note that the conditional hazards $\tilde{r}^k$ and $\tilde{R}^k$ do not depend on k. This follows from the assumption that the risk to experience events is not affected by previous events. Thus, we can define

$$\tilde{R}_t(s) := \tilde{R}^k(s \mid T_{k-1} = t) = R(s + t) - R(t). \tag{2.42}$$

For a specific time to recurrent event model, closed form solutions can be found for $\tilde{R}$, derived from r.

Simulations of recurrent events data based on this model are based on the conditional survival distribution $\tilde{S}(u \mid T_{k-1} = t) = \exp\{-\tilde{R}^k(u \mid T_{k-1} = t)\}$. For $k > 1$, the conditional random variable $\exp\{-\tilde{R}^k(I_k \mid T_{k-1} = t)\} \sim \mathcal{U}(0, 1)$. Therefore

$$I_k \mid T_{k-1} = t \sim \tilde{R}_t^{-1}\left(-\log(\xi)\right),$$

with $\xi \sim \mathcal{U}(0,1)$. For $k = 1$, $I_1 = T_1 \sim R^{-1}(-\log(\xi))$. The steps to simulate recurrent events are in Algorithm (2.3)

Algorithm 2.3 Simulation of recurrent events

1: Specify $r(t)$ as a function of total time.
2: Derive $\tilde{R}_t$ and $\tilde{R}_t^{-1}$.
3: For $k = 1$:
   (a) sample $\xi_1 \sim \mathcal{U}[0,1]$,
   (b) compute $t_k = R^{-1}(-\log(\xi_1))$
4: For $k = 2, \ldots, K$:
   (a) sample $\xi_k \sim \mathcal{U}[0,1]$,
   (b) compute $t_k = t_{k-1} + \tilde{R}_{t_{k-1}}^{-1}(-\log(\xi_k))$

For the Weibull(κ, ρ) distribution we have the following hazard, cumulative hazards and the inverse cumulative hazard,

$$\begin{aligned} r(t) &= \kappa\rho(\rho t)^{\kappa-1} \\ R(t) &= (\rho t)^{\kappa} \\ \tilde{R}_t(s) &= \rho^{\kappa}\left\{(t+s)^{\kappa} - t^{\kappa}\right\} \\ \tilde{R}_t^{-1}(s) &= \frac{1}{\rho}\left\{s + (\rho t)^{\kappa}\right\}^{1/\kappa} - t. \end{aligned}$$

Including baseline covariates and a random effect is straightforward to simulate data from the AG model of Equations (2.37) and (2.39). For instance, to simulate data with random effects u_i from the model of Equation (2.39), let $R_0(t) = \int_0^t r_0(s)\mathrm{d}s$ denote the baseline cumulative intensity function. Assuming we have derived the (inverse) conditional cumulative baseline hazard of inter-event times $\widetilde{R}_{0,t}(s)$ and $\widetilde{R}_{0,t}^{-1}(s)$, then the (inverse) conditional cumulative hazard of inter-event times corresponding to Equation

(2.39) can be derived for any realization $(\mathbf{w}_i, u_i)$ of $(\mathbf{w}, u)$:

$$\begin{aligned}\widetilde{R}_t(s) &= R(s+t) - R(t) \\ &= \{R_0(s+t) - R_0(t)\}\, u e^{\mathbf{w}^\top \boldsymbol{\gamma}} \\ &= \widetilde{R}_{0,t}(s) u e^{\mathbf{w}^\top \boldsymbol{\gamma}} \\ \widetilde{R}_t^{-1}(s) &= \widetilde{R}_{0,t}^{-1}\left(\frac{s}{u} e^{-\mathbf{w}^\top \boldsymbol{\gamma}}\right)\end{aligned}$$

Appendix A.3 contains sample code to simulate a data set with recurrent events and baseline covariates wit the Andersen–Gill model, with option to use random effects and several options for the baseline intensity (Weibull, log-logistic, Gompretz, Makeham and bathtub).

2.3 Prediction

Prediction means literally the stating beforehand of what will happen at some future (Aitchison & Dunsmore, 1975). In Statistics the concept of prediction is used in a broader sense. In this section we will explain the different purposes of making predictions in the context of longitudinal and survival analysis models. Even though an exhaustive taxonomy of types prediction in Statistics would be interesting it is out of our scope.

As introduced in Section 2.1, in longitudinal studies, data of the quantitative outcome $y_i(t)$ is collected for a sample of n subjects at discrete time point, t_{ij} for $j = 1, \ldots, n_i$ and $i = 1, \ldots, n$. Hence, we observe y_{ij}. Once a model is fitted to the data, say $\{\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{b}}_i\}$, it can be used to make different types of predictions. In general the form of predictions will be as shown in Equation (2.43).

$$\widehat{y}_i(t) = \mathbf{x}_i^\top(t)\widehat{\boldsymbol{\beta}} + \mathbf{z}_i^\top(t)\widehat{\mathbf{b}}_i. \tag{2.43}$$

The difference in what is meant by prediction will depend on whether the same sample used to fit the model is scored with the fitted model or a completely new sample, also the temporal frame, whether we refer to the past or the future relative to the

last observation time point, t_{in_i}. Note that predictions based on Equation (2.43) assume that $\mathbf{x}_i(t)$ and $\mathbf{z}_i(t)$ can be known at any time t when some of their elements are time-varying covariates, or otherwise predicted on their own by a separate prediction process. This is not a problem when all are baseline covariates and $time$ is the only time-varying element of $\mathbf{x}_i$ and $\mathbf{z}_i$. We identify the following types of prediction in longitudinal data.

1. Reconstruction of the past subject-specific longitudinal outcome profiles, i.e. for the continuous time interval $0 \leq t \leq t_{in_i}$. We call this **In-sample prediction**.
2. Based on observations $\mathbf{x}_i(t)$ and $\mathbf{z}_i(t)$ of a completely new sample, reconstruct the past longitudinal outcome profiles, i.e. for $0 \leq t \leq t_{in_i}$. We call this **Out-of-sample prediction**.
3. Estimate the future values based on past observations, whether In-sample or Out-of-sample, i.e. for $t_{in_i} < t \leq t^* < \infty$. We call this **forecasting** and it has a similar connotation as in the Time Series context.

With survival analysis data, the outcome is the tuple $\{t_i, \delta_i\}$ for $i = 1, \ldots, n$, this is the total follow up time and the indicator that tells whether the follow up was interrupted because the event is observed or censored. A useful way to characterize the outcome is in terms of the survival probability and predictions can be based on its estimate. Suppose $\widehat{\theta}$ contains all the parameter estimates of the fitted model. Then the estimated survival probability is

$$\widehat{S}_i(t) = S_i(t \mid \mathbf{w}_i(t), \widehat{\theta}), \tag{2.44}$$

where the specific functional form of (2.44) depends on several features, like the type of survival analysis model fit, the choices about the baseline hazard, the existence of time-varying covariates, if a random effect is in the model, if data are clustered, etc. Just as in the case of longitudinal data, in the presence of time-varying covariates, predictions based on (2.44) assume that the covariate vectors $\mathbf{w}_i(t)$ are available for all subjects at all times t or their predictions otherwise.

The types of predictions with survival analysis models can be classified in a similar way as in longitudinal data. Let τ denote the end of the study. We can distinguish the following types of prediction in survival analysis.

1. **In-sample prediction**. Estimate the individual survival probability from origin to the end of the study, i.e. $0 \leq t \leq \tau$ for the same sample used to fit the model.
2. **Out-of-sample prediction**. Provided the $\mathbf{w}_i(t)$ is available for a completely new sample different from the one whose data was used to fit the model, estimate the survival probability for $0 \leq t \leq \tau$.
3. **Forecasting** Even though we could extend the survival probability estimates for any time $t < \infty$, these probabilities are meaningless when we already know the subjects status at τ. Instead the **residual survival probabilities** provide with updated and more information regarding the likelihood of observing the event in the future. The residual survival probability is defined as the conditional probability of surviving time $s > \tau$ given survival up to τ, this is

$$\pi_i(s \mid \tau) = \Pr(T_i \geq s \mid T_i > \tau, \cdot) = \frac{S_i(s \mid \cdot)}{S_i(\tau \mid \cdot)}, \tag{2.45}$$

   where conditioning on covariates and estimates is omitted but assumed. This topic is addressed again in Section 3.2.2 for joint modelling longitudinal and time-to-event data. Residual survival probabilities can be estimated In-sample and Out-of-sample provided covariate data of $\mathbf{w}_i(t)$ is available for $t > \tau$.

A final comment about predictions with survival analysis models is that there are other ways to use the fitted model to predict the time-to-event outcome:

- For a given percentile of the probability of survival (p) we can obtain the estimated survival time

$$\widehat{T}_{i(p)} = S_i^{-1}(p \mid \mathbf{w}_i(t), \widehat{\theta}).$$

  Usually, $p = 0.5$ is a relevant choice giving the estimated median survival time.
- For a given time point, say t^*, it might be of interest to know how well the model separates events from non-events at different values of $\widehat{S}_i(t)$ as threshold. This dynamic classification ability of survival analysis models can be analyzed by extending the Receiver Operating Characteristic Curve (ROC) methodology to a dynamic model (Rizopoulos, 2011).

Predictions of the time-to-event and classification of events is out of the scope of this thesis, but we refer the reader to (Rizopoulos, 2011) for further details.

In Section 2.3.1 we address how to assess the accuracy of predictions made with LMM and survival analysis models taking $\widehat{y}_i(t)$ and $\widehat{S}_i(t)$ as the predicted outcomes. The reason we include this topic is because in Chapter 5 we use these prediction assessment methods for assessing the accuracy of predictions made by joint modelling longitudinal and time-to-event data.

2.3.1 Accuracy of prediction

The distance between the fitted value of the outcome and the actual outcome is central to quantify overall model performance. Typically the squared distances are used: the mean squared error, $(y - \widehat{y})^2$, for continuous outcomes and the Brier score (Brier, 1950), $(\mathbb{1}(Y = 1) - \widehat{\Pr}(Y = 1))^2$ (with $\mathbb{1}(z) = 1$ if z is true and 0 otherwise) for binary outcomes. These distances between observed and predicted outcomes are related to the concept of "goodness-of-fit" of the model, with better models having smaller distances between predicted and observed outcomes. The main difference between goodness-of-fit and predictive performance of a model is that the former is evaluated in the same data used to fit the model, while assessment of the latter requires either a new dataset or cross-validation (Steyerberg *et al.*, 2010).

Accuracy of prediction in linear mixed models

In general, the **Mean squared error** (MSE) of an estimator $\widehat{\theta} = f(X)$ with respect to an unknown parameter θ is defined by

$$\mathrm{MSE}(\theta) = \mathbb{E}_\theta \left(\widehat{\theta} - \theta\right)^2.$$

The MSE is a property of an estimator, so it is a function of the data. The MSE can be written in terms of the bias and variance of the estimator as

$$\mathrm{MSE}(\theta) = \left(\mathrm{bias}(\widehat{\theta})\right)^2 + \mathrm{var}\left(\widehat{\theta}\right).$$

This decomposition allows to assess how accurate and precise estimators are given their MSE.

The prediction MSE is defied in similar terms as the squared difference between data points y_i and their predictions $\widehat{y}_i$. In the context of longitudinal data that we have n_i data points for each subject in a sample of size n, we define the prediction MSE as described by Equation (2.46).

$$\begin{aligned} \text{MSE} &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}(y_{ij}-\widehat{y}_{ij})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\text{MSE}_i. \end{aligned} \tag{2.46}$$

This is, the MSE with longitudinal data averages across the n subjects of the sample the prediction MSE of the subject-specific profiles.

Accuracy of prediction in survival analysis models

Several measures have been proposed to evaluate the predictive ability of survival analysis models, although so far there is no consensus of opinion on which is the most appropriate one (Choodari-Oskooei *et al.*, 2012a). These measures attempt to assess different aspects of the predictive ability of survival models and they have been systematically studied and compared in order to understand their strengths and shortcomings. For instance, (Choodari-Oskooei *et al.*, 2012a) and (Choodari-Oskooei *et al.*, 2012b) classified these measures in three categories: explained variation, explained randomness and predictive accuracy. They compared the measures of each category by applying a set of criteria that a measure of predictive ability should possess in the context of survival analysis including (1) independence from censoring; (2) monotonicity; (3) interpretability; (4) robustness against influential observations; (5) ability of the model to account for the variability/uncertainty of the outcome; and (6) ability of the model to predict the time-to-event outcome. Rahman *et al.* (2017) classified the measures with a different taxonomy: overall performance, discrimination and calibration, and compared the measures of each group.

The performance of a mathematical model predicting a dichotomous outcome is typically assessed by quantifying their *discrimination* and *calibration* (Pencina & D'Agostino,

2004; Rahman *et al.*, 2017; Steyerberg *et al.*, 2010). Discrimination quantifies the ability of the model to correctly classify subjects into one of the two categories (events and non-events for time-to-event data). A model is said to have perfect discrimination if it classifies each subject to the group they truly belong to. The area under the receiver operating characteristic curve (ROC) (Fawcett, 2006) is one of the most popular measures of discrimination (Pencina & D'Agostino, 2004). In the context of time-to-event data, some discrimination measures are:

- (Royston & Sauerbrei, 2004), D. This measure computes the log hazard ratio between two equal sized prognostic groups formed by dichotomizing the prognostic index at its median. It assumes that the prognostic index is normally distributed.
- Harrell *et al.* (1996) proposed a rank order statistic for an arbitrary time by comparing the prognostic index (linear predictor) and the observed event times of every uncensored pairs of subjects. It is interpreted as the probability that the model correctly ranks the estimated risks of a randomly selected pair of subjects given their observed event times.
- Gönen & Heller (2005) proposed a measure as the reverse of that proposed by Harrell *et al.* (1996), i.e. the probability that the observed event times of a randomly chosen pair is correctly ordered given their estimated risks.

Calibration assesses how closely the predicted probabilities agree numerically with the actual outcomes, $\mathbb{1}(Y = 1)$. Calibration measures include the Brier score (Brier, 1950), the calibration slope (van Houwelingen, 2000). According the the classification in Choodari-Oskooei *et al.* (2012a), the Brier score is a predictive accuracy measure.

In this thesis, we focus on the Integrated Brier Score (IBS) (Graf *et al.*, 1999). We decided to use the IBS, since it is based on a squared loss function analogous to the MSE and it provides a summary over a relevant time interval $(0, t^*)$ rather than a specific time point.

Brier Score and Integrated Brier Score Brier (1950) proposed a scoring rule for binary outcomes as the mean squared difference between the actual outcome and its estimated probability. In the survival analysis context, the **Brier score** for a particular subject, i, is defined as the squared differences between the survival status at a given evaluation time t^*, $\mathbb{1}(T_i > t^*)$, and the estimated probability of survival up to this time given a set of covariates $\mathbf{w}_i$, $\widehat{S}_i(t^* \mid \mathbf{w}_i)$, Graf *et al.* (1999), Steyerberg *et al.* (2010), van Houwelingen & Putter (2011). Figures 2.3a and 2.3b show the predicted survival curves for the subjects of a simulated data set of sample size $n = 100$, and the region of interest for the BS for a specific subject of such data. In practice, the BS of each subject in the analyzed sample would look similar to Figure 2.3b, with the green line representing the distance between $(\widehat{S}_i(t_k^*))$ and $\mathbb{1}(T_i > t_k^*)$ and $\text{BS}(t^*)$ being the squared of the length of this line. For instance, for the subject represented in Figure 2.3b, $\text{BS}(1.4) = 0.07$.

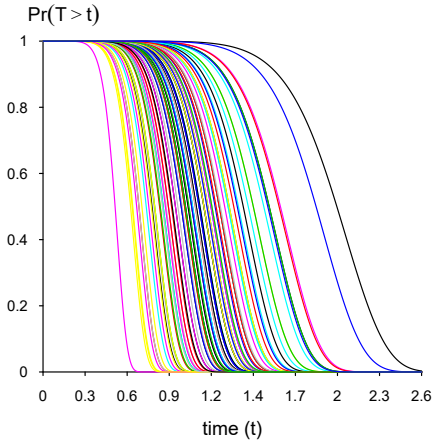
The Brier score at t^* in a sample of n subjects is the mean of the BS of all the subjects in the data set:

$$\text{BS}(t^*) = \frac{1}{n}\sum_{i=1}^{n} \text{BS}_i = \frac{1}{n}\sum_{i=1}^{n} \Big(\mathbb{1}(T_i > t^*) - \widehat{S}_i(t^* \mid \mathbf{w}_i)\Big)^2. \tag{2.47}$$

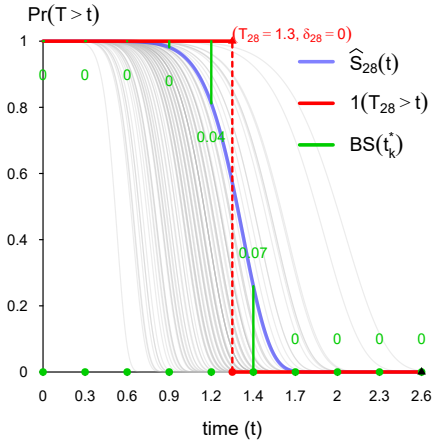
In survival analysis it is more informative to assess the predictive ability of the model for a relevant time interval (for instance from the beginning to the end of a study, from diagnosis up to a relevant threshold after treatment) rather than restricting to specific fixed time points. Graf *et al.* (1999) introduced an *integrated Brier score* by extending the idea of the BS to cover a time interval. The IBS is obtained by integrating the BS over time for $t \in [0, t^*]$. The IBS for a sample of n subjects is the mean of the individual IBS and is computed by

$$\begin{aligned} \text{IBS}(t^*) &= \frac{1}{n}\sum_{i=1}^{n} \text{IBS}_i(t^*) \\ &= \frac{1}{n}\sum_{i=1}^{n} \int_0^{t^*} \Big(\mathbb{1}(T_i > t) - \widehat{S}_i(t \mid \mathbf{w}_i)\Big)^2 \mathrm{d}t. \end{aligned} \tag{2.48}$$

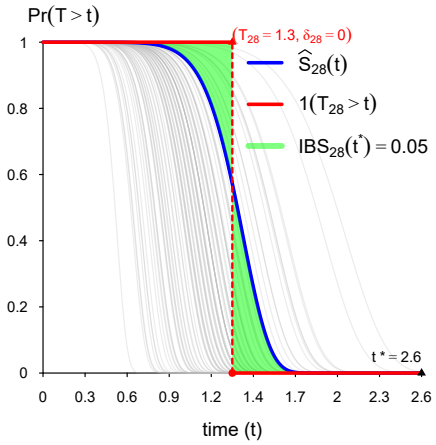
Graf *et al.* (1999) suggested integrating the BS with respect to some weight function



(a) Survival curves $(\widehat{S}_i(t))$ for all the subjects of a simulated data set.



(b) Brier score. — is the survival curve of an instance of the sample; — is the indicator $\mathbb{1}(T_i > t_k^*)$ which takes the value of 1 as long as subject i is alive and then drops to 0. Each | represents the difference between $(\widehat{S}_i(t_k^*))$ and $\mathbb{1}(T_i > t_k^*)$ at time points $t_k^* = 0, 0.3, \ldots, 2.6$, with $\mathrm{BS}(t_k^*)$ being the square of this difference.



(c) Inegrated Bier score. The green shaded region in is the area where the Integrated Brier score is computed in the time interval $(0, t^*)$

Figure 2.3: (a) Survival curves, (b) Brier Score, and (c) Integrated Brier score.

$\mathcal{W}(t)$. That is

$$\text{IBS}(t^*) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{t^*}\Big(\mathbb{1}(T_i > t) - \widehat{S}_i(t \mid \mathbf{w}_i)\Big)^2 \mathrm{d}\mathcal{W}t, \tag{2.49}$$

considering $\mathcal{W}(t) = t/t^*$ and $\mathcal{W}(t) = (1-\widehat{S}(t))/(1-\widehat{S}(t^*))$ as two options leading to

$$\text{IBS}(t^*) = \frac{1}{t^*}\int_0^{t^*}\frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{1}(T_i > t) - \widehat{S}_i(t \mid \mathbf{w}_i)\Big)^2 \mathrm{d}t \quad \text{and} \tag{2.50}$$

$$\text{IBS}(t^*) = \frac{1}{1-\widehat{S}(t^*)}\int_0^{t^*}\frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{1}(T_i > t) - \widehat{S}_i(t \mid \mathbf{w}_i)\Big)^2 f(t)\mathrm{d}t. \tag{2.51}$$

In this thesis we used the IBS of Equation (2.48) because it is more intuitive. The green region of Figure 2.3c corresponds the area where the IBS is being calculated for the same subject.

Censoring is usually accommodated for in two steps (Gerds & Schumacher, 2006; Graf *et al.*, 1999; van Houwelingen & Putter, 2011) :

- Deleting observations whose event status cannot be determined by the time point of interest, i.e. all observations censored.
- Weighting the observations by the probability of not being censored by the time point of interest.

Recall that with time-to-event data we observe $T_i = \min(T_i^*, C_i)$ and $\delta_i = \mathbb{1}(T_i^* \leq C_i)$ for each individual. Suppose it is of interest to compute the BS at time point t^*. For this fixed time point, t^*, the contributions to the BS of each individual can be split into three categories:

- Category 1: $T_i \leq t^*$ and $\delta_i = 1$. For these individuals the event occurred by time t^*, so their event status after t^* is $\mathbb{1}(T_i > t^*) = 0$. Individuals in this category contribute to the Brier score with $\Big\{0 - \widehat{S}(t^*|\mathbf{w}_i)\Big\}^2$.

- Category 2: $T_i > t^*$ and ($\delta_i = 1$ or $\delta_i = 0$). These are individuals still at risk of the event at time t^*, i.e. they are event-free and uncensored so their event status after t^* is $\mathbb{1}(T_i > t^*) = 1$, and their contribution to the Brier score is $\left\{1 - \widehat{S}(t^*|\mathbf{w}_i)\right\}^2$.
- Category 3: $T_i \leq t^*$ and $\delta_i = 0$. These are individuals censored by time t^* so their event status after t^* is unknown and do not contribute to the Brier score.

In order to accommodate censoring, the individual contributions are weighted, with weights being the probability of not being censored by the last observed time. Denote by $G(t) = \Pr(T_i > t)$ the probability of not being censored by time t. Thus the contributions to the Brier score of individuals in category 1 get the weight $\frac{1}{\widehat{G}(T_i)}$, and those in category 2 $\frac{1}{\widehat{G}(t^*)}$. The estimates $\widehat{G}(t)$ are the Kaplan–Meier estimates (Kaplan & Meier, 1958) of the censoring process (that is the KM for $\{t_i, 1 - \delta_i\}$). Therefore, the Brier score adjusted for censoring is given by,

$$\begin{aligned}
\mathrm{BS}^C(t^*) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\left\{0 - \widehat{S}(t^*|\mathbf{w}_i)\right\}^2 \frac{\mathbb{1}(T_i \leq t^*, \delta_i)}{\widehat{G}(T_i)} + \right. \\
&\qquad \left. \left\{1 - \widehat{S}(t^*|\mathbf{w}_i)\right\}^2 \frac{\mathbb{1}(T_i > t^*)}{\widehat{G}(t^*)}\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\left\{\mathbb{1}(T_i > t^*) - \widehat{S}(t^*|\mathbf{w}_i)\right\}^2 \frac{\mathbb{1}(T_i \leq t^*)\delta_i}{\widehat{G}(T_i)} + \right. \\
&\qquad \left. \left\{\mathbb{1}(T_i > t^*) - \widehat{S}(t^*|\mathbf{w}_i)\right\}^2 \frac{\mathbb{1}(T_i > t^*)}{\widehat{G}(t^*)}\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{BS}_i(t^*) W_i(t^*, \widehat{G}, T_i, \delta_i), \qquad (2.52)
\end{aligned}$$

where

$$W_i(t^*, \widehat{G}, T_i, \delta_i) = \frac{\mathbb{1}(T_i \leq t^*)\delta_i}{\widehat{G}(T_i)} + \frac{\mathbb{1}(T_i > t^*)}{\widehat{G}(t^*)} \qquad (2.53)$$

and

$$\widehat{G}(t) = \prod_{i:t_i \leq t}\left(1 - \frac{c_i}{\mathcal{C}(t)}\right),$$

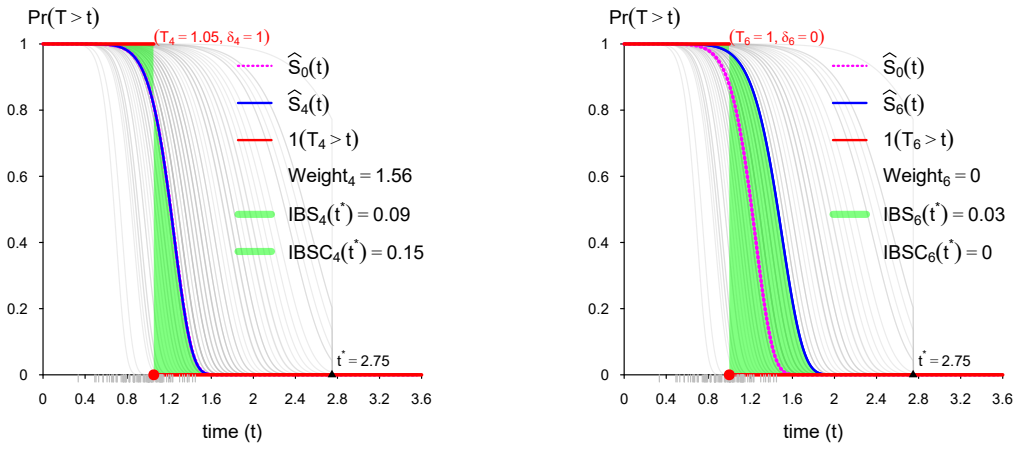
with c_i the number of censored individuals and $\mathcal{C}(t)$ the size of the set of subjects at

risk of censoring at time t.

The IBS can be weighted by $W_i(t^*, \widehat{G}, T_i, \delta_i)$ in a similar fashion to adjust for censoring. In this approach, censored observations will contribute their estimated event-free probabilities to the integrand up to the point t where the censoring occurs

$$\begin{aligned}\mathrm{IBS}^C(t^*) =& \frac{1}{n}\sum_{i=1}^{n}\int_0^{t^*}\Big(\mathbb{1}(T_i > t) - \widehat{S}_i(t \mid \mathbf{w}_i)\Big)^2 W_i(t^*, \widehat{G}, T_i, \delta_i)\mathrm{d}F(t) \\ =& \frac{1}{n}\sum_{i=1}^{n}\mathrm{IBS}_i(t^*)W_i(t^*, \widehat{G}, T_i, \delta_i). \end{aligned} \tag{2.54}$$

Figure 2.4 shows the IBS diagram of of two subjects from simulated data (left: observed; right: censored). The contribution to the IBS of the subject that observed the event has weight 1.56, and the censored of 0.



(a) Observed

(b) Censored

Figure 2.4: — Integrated Brier score (IBS) diagram of two subjects of a fictitious sample: (a) subject with observed event time and (b) subject with censored event time. The contribution to $\mathrm{IBS}(t^*)$ of the subject with observed event time has weight 1.56, and the subject with censored time has weight 0.00.

Appendix A.4 has sample code to compute the $\mathrm{BS}(t^*)$ and $\mathrm{IBS}(t^*)$ for the Cox model with baseline covariates and with an endogenous time-varying covariate, for an arbi-

trary time t^*. It contains also a function to weight individual contributions to accommodate censoring.

2.4 Penalized Likelihood Methods

Suppose we observe the data $(x_i, y_i),\ i = 1, \ldots, 5$, represented by • in Figure 2.5, and that y is a noisy version of the true underlying pattern function of x, $f(x)$. This is,

$$y_i = f(x_i) + \varepsilon_i,$$

where $\mathbb{E}(\varepsilon_i) = 0$ and $\text{var}(\varepsilon_i) = \sigma^2$.

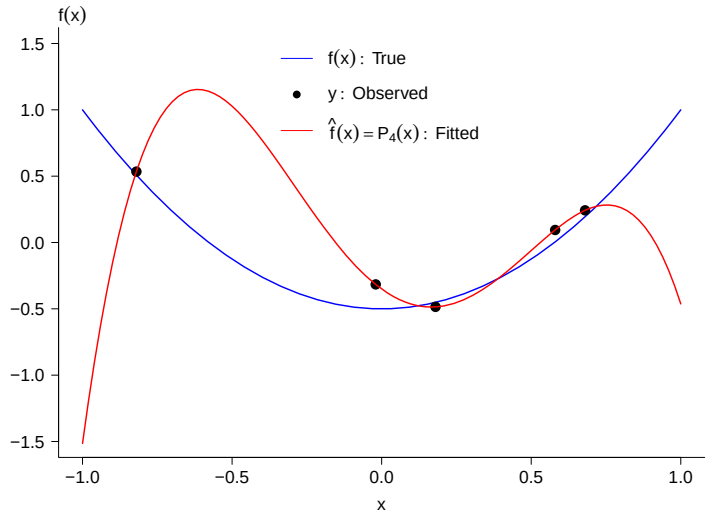


Figure 2.5: By fitting a fourth degree polynomial (—) to the observed data (•) the predictions equal the observed values at each x_i, but can be too far away from the true underlying pattern (—).

Suppose a fourth degree polynomial,$f(x_i) = f(x_i, \boldsymbol{\beta})$, is fitted to the data, $\hat{f}(x_i)$, to represent the x_i, y_i relationship and the polynomial coefficients are estimated by

minimizing the MSE,

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i, \boldsymbol{\beta}))^2.$$

With only five data points, $\hat{f}(x_i)$ fits the data perfectly, as shown by — in Figure 2.5. The fit is perfect in the sense that the predictions are unbiased since the predictions equal the observed values at the observation times, but the variance is large. Apparently, $\hat{f}$ is aiming for the observed data, y_i, and not the true underlying pattern, — $f(x)$. This is known as *overfitting* and it is said that the model overfits the data. By introducing a penalty to the objective function ($\lambda > 0$),

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i, \boldsymbol{\beta}))^2 + \lambda\sum_{j=1}^{4}\beta_j^2,$$

it is possible to obtain a more parsimonious prediction. By penalizing the objective function we obtain biased predictions, but with smaller error. To see this, consider the decomposition of the expected prediction error (EPE) of a regression fit $\hat{f}(x)$ at an input point $x = x_0$:

$$\begin{aligned}\text{EPE} &= \mathbb{E}\left\{\left[Y - \hat{f}(x_0)\right]^2\right\} \\ &= \underbrace{\mathbb{E}\left\{[Y - f(x_0)]^2\right\}}_{\text{irreductible error } (\sigma^2)} + \underbrace{\left(\mathbb{E}\{\hat{f}(x_0)\} - f(x_0)\right)^2}_{\{\text{bias}\hat{f}(x_0)\}^2} + \underbrace{\mathbb{E}\left\{\left[\hat{f}(x_0) - \mathbb{E}\{\hat{f}(x_0)\}\right]^2\right\}}_{\text{var}\{\hat{f}(x_0)\}}\end{aligned}$$

The first term is the variance of the outcome around its true mean, and cannot be avoided no matter how well we estimate $f(x_0)$, unless $\sigma^2 = 0$. The bias is the magnitude by which the average of the estimate differs from the true mean. The last term is the variance of the estimate, i.e. the expected squared deviation of $\hat{f}(x_0)$ around its mean. The mean squared error is defined as the sum of the last two terms:

$$\text{MSE} = \left\{\text{bias}(\hat{f})\right\}^2 + \text{var}(\hat{f}).$$

Typically the more complex we make the model, $\hat{f}$, the lower the bias, but the higher the variance, which might result in large MSE (Seber & Lee, 2012).

Figure 2.6 shows that as we increase the value of the penalty (left to right and top to bottom) we get more parsimonious predictions until eventually the prediction is flat. It looks that from the six chosen values for the penalty, when $\lambda = 0.1$ the predictions are closest to the true underlying pattern.

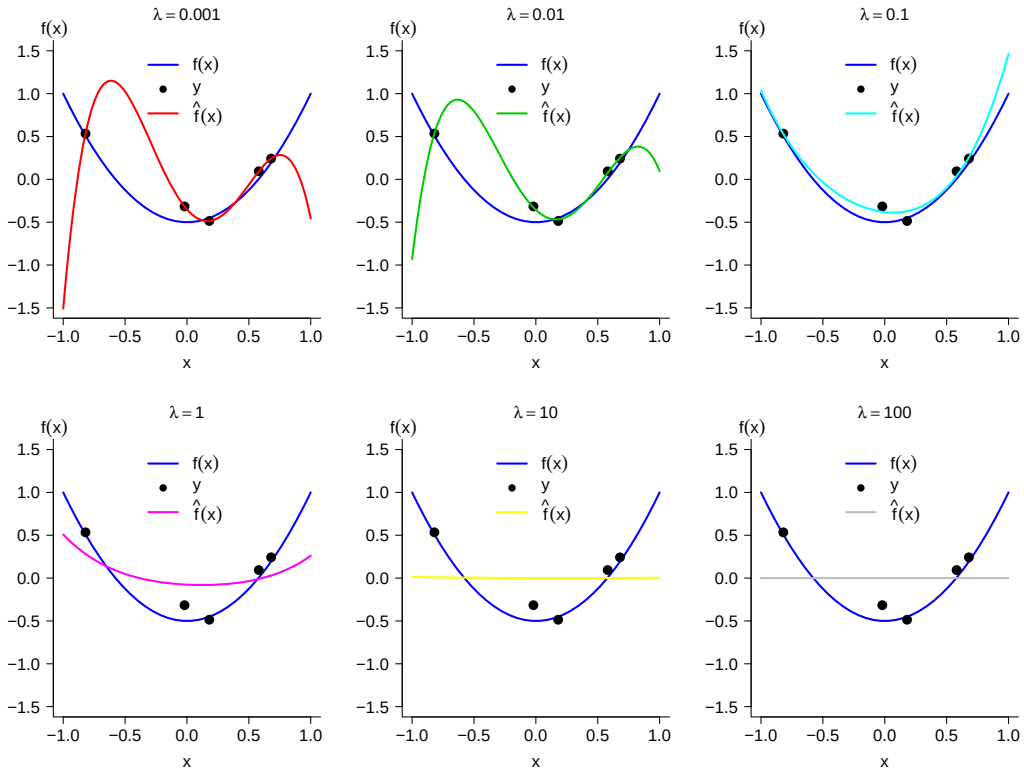


Figure 2.6: Fourth degree polynomial fit obtained by penalizing the objective function for six values of the penalty (from left to right and top to bottom): $\log_{10}(\lambda) = (-3, -2, -1, 0, 1, 2)$.

Penalization is a method for circumventing problems in the stability of parameter estimates that arise when the likelihood is relatively flat, making determination of the MLE difficult by means of standard or profile approaches. Penalization is also known as shrinkage, semi-Bayes, or partial-Bayes estimation. It can be viewed as a method for introducing some tolerable degree of bias in exchange of reduction in the variability of parameter estimates Cole *et al.* (2014).

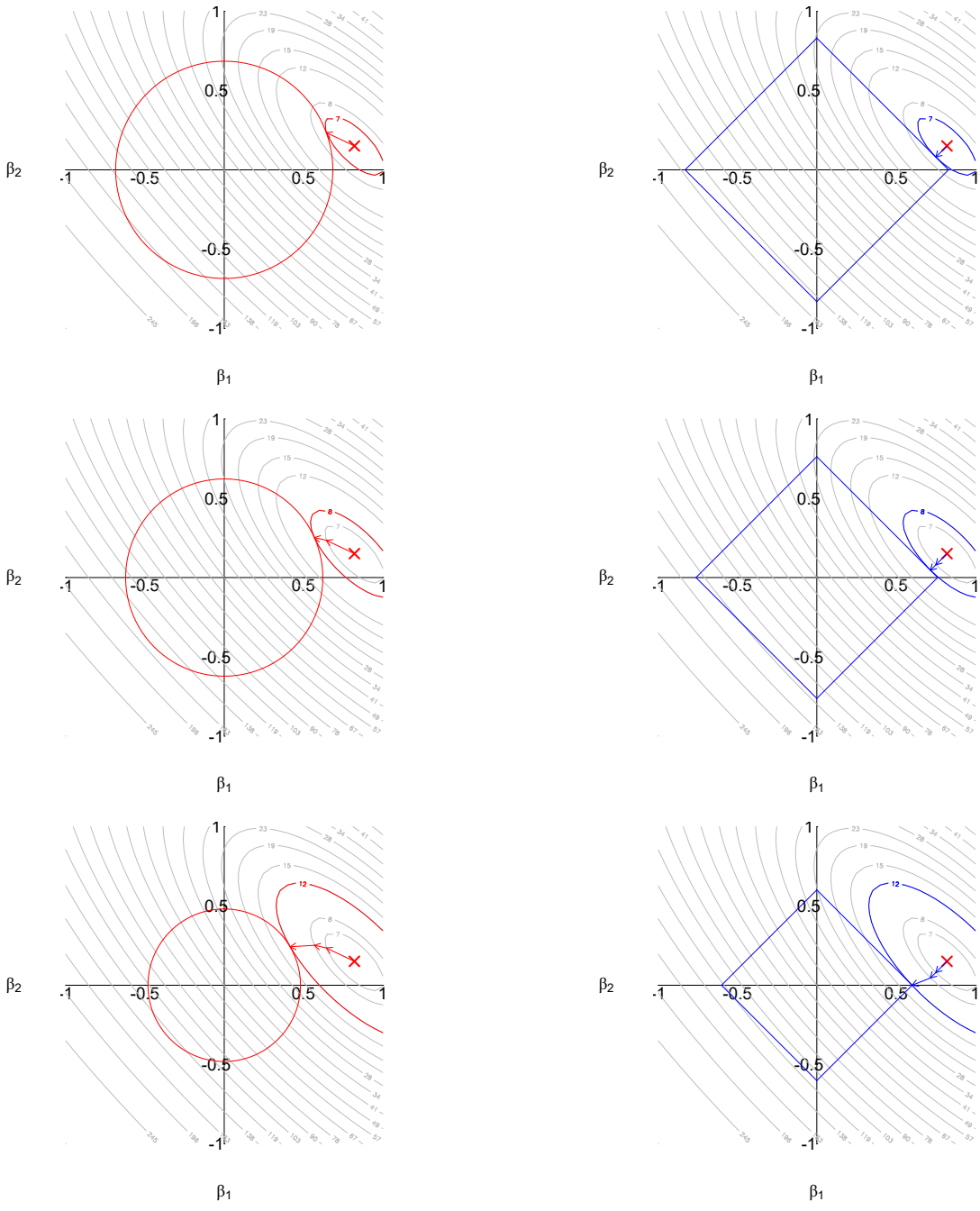
Maximizing a penalized likelihood reformulates the free optimization setting of maximum likelihood estimation into a constrained optimization one, where the constraint or penalty impose boundaries in the parameter space, i.e. it restricts the possible parameters' values. The immediate consequence of maximizing a penalized likelihood is that parameter estimates are no longer unbiased due to the constrain imposed to the parameter space, but such estimates have smaller standard errors. By imposing a penalty to the likelihood we aim to obtain estimates with much smaller standard errors in return of sacrificing a small amount of bias.

2.4.1 Ridge and LASSO penalties

Choosing the form of the penalty or constraint for the objective function has different effects on the estimates. The most common penalties are L_2 and L_1-norm, which impose quadratic and linear constrains respectively. L_2-norm is called Ridge and L_1-norm is the Least Absolute Shrinkage and Selection Operator (LASSO).

Figure 2.7 illustrates the effect of imposing Ridge and LASSO in the likelihood maximization problem with a sequence of contour plots of the likelihood for (β_1, β_2) and the corresponding penalty for different levels of the contour. Notice on the top plots of Figure 2.7a the arrow indicates the direction in which the estimates are shrinking corresponding to the tangent of the penalty and the contour at a level of 0.05. As the level of the contour increases, the radius of the circumference described by the quadratic penalty gets smaller and consequently the coefficients estimates. The Ridge penalty shrinks more the coefficients associated to the covariates with larger variances, but none of them gets as small as zero.

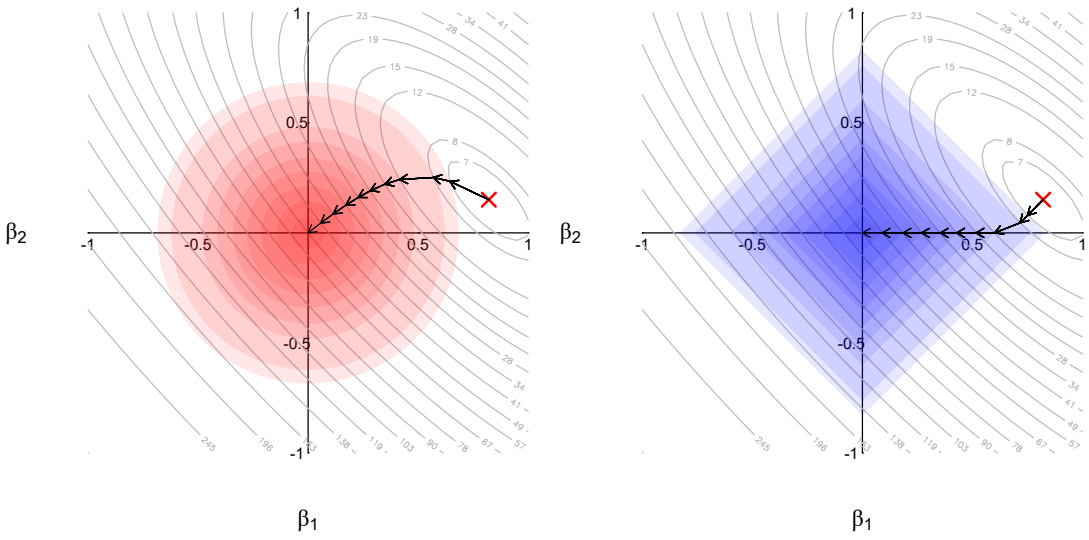
On the other hand, the LASSO penalty in Figure 2.7b shrinks the coefficients in direction to one of the axis. With the LASSO penalty some coefficients will be shrunk completely to zero and this is why it is a "selection operator". By imposing the LASSO penalty to the likelihood function (in the correct amount) it is possible to perform data-driven variable selection avoiding the problem of overfitting. Figures 2.8a and 2.8b illustrate the whole shrinkage path of Ridge and LASSO respectively. Note that LASSO aims for a corner solution of the constrained optimization problem while Ridge aims for an optimal linear combination of predictors or covariates.



(a) Ridge (b) Lasso

Figure 2.7: (a) L_2-norm penalty. The circumference represents the quadratic constraint to the optimization problem. The point of the circumference tangent to the smallest level curve represents the bias we sacrifice to reduce variance. (b) L_1-norm penalty. The rhomboid represents linear constraint to the optimization problem. The point of the edge tangent to the smallest level curve represents the bias we sacrifice to reduce variance.

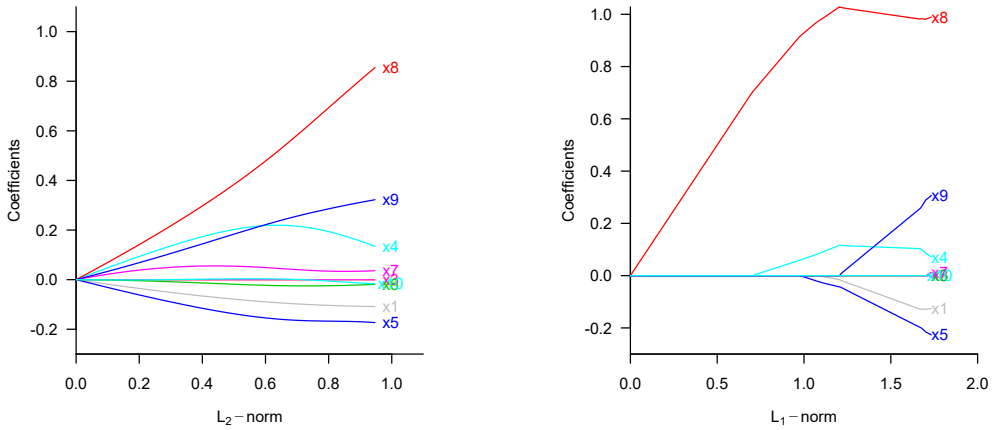
Finally, extending the number of possible predictor to $x_1, \ldots, x_8$ we have eight coefficients to estimate and Figures 2.9a and 2.9b show the shrinking process of the coefficiens as the value of the penalty gets closer to zero. Note that with the Ridge penalty all the coefficients shrink to zero but none of them is shrunk completely. On the other hand, LASSO completely shrinks the coefficients at a different value of the penalty.



(a) Ridge

(b) Lasso

Figure 2.8: (a) L_2-norm penalty; (b) L_1-norm penalty. The arrows show path in which the coefficients are being shrunk, with each arrowhead indicating the point where the penalty is tangent to the level curve



(a) Ridge (b) Lasso

Figure 2.9: (a) L_2-norm penalty; (b) L_1-norm penalty. As the norm decreases, the coefficients are shrunk towards zero at different rates. For a given value of the norm, Ridge regression results in a linear combination of all the coefficients, and LASSO in a subset of non-zero coefficients and another subset of coefficients completely shrunk to zero.

For a thorough discussion about regularization and shrinkage methods refer to Hastie *et al.* (2009).

2.5 Causal Inference

A randomized trial is a study in which a number of *similar* people are randomly assigned to 2 (or more) groups to test a specific drug, treatment or other intervention. One group (the experimental group) has the intervention being tested, the other (the comparison or control group) has an alternative intervention, a dummy intervention (placebo) or no intervention at all. The groups are followed-up to see how effective the experimental intervention was. Outcomes are measured at specific times and any difference in response between the groups is assessed statistically. The contrast of outcomes among the different treatment groups renders the causal effect of treatment. Randomized trials are considered the gold standard for investigating the effect of a *treatment* (or *exposure*) on an outcome.

In a properly designed randomized trial, before treatment allocation subjects are exchangeable regarding to which treatment level they can be assigned to, i.e. the potential outcome of the subjects should not differ systematically due to treatment allocation. This means, for instance, that the outcome of an experiment in which subject 1 is allocated to treatment A & subject 2 is allocated to treatment B should be indistinguishable from the experiment in which subject 1 is allocated to treatment B & subject 2 is allocated to treatment A. Immediately after subjects are randomly allocated to different treatment levels, the only difference among subjects must be the treatment group they are assigned to.

From the description of a randomized trial there are two important points that need to be emphasized because they render direct comparison between treatment groups possible, being key to causal effect estimation.

- Subjects being *similar* means that all their known and observed characteristics do not interfere with the effect of treatment. The existence of at least one characteristic that could potentially influence the effect of treatment on the outcome is known as *confounding* in the causal inference literature. Section 2.5.1 contains a broader discussion about confounding.
- There might be, unidentified or identified but unmeasured confounders, and that is why randomized treatment allocation is fundamental in this kind of studies. By

randomizing treatment allocation any potential link between such unknown/unmeasured confounders and treatment is broken, and the only difference between the subjects is the treatment level they are assigned to.

These two ideas together imply that in a randomized trial, potential outcomes are independent of confounders, which is a probabilistic statement about the relation between potential outcomes and confounders. The consequence of independence between potential outcomes and confounders is that after treatment has been allocated and subjects have been followed-up, the outcome differences among the treatment groups must be due the effect of treatment.

Often the data available for analysis do not come from randomized trials, but from observational studies, and the possible confounding structures can be complicated. Since randomization of treatment is not feasible in observational studies, it is rather used the idea of "controlling" for all possible sources of treatment-outcome confounding to be able to estimate the effects of treatment. Causal inference is the theoretical framework that attempts to articulate in mathematical language hypotheses about the confounding structures, and to state the necessary assumptions in order to control for confounders and move from conclusions about association to conclusions about causation.

To introduce the ideas in this section, consider the motivating example from Pearl *et al.* (2016). Imagine observational data from a study in which measures weekly hours of exercise and cholesterol level in people of different age. Plotting cholesterol as a function of hours of exercise we obtain the plot on the left panel of Figure (2.10), which exhibits a positive trend of cholesterol level with respect to hours of exercise, i.e. with more exercise the cholesterol level increases. This contradicts our common belief about the benefits of exercise on our health.

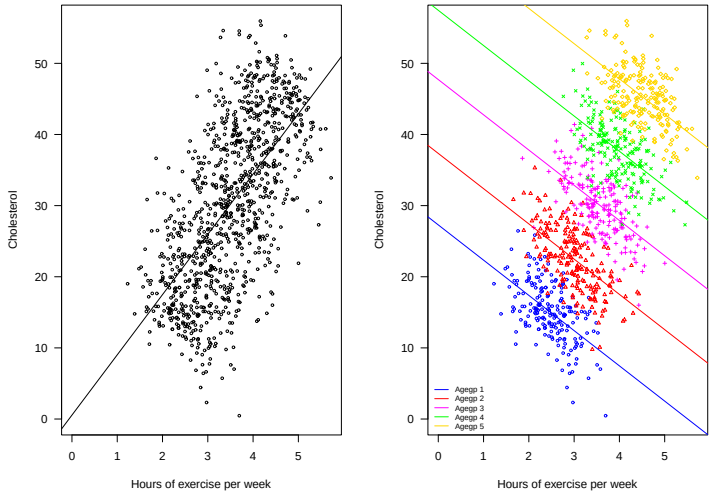


Figure 2.10: Left: Regression line of cholesterol level as function of hours of exercise; right: regression line of cholesterol level as function of hours of exercise stratifying by age group.

However, if now the data are analyzed taking into account the age of the participants in the study, the right side panel of Figure (2.10) shows that for each age group the relationship between exercise and cholesterol level is reversed. Which of the two analyses should we trust?

We have to turn to the story behind the data and try to elucidate the data generating process. If we know that older people, who are more likely to exercise are also more likely to have high cholesterol level regardless of exercise, then the different conclusions from the two analysis can be explained: age is a common cause of both exercise and cholesterol.

The general idea behind causal inference is to try to uncover the mechanism of the system that generated the observed data and estimate the strength of the causal paths from an "exposure" or "treatment" variable towards an outcome of interest. Suppose we observe data for a set of variables Z, an exposure variable X and an outcome Y. Typically, the joint distribution $P(Z, X, Y)$ is unknown, and statistical inference aims to learn some features of $P(Z, X, Y)$, like $\mathbb{E}(Y|X, Y)$ and $\mathrm{var}(Y|X, Y)$. The focus of

causal inference is to know the extent in which the joint distribution changes, say, from $P(Z, X, Y)$ to $P'(Z, X, Y)$ when we intervene on X.

By building a causal model for (Z, X, Y), we have the possibility of answering questions like the following:

- **Observational**. Given that X is observed to a value of x, what is $P(Y|X = x)$?
- **Intervention**. What will be the value of Y if X is forced to a value of x, $P(Y|\,\mathrm{do}(X))$?
- **Counterfactual**. What would have been the value of Y had X been set to x' instead of x?

In order to rigorously address causal inference questions with observational data, the following elements are required:

1. A working definition of "causation".
2. A method to formally articulate causal assumptions.
3. A method to link the structure of the causal model to features of the data.
4. A method to draw conclusions on the link between the causal model and the data.

The next section describes briefly these four requirements in the directed acyclic graphs (DAG) framework (Pearl, 2000), which is the one we mainly we based our analysis on in order to estimate the causal effect of frailty on mortality in the CARE75+ study.

2.5.1 Directed Acyclic Graphs

Directed acyclic graphs (DAGs), based in graph theory, provides a useful mathematical setting to articulate hypothetical causal links among the variables of a dataset. DAGs aid the formulation of (causal) statistical models by identifying the set of variables we should adjust for in order to be able give a valid causal interpretation to the estimates of the model.

A causal graph consists of a set of *vertices* (or *nodes*) and a set of *edges* (or *links*) that connect some pairs of vertices. The vertices in these graphs correspond to variables and the edges denote a certain relationship that holds in pairs of variables, the interpretation of which varies with the application (Pearl, 2000). Two variables connected by an edge are called *adjacent*.

Each edge in a graph can be either directed (marked by a single arrowhead on the edge), or undirected (no arrowheads). A graph with only directed edges and with no cycles is a directed acyclic graph.

$$C \longrightarrow X \longrightarrow Y$$

The node that a directed edge starts from is called the *ancestor* of all the nodes that come after; all the nodes that the edge goes into are the *descendants* (in the path $C \rightarrow X \rightarrow Y$, C and X are the ancestors of Y, and X and Y are the descendants of C). If two nodes are connected by an edge, we have a parent-child relationship (in $C \rightarrow X$, C is parent of X and X is child of C).

2.5.2 Structural Causal Models (SCM)

A *structural causal model* is a way of formally setting down the assumptions about the data generating process; it is a method for describing the relevant features of the system of interest and how they interact with each other. Specifically, a SCM describes how the mechanics of the system naturally assigns values to the variables.

In general, a SCM consists of:

- Two sets of variables: $U = \{U_1, \ldots, U_m\}$ and $V = \{V_1, \ldots, V_p\}$. U are exogenous or external to the system, for which it is decided not to explain how they are caused. U cannot be descendants of any other variables, in particular of any variable in V; they have no ancestors and are represented as root nodes in graphs. V are endogenous and are all variables in the model that are descendants of at least one exogenous variable.
- A set of functions $F = \{f_1, \ldots, f_p\}$ that assigns each variable in V a value based on the values of other variables in the model.

$P(u)$ and F induce a distribution of the observed variables, $P(v)$. Knowing the value of each U_i, then F determines with perfect certainty the value of every V_i.

Example: The salary, Y, that an employer pays to an individual with X years of schooling and Z years in the profession is assigned by the function $f_Y = 2X + 3Z$.

$$U = \{X, Z\}, \quad V = \{Y\}, \quad F = \{f_Y\}, \quad f_Y = 2X + 3Z.$$

$$X \longrightarrow Y \longleftarrow Z$$

Here X and Z are exogenous and Y is endogenous.

Definition 2.5.1 *(cause) A variable X is a direct cause of Y if there is an arrow pointing from X to Y on their DAG. All variables that appear in the function that assigns values to Y are* causes *of Y.*

Definition 2.5.2 *(graphical definition of causation) If, in a graphical model, G, a variable Y is the child of another variable X, then X is a direct cause of Y. If Y is a descendant of X, then X is a potential cause of Y.*

$$X \longrightarrow Y \qquad\qquad X \longrightarrow Z \longrightarrow Y$$

Basic structures in DAGs

The concept of conditional independence can be expressed visually in a DAG, and it captures the probabilistic information in a structural model. The following are the building blocks of DAGs from which the full the structure can be obtained:

$$\begin{aligned} \text{Chain:} \quad & X \longrightarrow C \longrightarrow Y \\ \text{Fork:} \quad & X \longleftarrow C \longrightarrow Y \\ \text{Collider:} \quad & X \longrightarrow C \longleftarrow Y \end{aligned}$$

The conditional probabilistic relationships encoded in these building blocks is summarized in 3 basic rules of DAGs.

Rule 1 (Conditional independence in chains) Two variables, X and Y, are conditionally independent given C, if there is only one unidirectional path between X and Y and C is any set of variables that intercepts that path.

Suppose three variables (X, C, Y) are related to each other according to the following DAG and structural equations:

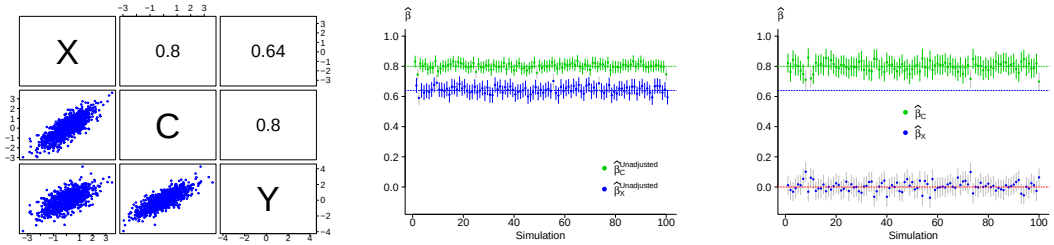
$$X \xrightarrow{0.8} C \xrightarrow{0.8} Y$$

$$\begin{aligned} Y &= 0.8C + \varepsilon_Y \\ C &= 0.8X + \varepsilon_C \\ X &\sim \mathcal{N}(0, 1) \\ \varepsilon_C &\sim \mathcal{N}(0, 0.36) \\ \varepsilon_Y &\sim \mathcal{N}(0, 0.36) \end{aligned}$$

If we want to estimate the effect of X on Y, we should not condition on C if C is in the path from X to Y.

$$Y = \beta_C C + \beta_X X + \varepsilon$$

Fitting the regression model where both X and C are predictors of Y will block the effect $X \rightarrow Y$ giving the impression that X and Y are uncorrelated.



(a) (b) (c)

Figure 2.11: Simulation of a basic chain structure: (a) Scatter plot of one simulated data set and correlations; (b) 95% confidence intervals of $\widehat{\beta}_C$ and $\widehat{\beta}_X$ of the models $Y = \beta_C C + \varepsilon$ and $Y = \beta_X X + \varepsilon$ (c) 95% confidence intervals of $\widehat{\beta}_C$ and $\widehat{\beta}_X$ of the model $Y = \beta_C C + \beta_X X + \varepsilon$, showing that conditioning on C, we cannot rule out $H_0 : \beta_X = 0$.

Figure 2.11a shows that $\text{cor}(X, C) = \text{cor}(Y, C) = 0.8$ and $\text{cor}(X, Y) = 0.64$. Figure 2.11c illustrates via 100 simulated datasets the **Rule in chains**: $Y \perp X \mid C$.

Rule 2 (Conditional independence in forks) If a variable C is a common cause of variables X and Y, and there is only one path between Y and C, then Y and X are independent conditional on C.

Suppose three variables (X, C, Y) are related to each other according to the following DAG and structural equations:

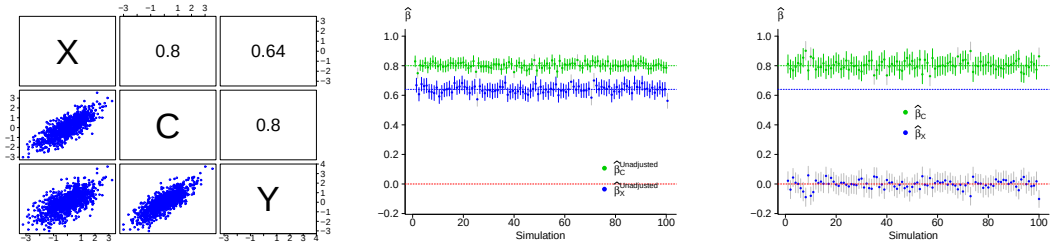
$$X \xleftarrow{0.8} C \xrightarrow{0.8} Y$$

$$\begin{aligned} Y &= 0.8\, C + \varepsilon_Y \\ X &= 0.8\, C + \varepsilon_X \\ C &\sim \mathcal{N}(0, 1) \\ \varepsilon_Y &\sim \mathcal{N}(0, 0.36) \\ \varepsilon_X &\sim \mathcal{N}(0, 0.36) \end{aligned}$$

If we want to estimate the effect of X on Y, we should not condition on C if C is in the path from $X$to Y.

$$Y = \beta_C C + \beta_X X + \varepsilon$$

Fitting the regression model where both X and C are predictors of Y will block the effect $X \rightarrow Y$ giving the impression that X and Y are uncorrelated.



(a) (b) (c)

Figure 2.12: Simulation of a basic fork structure: (a) Scatter plot of one simulated data set and correlations; (b) 95% confidence intervals of $\widehat{\beta}_C$ and $\widehat{\beta}_X$ of the models $Y = \beta_C C + \varepsilon$ and $Y = \beta_X X + \varepsilon$ (c) 95% confidence intervals of $\widehat{\beta}_C$ and $\widehat{\beta}_X$ of the model $Y = \beta_C C + \beta_X X + \varepsilon$, showing that conditioning on C, we cannot rule out $H_0 : \beta_X = 0$.

Figure 2.12a shows that $\text{cor}(X, C) = \text{cor}(Y, C) = 0.8$ and $\text{cor}(X, Y) = 0.64$. Figure 2.12c illustrates via 100 simulated datasets the **Rule in forks**: $Y \perp X \mid C$.

Rule 3 (Conditional dependence in colliders). If C is a common effect of X and Y and there is only one path among them, then C is a collision node. The causes of C, X and Y are (unconditionally) independent, but are conditionally dependent given C and any other descendant of C.

Suppose three variables (X, C, Y) are related to each other according to the following DAG and structural equations:

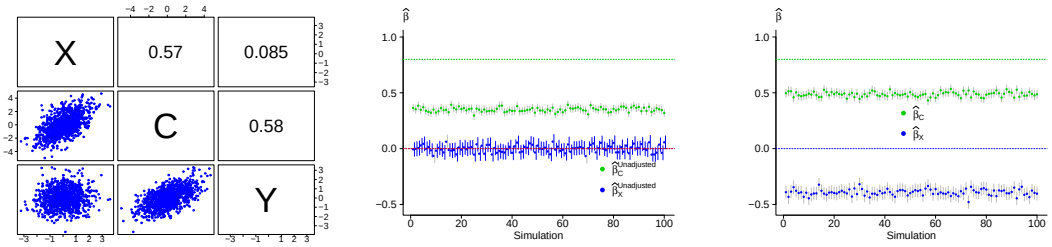
$$X \xrightarrow{0.8} C \xleftarrow{0.8} Y$$

$$\begin{aligned} C &= 0.8X + 0.8Y + \varepsilon_C \\ (X, Y) &\sim \mathcal{N}_2(\mathbf{0}, I_2) \\ C &\sim \mathcal{N}(0, 1) \\ \varepsilon_Y &\sim \mathcal{N}(0, 0.36) \\ \varepsilon_X &\sim \mathcal{N}(0, 0.36) \end{aligned}$$

If we want to estimate the effect of X on Y, we should not condition on C if C is in the path from $X$to Y.

$$Y = \beta_C C + \beta_X X + \varepsilon$$

Fitting the regression model where both X and C are predictors of Y will block the effect $X \rightarrow Y$ giving the impression that X and Y are uncorrelated.



(a) (b) (c)

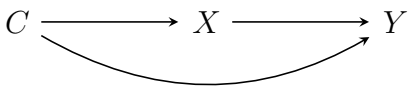
Figure 2.13: Simulation of a basic collider structure: (a) Scatter plot of one simulated data set and correlations; (b) 95% confidence intervals of $\widehat{\beta_C}$ and $\widehat{\beta_X}$ of the models $Y = \beta_C C + \varepsilon$ and $Y = \beta_X X + \varepsilon$ (c) 95% confidence intervals of $\widehat{\beta_C}$ and $\widehat{\beta_X}$ of the model $Y = \beta_C C + \beta_X X + \varepsilon$, showing that conditioning on C, then we rule out $H_0 : \beta_X = 0$, creating a spurious association between X and Y.

Figure 2.13a shows that $\text{cor}(X, C) = 0.57$, $\text{cor}(Y, C) = 0.58$, and $\text{cor}(X, Y) = 0.085$.

Rule in colliders: $Y \not\perp X \mid C$.

Notice that even though X and Y are (unconditionally) independent, conditioning on C produces a spurious association between X and Y. This is called *collider bias*.

Confounder. A variable that apparently changes the relationship between X and Y because it is related to both X and Y. The presence of confounding implies a violation to the assumption of independence between potential outcomes and confounders.



If the relationship between X and Y changes when C causes X and Y, this leads to an observed XY relationship that may be considered causal if C is not included in the analysis.

Case 1

Suppose three variables (X, C, Y) are related to each other according to the following DAG and structural equations:

$$C \xrightarrow{-0.8} X \xrightarrow{0.8} Y$$

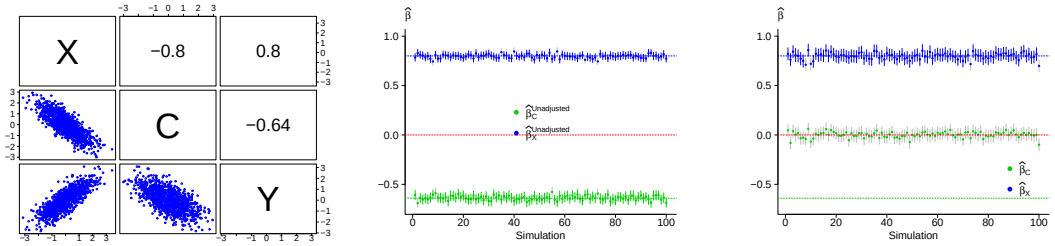
$$\begin{aligned} C &\sim \mathcal{N}(0, 1) \\ X &= -0.8C + \varepsilon_X \\ Y &= 0.8X + \varepsilon_Y \\ \varepsilon_X &\sim \mathcal{N}(0, 0.36) \\ \varepsilon_Y &\sim \mathcal{N}(0, 1) \end{aligned}$$

Figure 2.14a shows the correlation between C, X and Y for a simulated data set.

If we want to estimate the effect of X on Y, we must condition on C if C is a confounder (common cause) of $X$and Y. Ignoring the fact that C is a confounder can bias the estimates of the effect of X on Y. Consider the following two regression equations to estimate the effect of X on Y.

$$\begin{aligned} \text{Unadjusted:} \quad & Y = \beta_X X + \varepsilon \\ & Y = \beta_C C + \varepsilon \\ \text{Adjusted:} \quad & Y = \beta_X X + \beta_C C + \varepsilon \end{aligned}$$

The unadjusted and adjusted estimates of the effect of $X \to Y$ for 100 simulations are shown in Figure 2.14b and 2.14c.

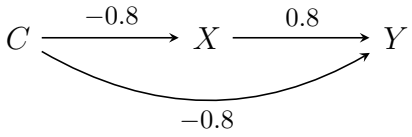


(a) (b) (c)

Figure 2.14: Simulation of a basic confounding structure: (a) Scatter plot of one simulated data set and correlations; (b) 95% confidence intervals of $\widehat{\beta}_X$ and $\widehat{\beta}_C$ of the univariate models $Y = \beta_X X + \varepsilon$ and $Y = \beta_C C + \varepsilon$; (c) 95% confidence intervals of $\widehat{\beta}_X$ and $\widehat{\beta}_C$ of the model $Y = \beta_C C + \beta_X X + \varepsilon$, showing that conditioning on C we cannot rule out $H_0 : \beta_X = 0$.

Case 2

Suppose three variables (X, C, Y) are related to each other according to the following DAG and structural equations:



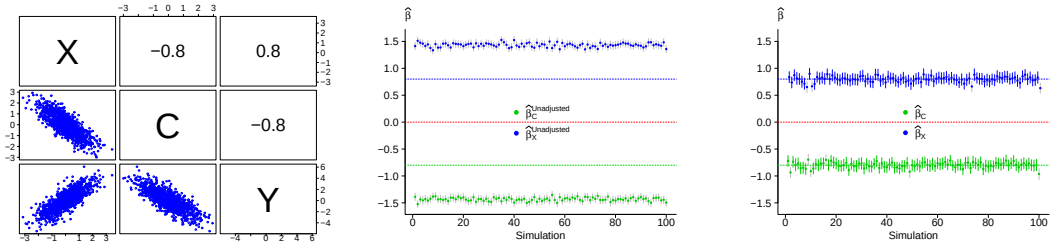
$$
\begin{aligned}
C &\sim \mathcal{N}(0, 1) \\
X &= -0.8C + \varepsilon_X \\
Y &= 0.8X - 0.8\,C + \varepsilon_Y \\
\varepsilon_X &\sim \mathcal{N}(0, 0.36) \\
\varepsilon_Y &\sim \mathcal{N}(0, 1)
\end{aligned}
$$

Figure 2.15a shows the correlation between C, X and Y for a simulated data set.

If we want to estimate the effect of X on Y, we must condition on C if C is a confounder (common cause) of $X$and Y. Ignoring the fact that C is a confounder can severely bias the estimates of the effect of X on Y. Consider the following two regression equations to estimate the effect of X on Y.

$$
\begin{aligned}
\text{Unadjusted:} \quad & Y = \beta_X X + \varepsilon \\
\text{Adjusted:} \quad & Y = \beta_X X + \beta_C C + \varepsilon
\end{aligned}
$$

The unadjusted and adjusted estimates of the effect of $X \rightarrow Y$ for 100 simulations are shown in Figure 2.15b and 2.15c.

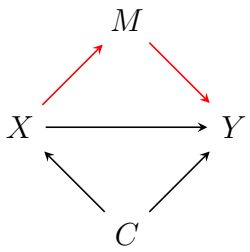


(a) (b) (c)

Figure 2.15: Simulation of a basic confounding structure: (a) Scatter plot of one simulated data set and correlations; (b) 95% confidence intervals of $\widehat{\beta}_X$ and $\widehat{\beta}_C$ of the univariate models $Y = \beta_X X + \varepsilon$ and $Y = \beta_C C + \varepsilon$, showing that failing to condition on the common cause C, results in wrong estimates of $\widehat{\beta}_X$; (c) 95% confidence intervals of $\widehat{\beta}_X$ and $\widehat{\beta}_C$ of the model $Y = \beta_C C + \beta_X X + \varepsilon$.

Mediation. Refers to the situation in which X causes M, which causes Y, also called a *mediating variable*. Here the third variable, M, is intermediate in the causal chain relating X and Y such that X causes M and M causes Y.

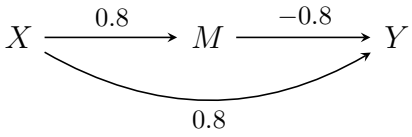
Both confounders and mediator account for the relationship XY. The difference is that a confounder explains the relation because it is related to both X and Y as a common cause $(X \leftarrow C \rightarrow Y)$, rather than as part of a causal mediation process. The mediator explains the XY relation because it transmits the effect of X on Y.



$X =$ Exposure
$M =$ Mediator
$C =$ Counfounder
$Y =$ Outcome

Mediators are also called intervening or intermediate variables to indicate their role as coming between X and Y. Other names include *process variable*, because M describes the process by which X affects Y; *surrogate* or *intermediate endpoints*, because they represent proximal measures of a distal outcome.

Suppose three variables (X, M, Y) are related to each other according to the following DAG and structural equations:



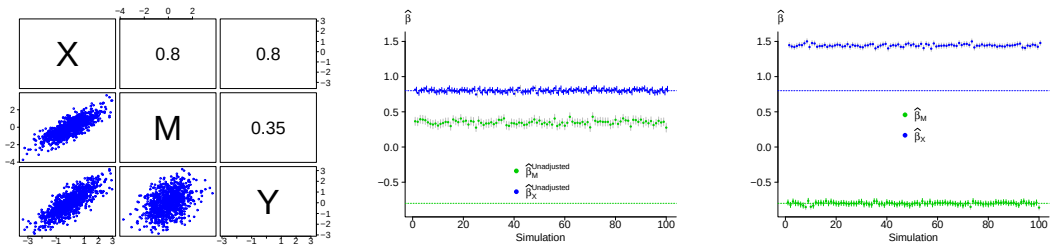
$$
\begin{aligned}
X &\sim \mathcal{N}(0, 1) \\
M &= 0.8\ X + \varepsilon_M \\
Y &= -0.8\ M + 0.8\ X + \varepsilon_Y \\
\varepsilon_M &\sim \mathcal{N}(0, 0.36) \\
\varepsilon_Y &\sim \mathcal{N}(0, 0.13)
\end{aligned}
$$

Figure 2.16a shows the correlation between M, X and Y for a simulated data set.

If we want to estimate the direct effect of X on Y, we must not condition on M if M is in the path between $X$and Y. Conditioning on M will block the effect $X \to Y$.
Consider the following regression equations.

$$
\begin{aligned}
\text{Unadjusted:} \quad & Y = \beta_X X + \varepsilon \\
& Y = \beta_M M + \varepsilon \\
\text{Adjusted:} \quad & Y = \beta_X X + \beta_M M + \varepsilon
\end{aligned}
$$

The unadjusted and adjusted estimates of the effect of $X \to Y$ and their 95% confidence intervals for 100 simulations are shown in Figure 2.16b and 2.16c. The consequence of conditioning on M is that the estimates of β_X are largely biased (Figure 2.16c). Ignoring M will produce unbiased estimates of β_X, (Figure 2.16b).



(a) (b) (c)

Figure 2.16: Simulation of a basic mediation structure: (a) Scatter plot of one simulated data set and correlations; (b) 95% confidence intervals of $\widehat{\beta}_X$ and $\widehat{\beta}_M$ of the univariate models $Y = \beta_X X + \varepsilon_Y$ and $Y = \beta_M M + \varepsilon_Y$; (c) 95% confidence intervals of $\widehat{\beta}_X$ and $\widehat{\beta}_M$ of the model $Y = \beta_M M + \beta_X X + \varepsilon$, showing that conditioning on a mediating variable M results in a wrong estimate $\widehat{\beta}_X$.

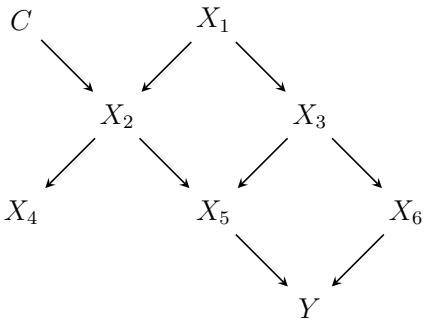
2.5.3 Rule of product decomposition

For any model whose graph is acyclic, the joint distribution of the variables in the model is given by the product of the conditional distributions $P(child|parents)$ over all "families" in the graph. We write this rule as

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P\left(x_i \mid \mathrm{pa}(x_i)\right),$$

where $\mathrm{pa}(x_i)$ stands for the parents of X_i.

Figure (2.17) is a DAG showing the relationships between variables $Y, C, X_1, \ldots, X_6$ and the factorization of their joint distribution according to the conditional independencies depicted in the DAG.



$$\begin{aligned} &P(Y, C, X_1, \ldots, X_6) \\ &\quad = P(C)P(X_1)P(X_2 \mid C, X_1)P(X_3 \mid X_1) \\ &\qquad P(X_4 \mid X_2)P(X_5 \mid X_2, X_3)P(X_6 \mid X_3) \\ &\qquad P(Y \mid X_5, X_6). \end{aligned}$$

Figure 2.17: X_5 and X_6 are direct causes of Y, and the rest of the variables of the DAG are potential causes of Y.

Definition 2.5.3 *(d-separation) A path, p, is said to be d-separated (or blocked) by a set of nodes S, if and only if*

1. *p contains a chain of nodes $(A \to B \to C)$ or a fork $(A \leftarrow B \to C)$, such that the middle node $B \in S$ (i.e. B is conditioned on), or*
2. *p contains a collider $(A \to B \leftarrow C)$ such that the collision node $B \notin S$, and such that no descendant of B is in S.*

A set S is said to d-separate X from Y if S blocks every path from a node in X to a node in Y.

2.5.4 Model testing for causal search

If we have a DAG, G, that we believe might have generated a data set D, d-separation will tell us which variables in G must be independent conditional on which other variables. We can test for conditional independence from X to Y using data.

1. List all the paths between X and Y in G.
2. Identify the set S with all possible variables that conditioned on make X and Y, conditionally independent.
3. Estimate probabilities based on the data, D.
4. Investigate with estimated probabilities if the data supports the hypothesis that X and Y are independent conditioned on S. Otherwise, reject G as a candidate causal model for D.

We say that two variables, X and Y are causally related if a change in X has the potential to change Y. A distinctive aspect of causal analysis is separation of the exposure, X, from the confounders, Z. All the attention is focused on trying to learn about X. In the causal model we propose for the XY relationship we want to make sure to identify all the confounders for the joint distribution $P(X, Y)$. Even though the Z variables may have confounders with their relationship with Y, we treat them as nuisance parameters. The importance of Z limits to being controlled for appropriately according to their position in the data generating process so that we can learn about the XY relationship.

2.5.5 The effects of intervention

There is difference between intervening on a variable and conditioning on that variable. When intervening in a variable in a model, the value is fixed; we change the system, and the values of other variables often change as a result. When conditioning on a variable, we change nothing; we merely narrow our focus to the subset of cases in which the variable takes the value we are interested in. What changes, when conditioning on a variable, is our perception of the world, not the world itself (Pearl *et al.*, 2016).

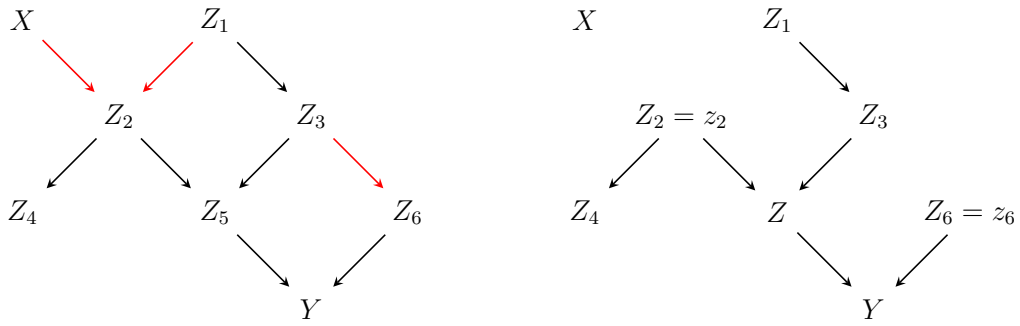


Figure 2.18: Left: red arrows in the DAG indicate that we can manipulate Z_2 and Z_6. Right: resulting DAG after manipulation, forcing Z_2 and Z_6 to take the values z_2 and z_6, respectively. By doing so, the arrows pointing to Z_2 and Z_6 are removed.

In an intervention, we force a variable to take a specific value. We denote the intervention on variable X to fix its value at $X = x$ as $\text{do}(X = x)$. So $P(Y = y|X = x)$ is the probability distribution of $Y = y$ among the subset of the population whose X value is x, while $P\{Y = y|\text{do}(X = x)\}$ is the probability that $Y = y$ when we intervene to make $X = x$, i.e. the probability distribution of Y if everyone in the population had their value X fixed at x.

2.5.6 The adjustment formula

Suppose we want to find out how effective a drug is in the population. We imagine a hypothetical intervention by which we administer the drug uniformly to the entire population and compare the recovery rate to what we would obtain under the complementary intervention, where we prevent everyone from using the drug. Denoting the first situation with $\text{do}(X = 1)$, the second with $\text{do}(X = 0)$ and the situation where a person recovers by $Y = 1$ we compute the *average causal effect* or *causal effect difference* with

$$\text{ACE} = P\{Y = 1|\text{do}(X = 1)\} - P\{Y = 1|\text{do}(X = 0)\}$$

If X and Y take more than one value, then use $P\{Y = y|\text{do}(X = x)\}$ for two arbitrary values x, y.

Formula for the causal effect in terms of preintervention probabilities (*adjustment formula)*

$$P\{Y = y|\text{do}(X = x)\} = \sum_z P(Y = y|X = x, Z = z)P(Z = z). \tag{2.55}$$

This formula computes the association between X and Y for each value of Z, then averages over those values.

To adjust or not to adjust?

What set of variables Z can legitimately be included in the adjustment formula? The intervention procedure, which led to the adjustment formula, dictates that Z should coincide with the parents of X, because it is the influence of these parents that we neutralize when we fix X by external manipulation.

Denoting the parents of X by $\text{pa}(X)$ we can therefore write a general adjustment formula and summarize it in a rule.

Rule 4 (The causal effect rule). Given a graph G in which a set of variables $\text{pa}(X)$ are designated as the parents of X, the causal effect of X is given by

$$P\{Y = y|\text{do}(X = x)\} = \sum_z P\{Y = y|X = x, \text{pa}(X) = z\}P\{\text{pa}(X) = z\}. \tag{2.56}$$

Multiply and divide by $P\{X = x|\,\text{pa}(X) = z\}$

$$P\{Y = y|\text{do}(X = x)\} = \sum_z \frac{P\{Y = y, X = x, \text{pa}(X) = z\}}{P\{X = x|\,\text{pa}(X) = z\}},$$

which displays the role of the parents of X in predicting the results of interventions. The factor $P(X = x \mid \text{pa}(X) = z)$ is the *propensity score* (also called balancing score). Note that

- Preintervention: $P(x, y, z) = P(z)P(x|z)P(y|x, z)$
- Postintervention: $P\{y, z|\text{do}(x)\} = P(z)P(y|x, z)$.

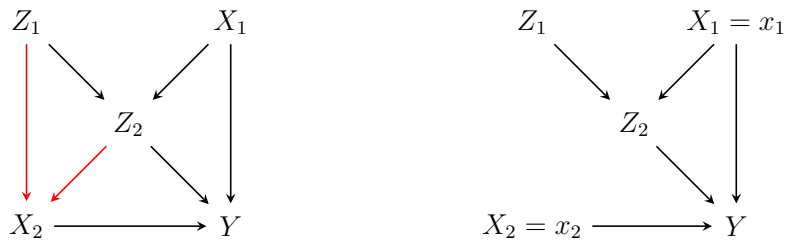


Figure 2.19: Left: red arrows in the DAG indicate that we can manipulate X_1 and X_2. Right: resulting DAG after manipulation, forcing X_1 and X_2 to take the values x_1 and x_2, respectively. By doing so, the arrows pointing to X_2 are removed.

Combining preintervention and postintervention we obtain

$$P\{z, y|\text{do}(x)\} = P(z)P(y|x, z) = P(z)\frac{P(x, y, z)}{P(z)P(x|z)} = \frac{P(x, y, z)}{P(x|z)}.$$

This means that $P(x|z)$ is all we need to know in order to predict the effect of an intervention $\text{do}(x)$ from nonexperimental data governed by the distribution $P(x, y, z)$.

Finally, the postintervention distribution can be generalized to multiple interventions. As an example, consider the DAG of on the left side of Figure 2.19 for the relationships in $V = \{Y, Z_1, Z_2, X_1, X_2\}$. Suppose we intervene to fix X_1 and X_2 to x_1 and x_2, as shown on the right side of Figure 2.19. The postintervention distribution of V is given by

$$P\{z_1, z_2, y \mid \text{do}(x_1, x_2)\} = P(z_1)P(z_2|x_1, z_1)P(y|x_1, x_2, z_2),$$

which allows us to extend the adjustment formula in Equation (2.55) to multiple interventions.

$$P\{y \mid \text{do}(x_1, x_2)\} = \sum_{z_2}\sum_{z_1} P(z_1)P(z_2|x_1, z_1)P(y|x_1, x_2, z_2) \tag{2.57}$$

The expression in Equation (2.57) is in agreement with G-computation formula proposed by Robins (1986), which is a generalization derived from a more complicated set of assumptions on counterfactuals. This formula dictates an adjustment for Z_2 that might be affected by a previous exposure variable, say X_1 (Pearl, 2010).

How do we choose among different competing causal models? The answer to this

question can be given in two steps:

1. The first part is to make a statement about the plausibility of the effect of X on Y. When we have several competing causal models we should organize them in a compact representation and see whether they agree or not with a certain causal query, e.g. there is an effect of X on Y. If they do agree, then we can make a strong causal statement of the exposure-effect relationship we are trying to address.

2. Once having made a statement about X having an effect on Y, we are interested in an estimate of the size of such effect. We can make a subset of the competing models that agree on such query and compare their performance on the data. Penny *et al.* (2004) describe the use of Bayes factor (Kass & Raftery, 1995) for comparing dynamic causal models. Penny (2012) provides some evidence from simulation studies of the performance of AIC, BIC and variational Free Energy (based on the Kulback-Leibler divergence) for selecting competing causal models. The estimated magnitude of the effect of X on Y will be given by the estimate of the selected model.

Chapter 3

Joint models of longitudinal and time-to-event data

In follow-up studies usually different types of outcomes are collected for each sample unit, which may include multiple longitudinal repeated measures and the time until an event of particular interest occurs. The research questions of interest are typically formulated for separate analyses of the recorded outcomes, nonetheless sometimes due to the characteristics of the data generating process of the phenomena under study, joint modelling of the different outcomes is more convenient. In this thesis the focus is on joint modelling longitudinal measures of a quantitative outcome, recurrent events and a terminal event.

In Sections 2.2.1 and 2.1.3 we introduced the challenges in survival analysis models with time-varying covariates and missing data in longitudinal studies that motivate joint modelling longitudinal and time-to-event data. In Sections 3.1 and 3.1.4 we explaine how by joint modelling the longitudinal and time-to-event outcomes it is possible to accommodate endogenous time-varying covariates in survival analysis models and to account for the bias induced by data missing not at random in longitudinal studies. However, there are other situations in which joint outcome modelling is important.

In some cases the time-to-event outcome does not refer to death, but to an event that causes a major change in the dynamics of the quantitative outcome and this gives rise

to research questions and features of the data for which joint modelling is necessary. For instance:

- Prostate specific antigens (PSA) are used for monitoring patients treated for prostate cancer. If cancer relapse occurs, the dynamics of PSA changes because of the set-up of new treatments. In this case the relevant event is relapse.
- CD4 and HIV viral load are used for monitoring HIV patients. These two markers change dramatically after switching to the AIDS stage or after the initiation of antiretroviral treatment. In this context, the relevant event is the change of treatment.

In these two examples, as the event status (relapse or change of treatment) changes the dynamics of the quantitative response changes as well, causing its distribution after the event to be different from its distribution before the event. Joint modelling allows to estimate the quantitative response change over time conditionally on the time of the event.

Often interest lies in exploring the joint distribution of a longitudinal and time-to-event outcome to understand how they relate to each other. For instance, Ibrahim *et al.* (2010) discuss the importance of joint modelling quality of life (QOL) and mortality in cancer patients since one might argue that for a patient, improvement in QOL is often more important than any modest survival benefit in treatment decisions. Therefore, it is of great interest in cancer clinical trials to characterize the association between time-to-event and QOL through joint modelling and to understand the tradeoffs between QOL and survival. A specific treatment protocol with chemotherapy/radiotherapy may prolong survival or relapse, but the QOL in that prolonged period may be poor, and thus the clinician must decide whether such a benefit is worth it for the patient.

Several approaches have been proposed for the statistical analysis of joint models for longitudinal and time-to-event data, which can be grouped in likelihood maximization and Bayesian methods. Two types of joint models have been proposed based on likelihood maximization: Shared Random Effects Joint Models (SREJM) and Latent Class Joint Models (LCJM), combine a linear mixed model for the evolution for the longitudinal outcome and a survival model for the time-to-event outcome. The difference between these two alternative models is the latent structure that defines the association

between the two outcomes. In a SREJM a function of the random effects is introduced as an explanatory variable in the survival submodel. In the LCJM the population is assumed to be heterogeneous but comprised of subpopulations with different patterns of change for the longitudinal outcome and different risk profiles for the time-to-event outcome. Commenges & Jacqmin-Gadda (2015). Hickey *et al.* (2016) and Hickey *et al.* (2018) provide recent review for the implementation of joint models for longitudinal and time-to-event outcomes, including the estimation approaches utilized. This thesis uses the Shared Random Effects Joint Model since it is the approach most widely used.

3.1 Shared Random Effects Joint Model for Longitudinal and Time-to-Event Data

Let T_i denote the recorded failure time for the i^{th} subject $(i = 1, \ldots, n)$, which is taken to be the minimum of the "true" event time, T_i^*, and the censoring time, C_i, that is, $T_i = \min(T_i^*, C_i)$. Denote the event indicator by $\delta_i = \mathbb{1}(T_i^* \leq C_i)$, where $\mathbb{1}(z)$ is the indicator function taking the value of 1 if z is true, and 0 if z is false. Let $y_i(t)$ denote the observed value of the longitudinal outcome for subject i at time t, which is an error prone version of the true and unobservable $m_i(t)$ (See Chapter 1 for the relationship between hypothetical constructs, measurements and measurement error). This is,

$$y_i(t) = m_i(t) + \varepsilon_i(t),$$

where $\varepsilon_i(t)$ denotes the measurement error.

In studies with longitudinal data, $y_i(t)$ is not measured continuously at every time t, but only at specific follow-up time points $t_{ij},\ j = 1, \ldots, n_i,\ i = 1, \ldots, n$, which might be different for each subject. We will denote by y_{ij} the value of the j^{th} repeated measure of the longitudinal outcome for subject i taken at time point t_{ij}. Hence the vector of observed repeated measures of the longitudinal outcome for subject i consists of $\mathbf{y}_i = \{y_{ij}; j = 1, \ldots, n_i\}$. Figure 3.1 illustrates what these data would look like (•) for two hypothetical subjects: of the longitudinal outcome and died shortly after the

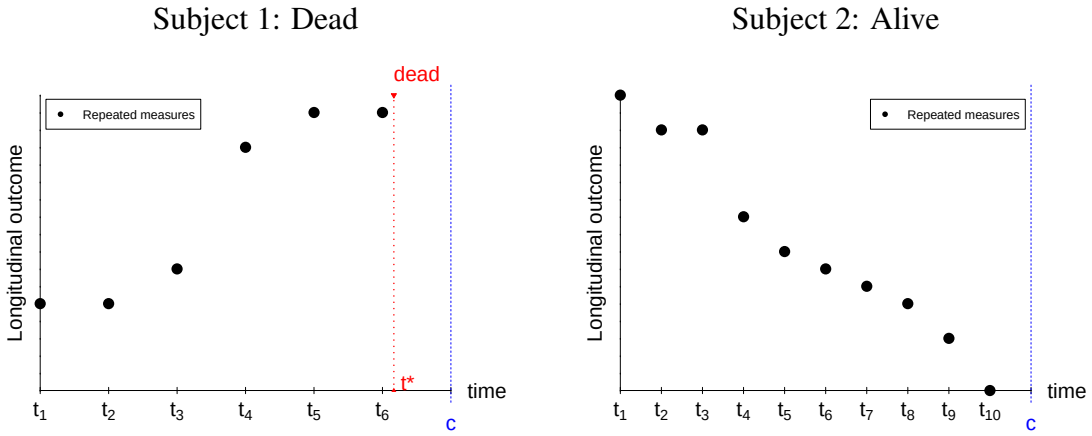


Figure 3.1: Illustration of longitudinal and time-to-event data of two fictitious subjects. Subject 1 (left) has 6 repeated measures of the longitudinal outcome (•) and died at time t^*; Subject 2 (right) has 10 repeated measures (•) and is alive by the censoring time C.

sixth measurement at time t^* (left), and Subject 2 provided 10 repeated measures and was still alive by the administrative censoring time C.

In the joint modelling framework it is conjectured that the longitudinal process, $m_i(t)$ is directly associated with the hazard rate $h_i(t)$ of a terminal event (or time-to-event outcome). The intuitive idea behind joint models for longitudinal and time-to-event data is depicted in Figure 3.2 (Rizopoulos, 2012).

The left panel of Figure 3.2 shows the true and unobservable trajectory of the longitudinal outcome $m_i(t)$ (- -) and the repeated measures of its noisy version, y_{ij} (•), taken at time-points t_{ij}. In the extended Cox model with $y_i(t)$ being a time-varying covariate, estimates of $h_i(t)$ (–•–) are based on the crude repeated measurements $(y_{ij}$ under the assumption that $y_i(t)$, remains constant between consecutive measurements. The effect of measurement error on the hazard rate can be seen by comparing the estimates of $h_i(t_{ij})$ based on y_{ij} against the true unobserved hazard rate (—) on the right panel. Moreover, when y_{ij} is an endogenous time-varying covariate in a survival analysis model and there is interest in predicting the hazard all along the follow-up period $(0 \leq t \leq t^*)$, it is important to know its value at all times, t, and not only at the data collection time points, t_{ij}. Thus we need to be able to reconstruct the complete longi-

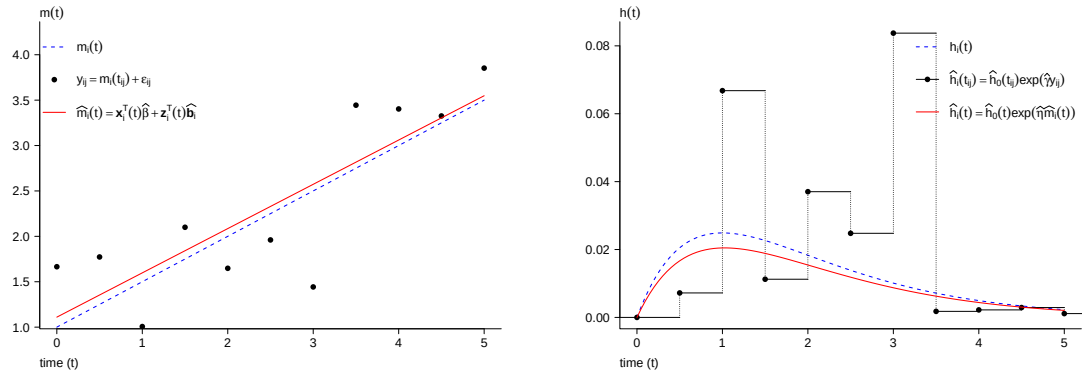


Figure 3.2: Longitudinal outcome process (- - true & unobserved, • observed and — estimated) and hazard function (–•– Cox model, - - evaluated at the true parameters and — estimated with joint model).

tudinal history for each subject, $\mathcal{M}_i(t)$. The joint modelling framework postulates the simultaneous estimation of the parameters of both hazard rate and longitudinal outcome process. The longitudinal outcome trajectory, $\mathcal{M}_i(t)$, is estimated by a linear mixed model (— left panel) to account for measurement error. By doing so it is possible to relax the assumption of piecewise constant time-varying covariates and it is possible to estimate both longitudinal outcome and hazard rate at all t and correct for the effect of measurement error on the hazard (— in right panel).

In addition to data for the longitudinal and time-to-event outcomes, data about baseline and possibly time-varying covariates are often collected. Let $\mathbf{x}_i(t)$ be the p-vector of baseline covariates (possibly time-varying) of subject i at time t associated to the fixed effects longitudinal outcome, and $\mathbf{z}_i(t)$ the covariates associated with a q-vector of random effects. The same as for $y_i(t)$, covariate data are typically known for specific time-points, $t_{ij}, j = 1, \ldots, n_i$. We will denote by $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ the values of the covariate vectors at the specific time-point t_{ij}. Define X_i as the $n_i \times p$ fixed effects design matrix of subject i conformed by stacking the n_i vectors $\mathbf{x}_{ij}^\top$, and similarly the $n_i \times q$ random effects design matrix Z_i conformed by the n_i vectors $\mathbf{z}_{ij}^\top$. Let $\mathbf{w}_i$ be a vector of baseline covariates associated to the time-to-event outcome. For simplicity, in this thesis we will consider $\mathbf{w}_i$ to be time-independent, although this is not strictly necessary and $\mathbf{w}_i$ may contain exogenous time-varying covariates. The vectors $\mathbf{x}_i(t)$ and $\mathbf{w}_i$ might have

covariates in common.

The joint modelling framework for longitudinal and time-to-event data is based on the fundamental assumption that the two outcomes are conditionally independent given the random effects. It requires the specification of a regression submodel for each outcome, a covariance structure for random effects which are assumed to act in both submodels, and *link* functions connecting both submodels. Equations (3.1a)–(3.1b) describe the standard joint model for a longitudinal and a time-to-event outcome.

$$\begin{cases} \underset{\text{(Longitudinal)}}{y_i(t \mid \mathbf{b}_i)} = m_i(t) + \varepsilon_i(t) = \mathbf{x}_i^\top(t)\boldsymbol{\beta} + \mathbf{z}_i^\top(t)\mathbf{b}_i + \varepsilon_i(t) & \text{(3.1a)} \\ \underset{\text{(Terminal)}}{h_i(t \mid \mathbf{b}_i)} = h_0(t)\exp\left\{\mathbf{w}_i^\top\boldsymbol{\gamma} + \left[g\left(\mathbf{b}_i, t\right)\right]^\top\boldsymbol{\eta}\right\}. & \text{(3.1b)} \end{cases}$$

where $\varepsilon_i(t) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ for all t and all i is the within-subject measurement error assumed independent and normally distributed with constant variance σ_ε^2. The vectors $\mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}, B)$ are the subject-specific random effects assumed to follow a multivariate normal distribution centered at zero with covariance matrix B. We assume that measurement error and random effects are independent, $\varepsilon_i(t) \perp \mathbf{b}_i$, and that the repeated measures and the time-to-event outcomes are conditionally independent given the random effects, $\mathbf{y}_i \perp \{T_i, \delta_i\} \mid \mathbf{b}_i$. The baseline hazard, $h_0(t)$ is a positive real-valued function.

Equation 3.1a describes the linear mixed model formulation of the longitudinal outcome of subject i, where $m_i(t) = \mathbf{x}_i^\top(t)\boldsymbol{\beta} + \mathbf{z}_i^\top(t)\mathbf{b}_i$ represents the true value of the longitudinal outcome at time t. Here, $y_i(t), m_i(t)$ and $\varepsilon_i(t)$ are scalars, and $\mathbf{x}_i(t)$ and $\mathbf{z}_i(t)$ are column vectors. The vectors $\mathbf{x}_i^\top(t)$ and $\mathbf{z}_i(t)$ have associated regression coefficients $\boldsymbol{\beta}$ and $\mathbf{b}_i$, respectively.

Equation 3.1b describes the regression submodel for the hazard rate of of a time-to-event outcome, where $\mathbf{w}_i$ is the vector of baseline covariates for subject i with associated vector of regression coefficients $\boldsymbol{\gamma}$, and $h_0(t)$ denoting the baseline hazard function. The additional term $g(\mathbf{b}_i, t)$ represents the form of the association between the two outcomes, where the function g is the *link function*, possibly vector valued. Any function of the random effects, $\mathbf{b}_i$ can be considered and it is chosen depending on the context and purpose of the study (Commenges & Jacqmin-Gadda, 2015).

The strength of such association is quantified by the regression coefficient's vector $\boldsymbol{\eta}$. There are no restrictions for g and in practice the identity function is often considered. The most important part of the link function are the random effects because of the conditional independence assumption of joint modelling. In Section 3.1.1 we discuss common choices for the link function.

3.1.1 Link Function

The link function that connects the marker with the survival outcomes could be in principle any function of the random effects, $\mathbf{b}_i$, and the fixed effects. Some examples are the following:

- $g(\mathbf{b}_i, t) = \mathbf{b}_i$. The risk of event depends only on the individual random effects. This is the most common link function.
- $g(\mathbf{b}_i, t) = m_i(t) = \mathbf{x}_i(t)\boldsymbol{\beta} + \mathbf{z}_i(t)\mathbf{b}_i$. This function is also a commonly used. It assumes that the instantaneous risk of event at t depends on the longitudinal outcome at t free of measurement error.
- $g(\mathbf{b}_i, t)^\top = \left(m_i(t), \frac{d}{dt}m_i(t)\right)$. It assumes dependence of the terminal event on the current value and the trend. For example, the risk of prostate cancer relapse depends on the level of a biomarker and its most recent change.
- $g(\mathbf{b}_i, t) = \mathbf{z}_i(t)\mathbf{b}_i$. Assumes that the event risk at t is a function of the individual deviation of the longitudinal outcome at t from the population average.

3.1.2 Baseline Hazard

In the PH Cox model, it is not required to specify the baseline hazard and in practice it is customary left completely unspecified. In the joint modelling framework, it is important to estimate the baseline hazard to avoid underestimation of the standard errors of the parameter estimates (Hsieh *et al.*, 2006; Rizopoulos, 2012). One option is to assume a baseline hazard governed by a known parametric distribution, for instance Weibull or Gamma. Another alternative is a flexible or nonparametric

specification of the baseline hazard, being the most common piece-wise constant and spline-approximated hazards.

3.1.3 Estimation

The unknown quantities of the joint model formulation that need to be estimated are

$$\boldsymbol{\theta} = \left(h_0(t), \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma_\varepsilon^2, \texttt{vech}(B)\right).$$

Suppose that data, $\mathcal{D} = \{t_i, \delta_i, \mathbf{y}_i;\ i = 1, \ldots, n\}$, of both longitudinal and time-to-event outcomes are collected on subjects $i = 1, \ldots, n$. The estimation of $\boldsymbol{\theta}$ is based on maximum likelihood principles by maximizing the log-likelihood function of the joint distribution of the longitudinal and the time-to-event outcomes, $\{\mathbf{y}_i, t_i, \delta_i\}$:

$$\ell(\boldsymbol{\theta}|\mathcal{D}) = \sum_{i=1}^{n} \log\left(\int_{\mathbb{R}^q} f(\mathbf{y}_i \mid \mathbf{b}_i) f(t_i, \delta_i \mid \mathbf{b}_i) f(\mathbf{b}_i) \mathrm{d}\mathbf{b}_i\right), \text{ where} \tag{3.2}$$

$$f(\mathbf{y}_i \mid \mathbf{b}_i) = (2\pi\sigma_\varepsilon^2)^{-n_i/2} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2} \left\|\mathbf{y}_i - X_i\boldsymbol{\beta} - Z_i\mathbf{b}_i\right\|^2\right\}, \tag{3.2a}$$

$$f(t_i, \delta_i \mid \mathbf{b}_i) = \left[h_i\left(t_i \mid \mathcal{M}_i(t_i)\right)\right]^{\delta_i} S_i\left(t_i \mid \mathcal{M}_i(t_i)\right), \tag{3.2b}$$

$$f(\mathbf{b}_i) = (2\pi)^{-q/2}|B|^{-1/2} \exp\left\{-\frac{1}{2}\|B^{-1/2}\mathbf{b}_i\|^2\right\}, \tag{3.2c}$$

with

$$h_i\left(t_i \mid \mathcal{M}_i(t_i)\right) = h_0(t_i) \exp\left\{\mathbf{w}_i^\top \boldsymbol{\gamma} + \left[g\left(\mathbf{x}_i^\top(t_i)\boldsymbol{\beta} + \mathbf{z}_i^\top(t_i)\mathbf{b}_i\right)\right]^\top \boldsymbol{\eta}\right\},$$

$$S_i\left(t_i \mid \mathcal{M}_i(t_i)\right) = \int_0^{t_i} h_i\left(t \mid \mathcal{M}_i(t_i)\right) \mathrm{d}t.$$

Note that in the log-likelihood of Equation (3.2), the hazard rate $h_i(t_i|\mathcal{M}_i(t_i))$ depends on the current value of $m_i(t)$ through $g(m_i(t))$. However, the survival function $S_i(t_i|\mathcal{M}_i(t_i))$ depends on knowing the whole trajectory of the longitudinal outcome up to time t_i, i.e. $\mathcal{M}_i(t_i)$. This shows the need for recovering the full path of the longitudinal outcome which is approximated by the linear mixed model (Equation (3.1a)) in the joint modelling framework.

Several estimation approaches have been proposed. For instance, Wulfsohn & Tsiatis (1997) used the Expectation-Maximization (EM) algorithm Dempster *et al.* (1977). (Rizopoulos, 2012) proposed a hybrid optimization procedure of the log-likelihood function of Equation (3.2) starting with the EM algorithm (Dempster *et al.*, 1977) for a fixed number of iterations and switching to a quasi-Newton algorithm until convergence. This procedure is implemented in function `jointModel()` of the R library `JM`. Rondeau *et al.* (2003, 2007) followed a penalized maximum likelihood approach using the Marquardt algorithm (Marquardt, 1963) to optimize the likelihood function, which combines the Newton–Raphson and steepest descent algorithms, and it is implemented in the `frailtyPenal()` function of the R package `frailtypack` (Rondeau *et al.*, 2012).

Markov chain Monte Carlo methods (Gibbs sampling and Metropolis–Hastings algorithms) have been employed for Bayesian estimation of joint models. See for instance, Fawcett & Thomas (1996), R. Brown & G. Ibrahim (2003), Ibrahim *et al.* (2004), Das *et al.* (2012) and Rizopoulos (2014).

As discussed in Section 2.2.1 with endongenous covariates the partial likelihood function of the Cox model is no longer valid and the regression coefficient estimates are biased. To illustrate this point and the need of joint modelling to obtain unbiased regression coefficient estimates, consider as an example Figure 3.3 produced with 100 simulated data sets from the joint model described by Equations (3.3a)–(3.3b), where the link is $g(\mathbf{b}_i, t) = m_i(t)$. Appendix B.1 describes the full simulation scheme and contains plots of other scenarios.

$$M = \begin{cases} \underset{\text{(Longitudinal)}}{y_i(t \mid \mathbf{b}_i)} = \underbrace{(\beta_0 + b_{i0}) + (\beta_t + b_{i1})t + \mathbf{w}_i^\top \boldsymbol{\beta}}_{m_i(t)} + \varepsilon_i(t) & \text{(3.3a)} \\ \underset{\text{(Terminal)}}{h_i(t \mid \mathbf{b}_i)} = h_0(t) \exp\{\mathbf{w}_i^\top \boldsymbol{\gamma} + \eta m_i(t)\} & \text{(3.3b)} \end{cases}$$

We estimated η, the regression coefficient of the time-varying covariate $m_i(t)$ by fitting (a) the extended Cox model with only the time-varying covariate, (b) the extended Cox model adjusting for $\mathbf{w}_i$, and (c) the joint model. The boxplots (top left) are the 100 estimates $\widehat{\eta}$ of all the simulations (univariate Cox model, Cox model adjusting for $\mathbf{w}_i$ and joint model) with a horizontal line indicating the true value of the regression

coefficient η. The vertical lines of the other three plots represent the range of the 95% confidence intervals of $\widehat{\eta}$ estimated with the three fitted models. Notice on the boxplots and the 95% interval plots of the Cox model's estimates (top right and bottom left) that the estimates $\widehat{\eta}$ are biased (true value represented by - -), whether or not we adjust for $\mathbf{w}_i$ and their interval estimates have coverage probability of 0. On the other hand, the estimates $\widehat{\eta}$ of the fitted joint model are unbiased and their 95% confidence intervals have coverage probability of 0.97. Confidence intervals are calculated based on the asymptotic properties of the estimators $\widehat{\eta} \pm 1.96\,\mathrm{se}(\widehat{\eta})$.

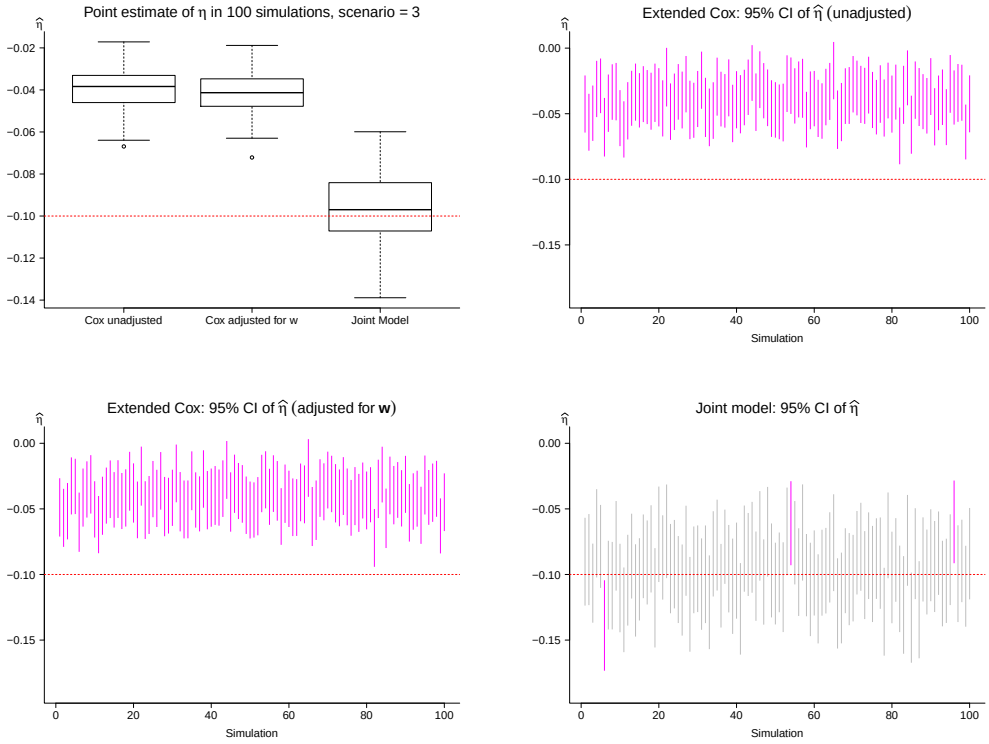


Figure 3.3: The boxplot shows the parameter's estimates of η from 100 simulated datasets, obtained from fitting the extended Cox model (unadjusted and adjusted for baseline covariates) and a joint model with - - - drawn at the true value. Notice that the estimates from the Cox model are biased. The vertical lines on the other three plots represent the 95% interval estimates of η (| if overlaps with the true value and | otherwise). In these 100 simulations the interval estimates obtained from the Cox model are far from the true value of η and none of the confidence intervals covers the true value, while 97 of the confidence intervals obtained from the joint model overlap with the true value of η.

Several extensions have been proposed to the joint model of Equations (3.1a)–(3.1b) including several families of parameterization of the association between the longitudinal and time-to-event outcomes (for instance lagged and cumulative effects), replacing the relative hazard model by competing risk or accelerated failure time models, joint models for multiple time-to-event and longitudinal outcomes. See for instance Hickey *et al.* (2018), Hickey *et al.* (2016) and Rizopoulos (2012), for an overview of these extensions. Li & Luo (2017) proposed a functional joint model for longitudinal and time-to-event data to account for functional predictors in both longitudinal and survival submodels. These topics are out of the scope of this thesis and we mention them only to provide with a scope of possible extensions of joint modelling longitudinal and time-to-event data.

3.1.4 Connection with the missing data framework

From the joint modelling perspective the missing data process of longitudinal data can be interpreted as the occurrence an individual level event that corresponds to an interruption of the longitudinal outcome process. This is because either further measures can no longer be collected or their distribution changes after the event has occurred.

For each subject, i $(i = 1, \ldots, n)$, let $\mathbf{y}_i$ denote the corresponding complete data response vector, as defined in Section 2.1.3. Define $\mathbf{y}_i^{\text{o}}$ and $\mathbf{y}_i^{\text{m}}$ as the observed and missing parts of the longitudinal response vector, respectively, as follows

$$\mathbf{y}_i^{\text{o}} = \left\{ y_i(t_{ij}^{\text{o}}) : t_{ij}^{\text{o}} \leq T_i, j = 1, \ldots, n_i \right\} \text{ and } \mathbf{y}_i^{\text{m}} = \left\{ y_i(t_{ij}^{\text{m}}) : t_{ij}^{\text{m}} \leq T_i, j = 1, \ldots, n_i^* \right\},$$

where $\mathbf{y}_i^{\text{o}}$ contains the longitudinal outcome measures of subject i just before the event time, and $\mathbf{y}_i^{\text{m}}$ represents the measurements that would have been collected until the end of the end of the study, $n_i^* \geq n_i$.

Following the arguments of Section 2.1.3 the distribution of the dropout mechanism can be expressed as the conditional distribution of the time-to-event T_i given the complete vector of longitudinal responses $\mathbf{y}_i$. Due to the conditional independence of

the longitudinal outcome and the time-to-event given the random effects, the dropout mechanism distribution is given by Equation (3.4).

$$\begin{aligned} f(t_i \mid \mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}; \boldsymbol{\theta}) &= \int_{\mathbf{b}_i} f(t_i, \mathbf{b}_i \mid \mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}; \boldsymbol{\theta}) \mathrm{d}\mathbf{b}_i \\ &= \int_{\mathbf{b}_i} f(t_i \mid \mathbf{b}_i, \mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}; \boldsymbol{\theta}) f(\mathbf{b}_i \mid \mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}; \boldsymbol{\theta}) \mathrm{d}\mathbf{b}_i \\ &= \int_{\mathbf{b}_i} f(t_i \mid \mathbf{b}_i; \boldsymbol{\theta}) f(\mathbf{b}_i \mid \mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}; \boldsymbol{\theta}) \mathrm{d}\mathbf{b}_i \end{aligned} \tag{3.4}$$

Notice that the dropout mechanism distribution depends on $\mathbf{y}_i^{\text{o}}$ through the posterior distribution of the random effects, $f(\mathbf{b}_i \mid \mathbf{y}_i^{\text{o}}, \mathbf{y}_i^{\text{m}}; \boldsymbol{\theta})$, which means that joint models correspond to a MNAR mechanism.

A closer inspection of Equation (3.4) reveals that the key component behind the attrition mechanism in joint models is the random effects, $\mathbf{b}_i$,

$$\begin{cases} \text{Longitudinal} & y_i(t \mid \mathbf{b}_i) = \underbrace{\mathbf{x}_i^\top(t)\boldsymbol{\beta} + \mathbf{z}_i^\top(t)\mathbf{b}_i}_{m_i(t)} + \varepsilon_i(t) \\ \text{Terminal} & h_i(t \mid \mathbf{b}_i) = h_0(t) \exp\left\{\mathbf{w}_i^\top \boldsymbol{\gamma} + \eta_L m_i(t)\right\}. \end{cases}$$

Under this joint model, the longitudinal outcome and terminal event submodels share the same random effects, so joint model M belongs to the class of shared-parameter models. The likelihood to dropout is related to the longitudinal outcome profiles.

A relevant feature of joint models of longitudinal and time-to-event data is the connection of the association parameter, η_L, to the type of missing data mechanism. In particular, $\eta_L = 0$ corresponds to MCAR mechanism because once conditioning on available covariates, the dropout process does not depend on $\mathbf{y}_i^{\text{m}}$ or $\mathbf{y}_i^{\text{o}}$. Moreover, when $\eta_L = 0$ the parameters of the two submodels are disjoint sets, so the joint distribution of the dropout and longitudinal process can be factorized as

$$\begin{aligned} f(t_i, \delta_i, \mathbf{y}_i; \boldsymbol{\theta}) &= f(t_i, \delta_i; \boldsymbol{\theta}_t) f(\mathbf{y}_i; \boldsymbol{\theta}_y, \boldsymbol{\theta}_b) \\ &= f(t_i, \delta_i; \boldsymbol{\theta}) \int_{\mathbf{b}_i} f(\mathbf{y}_i \mid \mathbf{b}_i; \boldsymbol{\theta}_y) f(\mathbf{b}_i; \boldsymbol{\theta}_b) \mathrm{d}\mathbf{b}_i, \end{aligned}$$

which implies that $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_y$ can be estimated by separate models (Rizopoulos, 2012).

It is important to note that in practice, missingness can also occur because of censoring. The formulation of the joint model likelihood function is based on the assumption that the censoring process may depend on the history of the observed longitudinal outcome and covariates up to time t, but is independent of future measurements of the longutidinal outcome, i.e. we assumed that censoring is a MAR mechanism.

An additional feature of the shared-parameter models framework is that they can handle also intermittent missingness. Rizopoulos (2012). Thomadakis *et al.* (2019) explores in detail the consequences of biased estimates produced by data missing not at random and its connection with the joint modelling framework, exploring different submodels for the time-to-dropout. Their suggested parameterization of a joint model to completely account for the bias of missing data considers adding in the time-to-dropout submodel the current value of the longitudinal outcome $m_i(t)$ as well as the random effects $\mathbf{b}_i$.

3.2 Prediction with joint models of longitudinal and time-to-event data

By joint modelling longitudinal and time-to-event outcomes it is possible to gain a better understanding of the dynamics of how the two outcomes interact with each other since this is explicitly modelled through, $\boldsymbol{\eta}$, the coefficient of the link function of the random effects, g. Moreover, it is possible to obtain subject-specific predictions which are relevant for instance in medicine. In order to make individualized predictions with a joint model a prediction of the random effects is required (Rizopoulos, 2012). It is convenient to distinguish between *in-sample* and *out-of-sample* predictions (discussed in Section 2.3) since for the latter sometimes it might not be available the required data to predict the random effects.

3.2.1 Prediction of the Random Effects.

Suppose we have already estimated $\widehat{\boldsymbol{\theta}} = (\widehat{h}_0(t), \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\eta}}, \text{vech}(\widehat{B}))$. Recall from Section 3.1 the assumption of conditional independence between the longitudinal and

time-to-event outcomes given the random effects, and the assumption of normally distributed random effects with zero mean and covariance matrix B. Prediction of $\mathbf{b}_i$ is done by treating $\widehat{\boldsymbol{\theta}}$ as known parameters and using the observed data. We can follow an empirical Bayes approach to derive estimates of the random effects from the posterior distribution of $\mathbf{b}_i$ given the observed data:

$$\begin{aligned} f(\mathbf{b}_i \mid t_i, \delta_i, \mathbf{y}_i; \widehat{\boldsymbol{\theta}}) &= \frac{f\left(\mathbf{b}_i, t_i, \delta_i, \mathbf{y}_i; \widehat{\boldsymbol{\theta}}\right)}{f\left(t_i, \delta_i, \mathbf{y}_i; \widehat{\boldsymbol{\theta}}\right)} \\ &= \frac{1}{c} f\left(t_i, \delta_i \mid \mathbf{b}_i; \widehat{\boldsymbol{\theta}}\right) f\left(\mathbf{y}_i \mid \mathbf{b}_i; \widehat{\boldsymbol{\theta}}\right) f(\mathbf{b}_i), \end{aligned} \tag{3.5}$$

where

$$c = \int_{\mathbf{R}} f\left(t_i, \delta_i \mid \mathbf{b}_i; \widehat{\boldsymbol{\theta}}\right) f\left(\mathbf{y}_i \mid \mathbf{b}_i; \widehat{\boldsymbol{\theta}}\right) f(\mathbf{b}_i) \mathrm{d}\mathbf{b}_i$$

is a proportionality constant, and $f(t_i, \delta_i | \mathbf{b}_i; \widehat{\boldsymbol{\theta}})$, $f(\mathbf{y}_i | \mathbf{b}_i; \widehat{\boldsymbol{\theta}})$ and $f(\mathbf{b}_i)$ are the densities described in Equation (3.2) with the elements of $\boldsymbol{\theta}$ replaced by their corresponding estimate.

This posterior distribution does not have a closed-form solution. According to Hinkley & Cox (1979) it is asymptotically normal centred about the mode as $n_i \to \infty$. Rizopoulos (2011) use Markov chain Monte Carlo methods (MCMC) to approximate this posterior distribution. The predictions of the random effects can be taken to be the mean or the mode of the posterior distribution.

$$\widehat{\mathbf{b}}_i = \int_{\mathbb{R}} \mathbf{b}_i f(\mathbf{b}_i \mid t_i, \delta_i, \mathbf{y}_i; \widehat{\boldsymbol{\theta}}) \mathrm{d}\mathbf{b}_i \tag{3.6}$$

$$\widehat{\mathbf{b}}_i = \arg\max_{\mathbf{b}_i} \left\{ f(\mathbf{b}_i \mid t_i, \delta_i, \mathbf{y}_i \widehat{\boldsymbol{\theta}}) \right\} \tag{3.7}$$

Other numerical integration techniques can be used as an alternative to MCMC, for instance the Gauss quadrature technique, available for instance in the R function `integrate()` and `cubature::adaptIntegrate`.

3.2.2 In-Sample Predictions (Forecasting) for the Longitudinal Outcome and the Survival Probabilities.

An appealing characteristic of the joint model is the possibility to predict how individual response trajectories change over time and to make *dynamic predictions* of both longitudinal and time-to-event outcomes as more data is being collected. Figure 3.4 illustrates the idea that fitting a joint model will allow to make subject-specific predictions beyond t_{in_i}. We can predict the survival probability for the whole follow-up period and also the residual survival probability, given survival up to t_{in_i}. Figures 3.4a–3.4f illustrate that as more repeated measures are taken ($\times$), its in-sample prediction and forecast are updated (—), allowing for the prediction of the conditional survival probability given survival up to the most recent data collection point (—).

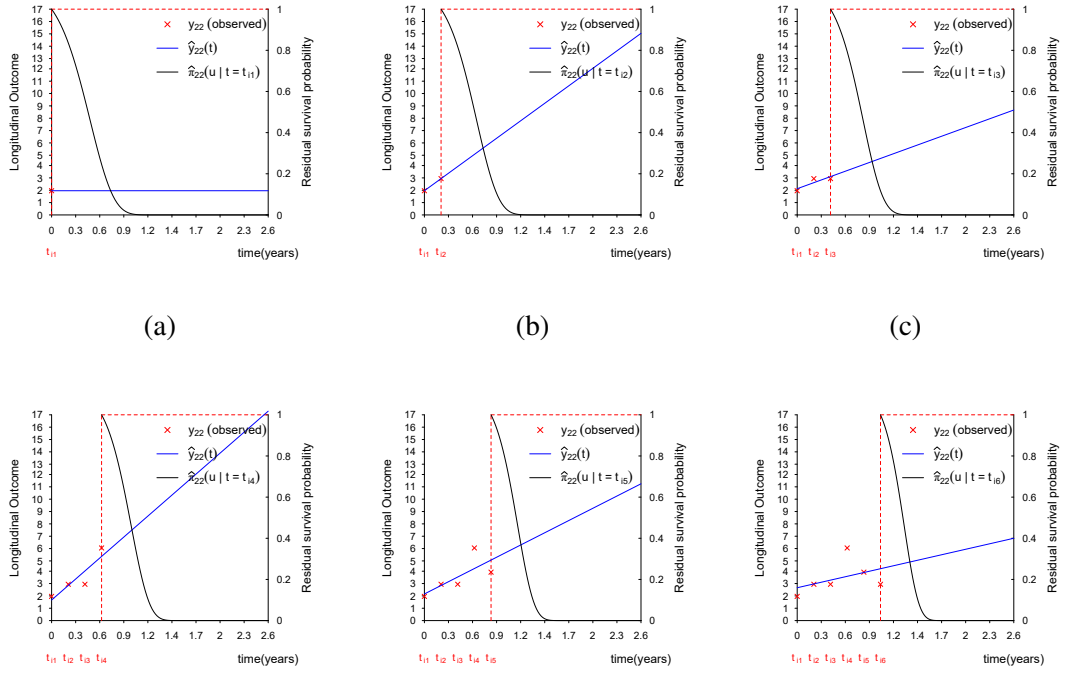


Figure 3.4: (×) are repeated measures of the longitudinal outcome of a fictional subject, y_{ij} at specific time points, $t_{i1}, \ldots t_{i6}$; (—) is the prediction $\widehat{y}_i(t)$ based on the available data up to $t_{ij,j=1,\ldots,6}$ and (—) the prediction of the residual survival probability $\widehat{\pi}_i(u|t), u > t \geq t_{in_i}$ given survival up to t_{ij}. This illustrates the idea that by joint modelling both outcomes are dynamically forecasted as data is being collected at time points $t_{ij},\ j = 1, \ldots, n_i$.

Suppose that we are interested in using $\widehat{\boldsymbol{\theta}} = (\widehat{h}_0(t), \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\eta}}, \text{vech}(\widehat{B}))$ to forecast the longitudinal outcome and the survival probability for subject i who has provided a set of longitudinal measurements $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})$ for whom it is known their survival status up to time $\tau_i \geq t_{in_i}$. The expected value of the longitudinal outcome at time $u > \tau_i$ for this subject given their observed responses up to time t_{in_i} is given by

$$\omega_i(u \mid \tau_i) = \mathbb{E}\left\{ y_i(u) \mid T_i^* > \tau_i, \mathbf{y}_i, \mathcal{D}; \widehat{\boldsymbol{\theta}} \right\},\ u > \tau_i \geq t_{in_i}.$$

Estimation of $\omega_i(u \mid \tau_i)$ can be done by noting that this expectation can be expressed

as

$$\begin{aligned}
\widehat{\omega}_i(u \mid \tau_i) &= \int_{\mathbb{R}} \mathbb{E}\left\{y_i(u) \mid T_i^* > \tau_i, \mathbf{y}_i, \mathbf{b}_i, \mathcal{D}; \widehat{\boldsymbol{\theta}}\right\} f(\mathbf{b}_i \mid T_i^* > \tau_i, \mathbf{y}_i; \widehat{\boldsymbol{\theta}}) \mathrm{d}\mathbf{b}_i \\
&= \int_{\mathbb{R}} \mathbb{E}\left\{y_i(u) \mid \mathbf{b}_i, \mathcal{D}; \widehat{\boldsymbol{\theta}}\right\} f(\mathbf{b}_i \mid T_i^* > \tau_i, \mathbf{y}_i; \widehat{\boldsymbol{\theta}}) \mathrm{d}\mathbf{b}_i \\
&= \int_{\mathbb{R}} \left\{\mathbf{x}_i^\top(u)\widehat{\boldsymbol{\beta}} + \mathbf{z}_i^\top(u)\mathbf{b}_i\right\} f(\mathbf{b}_i \mid T_i^* > \tau_i, \mathbf{y}_i; \widehat{\boldsymbol{\theta}}) \mathrm{d}\mathbf{b}_i \\
&= \mathbf{x}_i^\top(u)\widehat{\boldsymbol{\beta}} + \mathbf{z}_i^\top(u) \int_{\mathbb{R}} \mathbf{b}_i f(\mathbf{b}_i \mid t_i, \delta_i, \mathbf{y}_i; \widehat{\boldsymbol{\theta}}) \mathrm{d}\mathbf{b}_i \\
&= \mathbf{x}_i^\top(u)\widehat{\boldsymbol{\beta}} + \mathbf{z}_i^\top(u)\widehat{\mathbf{b}}_i, \qquad (3.8)
\end{aligned}$$

where $\widehat{\mathbf{b}}_i$ can be predicted as described in section 3.2.1.

According to the natural dependence between the longitudinal and time-to-event outcomes in joint models, a subject that provides longitudinal measurements up to time t_{in_i}, implies survival up to this time point. Therefore, it is more relevant to focus on the conditional probabilities of surviving time $u > \tau_i$ given survival up to $\tau_i \geq t_{in_i}$:

$$\pi_i(u \mid \tau_i) = \Pr\left\{T_i^* \geq u | T_i^* > \tau_i, \mathbf{w}_i, \mathcal{D}; \boldsymbol{\theta}, \mathbf{b}_i\right\} \qquad (3.9)$$

The conditional independence assumption between the two outcomes allows us to express the residual survival probability in the following way (conditioning on $\mathbf{w}_i$ is assumed but omitted from the notation):

$$\begin{aligned}
&\Pr\left\{T_i^* \geq u \mid T_i^* > \tau_i, \mathbf{y}_i, \mathcal{D}; \boldsymbol{\theta}\right\} \\
&= \int_{\mathbf{b}_i} \frac{\Pr\left\{T_i^* \geq u, T_i^* > \tau_i, \mathbf{y}_i, \mathbf{b}_i, \mathcal{D}; \boldsymbol{\theta}\right\}}{f(T_i^* > \tau_i, \mathbf{y}_i; \boldsymbol{\theta})} \mathrm{d}\mathbf{b}_i \\
&= \int_{\mathbf{b}_i} \Pr\left\{T_i^* \geq u, \mid T_i^* > \tau_i, \mathbf{y}_i, \mathbf{b}_i, \mathcal{D}; \boldsymbol{\theta}\right\} f(\mathbf{b}_i \mid T_i^* > \tau_i, \mathbf{y}_i; \boldsymbol{\theta}) \mathrm{d}\mathbf{b}_i \\
&= \int_{\mathbf{b}_i} \frac{S_i\left\{u \mid \mathcal{M}_i(u, \mathbf{b}_i, \boldsymbol{\theta}); \boldsymbol{\theta}\right\}}{S_i\left\{\tau_i \mid \mathcal{M}_i(\tau_i, \mathbf{b}_i, \boldsymbol{\theta}); \boldsymbol{\theta}\right\}} f(\mathbf{b}_i \mid T_i^* > \tau_i, \mathbf{y}_i; \boldsymbol{\theta}) \mathrm{d}\mathbf{b}_i \qquad (3.10)
\end{aligned}$$

where S is the survival function and $t, \mathcal{M}_i(t)$ is the trajectory of the longitudinal outcome up to time τ_i which is approximated by the linear mixed model.

Based on this expectation the estimate of $\pi_i(u \mid t)$ can be obtained using the empirical

Bayes estimate for $\mathbf{b}_i$ is (Rizopoulos, 2011) as follows

$$\widehat{\pi}_i(u|\tau_i) = \frac{\widehat{S}_i(u \mid \widehat{\mathbf{b}}_i; \widehat{\boldsymbol{\theta}})}{\widehat{S}_i(\tau_i \mid \widehat{\mathbf{b}}_i; \widehat{\boldsymbol{\theta}})} + O\left(1/n_i(t)\right), \tag{3.11}$$

where

$$\widehat{S}_i(t \mid \widehat{\mathbf{b}}_i; \widehat{\boldsymbol{\theta}}) = \widehat{S}_0(t)^{\exp(\mathbf{w}_i^\top \widehat{\boldsymbol{\gamma}} + [g(\widehat{\mathbf{b}}_i, t)]^\top \widehat{\boldsymbol{\eta}})} \tag{3.12}$$

is the estimated survival probability at time $t \geq 0$, $\widehat{\boldsymbol{\theta}}$ denotes the maximum likelihood estimates, $\widehat{\mathbf{b}}_i$ the prediction of the random effects for subject i and $n_i(t)$ the number of longitudinal responses for subject i by time t.

This provides a method to obtain point estimates of in-sample predictions. In order to derive prediction intervals for the longitudinal outcome and the (residual) survival probabilities we can evaluate Equations (3.8), (3.11) and (3.12) at the MCMC samples of the random effects, $\widehat{\boldsymbol{b}}_i^{(\ell)},\ \ell = 1, \ldots, L$,

$$\widehat{\omega}_i^{(\ell)}(u \mid \tau_i) = \mathbf{x}_i^\top(t)\widehat{\boldsymbol{\beta}} + \mathbf{z}_i^\top(t)\widehat{\mathbf{b}}_i^{(\ell)} \tag{3.13}$$

$$\widehat{\pi}_i^{(\ell)}(u|\tau_i) = \frac{\widehat{S}_i(u \mid \widehat{\mathbf{b}}_i^{(\ell)}; \widehat{\boldsymbol{\theta}})}{\widehat{S}_i(\tau_i \mid \widehat{\mathbf{b}}_i^{(\ell)}; \widehat{\boldsymbol{\theta}})} \tag{3.14}$$

$$\widehat{S}_i^{(\ell)}(t) = \widehat{S}_0(t)^{\exp(\mathbf{w}_i^\top \widehat{\boldsymbol{\gamma}} + [g(\widehat{\mathbf{b}}_i, t)]^\top \widehat{\boldsymbol{\eta}})},\ t \geq 0, \tag{3.15}$$

3.2.3 Out-of-Sample Predictions for the Longitudinal Outcome and the Survival Probabilities.

Suppose we want to use the joint model estimates, $\widehat{\boldsymbol{\theta}}$, to predict both $y_i(u)$ and $\pi_i(u|t), u > \tau_i \geq t_{n_i}$ for subjects of a totally new data set, $\mathcal{D}^{\text{N}}$. For simplicity, assume all subjects in $\mathcal{D}^{\text{N}}$ provide the set of covariates required to make predictions with $\widehat{\boldsymbol{\theta}}$, i.e. $\mathbf{x}_i(t), \mathbf{z}_i(t), \mathbf{w}_i$.

Some subjects in $\mathcal{D}^{\text{N}}$ might provide no longitudinal outcome measurements at all, others might provide only one longitudinal measurement at baseline (y_{i1} at $t_{i1} = 0$) and

the rest provide a series of repeated measurements $y_{i1}, \ldots, y_{in_i}$ at $t_{i1}, \ldots, t_{in_i}$. In general, we can base these predictions on the method summarized by Equations (3.8), (3.11) and (3.12), which require the prediction of the random effects, $\widehat{\mathbf{b}}_i$. Nonetheless, because subjects in the out-of-sample data set might have no longitudinal outcome measures, the estimation of the random effects requires special attention as this is based on the posterior density described in Equation (3.5).

Predictions for a new subject with no longitudinal measurements.

In this case it is not possible to evaluate the density of the longitudinal outcome, $f(\mathbf{y}_i \mid \mathbf{b}_i)$, in the posterior density of equation 3.5. So an individualized prediction for $\mathbf{b}_i$ cannot be obtained. In order to predict the random effects we simply plug in to Equations (3.8) and (3.9) the mean of the random effects, which by assumptions is zero.

If it is known that such a subject has survived up to time $\tau_i > 0$, prediction of the residual survival probability for $u > \tau_i$ might be relevant. Predictions in this case can be obtained by

$$\widehat{\omega}_i(u \mid \tau_i) = \mathbf{x}_i^\top(u)\widehat{\boldsymbol{\beta}} \tag{3.16}$$

$$\widehat{\pi}_i(u \mid \tau_i) = \frac{\widehat{S}_i(u; \widehat{\boldsymbol{\theta}})}{\widehat{S}_i(\tau_i; \widehat{\boldsymbol{\theta}})} \tag{3.17}$$

$$\widehat{S}_i(t) = \widehat{S}_0(t)^{\exp(\mathbf{w}_i^\top \widehat{\boldsymbol{\gamma}} + [g(\widehat{\mathbf{b}}_i, t)]^\top \widehat{\boldsymbol{\eta}})}, \; t \geq 0, \tag{3.18}$$

and taking the corresponding $(\alpha/2, \; 1 - \alpha/2)$ quantiles, where α is the significance level.

Prediction for a new individual that has one or more longitudinal measurements available.

If it is the case that the new subject we want to make predictions for has at least one longitudinal outcome measurement, we can follow the method described by Equations (3.8), (3.11) and (3.12) to forecast the longitudinal outcome and the (residual) survival probability, and Equation (3.5) to predict the random effects. The elements of the posterior of the posterior distribution of the random effects are the following, depending on whether one or more longitudinal outcome measures are available $(y_{i1}, \ldots, y_{in_i})$ and the last time such a subject is known to be event-free $(\tau_i \geq t_{in_i} \geq 0)$ and $\delta_i = 0$, with $\phi(\cdot; \boldsymbol{\mu}, \Sigma)$ being the density of the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ:

- $n_i = 1$ and $\tau_i = 0$, i.e. one longitudinal measure at baseline and no further information beyond this point regarding the event status of this subject.

$$\begin{cases} f(y_{i1} \mid \mathbf{b}_i; \widehat{\boldsymbol{\theta}}) = \phi(y_{i1};\ \mathbf{x}_{i1}^\top \widehat{\boldsymbol{\beta}} + \mathbf{z}_{i1}^\top \mathbf{b}_i,\ \widehat{\sigma}_\varepsilon^2) & \text{Longitudinal} \quad (3.19a) \\ f(\tau_i, \delta_i \mid \mathbf{b}_i; \widehat{\boldsymbol{\theta}}) = h_i^{\delta_i}(\tau_i) S_i(\tau_i) = 1 & \text{Terminal} \quad (3.19b) \\ f(\mathbf{b}_i; \widehat{\boldsymbol{\theta}}) = \phi(\mathbf{b}_i; \mathbf{0}, \widehat{B}) & \text{Random Effects} \quad (3.19c) \end{cases}$$

- $n_i = 1$ and $\tau_i > 0$, i.e. one longitudinal measure at baseline and it is known the event status of this subject after baseline.

$$\begin{cases} f(y_{i1} \mid \mathbf{b}_i; \widehat{\boldsymbol{\theta}}) = \phi(y_{i1};\ \mathbf{x}_{i1}^\top \widehat{\boldsymbol{\beta}} + \mathbf{z}_{i1}^\top \mathbf{b}_i,\ \widehat{\sigma}_\varepsilon^2) & \text{Longitudinal} \quad (3.20a) \\ f(\tau_i, \delta_i \mid \mathbf{b}_i; \widehat{\boldsymbol{\theta}}) = h_i^{\delta_i}(\tau_i) S_i(\tau_i) & \text{Terminal} \quad (3.20b) \\ f(\mathbf{b}_i; \widehat{\boldsymbol{\theta}}) = \phi(\mathbf{b}_i; \mathbf{0}, \widehat{B}) & \text{Random Effects} \quad (3.20c) \end{cases}$$

- $n_i > 1$ and $\tau_i \geq t_{in_i}$, i.e. repeated longitudinal measures are available and it is known the event status for this subject after the last repeated measure.

$$\begin{cases} f(\mathbf{y}_i \mid \mathbf{b}_i; \widehat{\boldsymbol{\theta}}) = \phi(\mathbf{y}_i;\ X_i \widehat{\boldsymbol{\beta}} + Z_i \mathbf{b}_i,\ \widehat{\sigma}_\varepsilon^2 I_{n_i}) & \text{Longitudinal} \quad (3.21a) \\ f(\tau_i, \delta_i \mid \mathbf{b}_i; \widehat{\boldsymbol{\theta}}) = h_i^{\delta_i}(\tau_i; \widehat{\boldsymbol{\theta}}) S_i(\tau_i; \widehat{\boldsymbol{\theta}}) & \text{Terminal} \quad (3.21b) \\ f(\mathbf{b}_i; \widehat{\boldsymbol{\theta}}) = \phi(\mathbf{b}_i; \mathbf{0}, \widehat{B}) & \text{Random Effects} \quad (3.21c) \end{cases}$$

In Chapter 5 we give an example of the mechanism to obtain out-of-sample predictions in a cross-validation setting for the particular case where the subjects in a "new" data set have exactly one longitudinal measure at baseline, $\tau_i = 0$ and the baseline hazard is modelled parametrically assuming that the time-to-event follows a Weibull(κ, ρ) distribution.

3.2.4 Accuracy of predictions made with joint models

Information criteria, such as the AIC and BIC are useful to assess the overall predictive ability of the joint model encompassing both the longitudinal and the time-to-event outcomes (Rizopoulos, 2012). However, in practice decision making based on a joint model's fit could benefit from obtaining predictions for each outcome, and in such a case it would be of interest to determine how accurately the model predicts each outcome, as discussed in Section 3 regarding the study of QOL and mortality of cancer patients in Ibrahim *et al.* (2010). A specific treatment protocol with chemotherapy/radiotherapy may extend survival or the time-to-relapse, but when the QOL in that prolonged is expected to be poor the clinician would be able to asses whether such a benefit is worth it for the patient. In this sense we consider it is important to assess prediction of joint models by how accurate both longitudinal and time-to-event outcomes are predicted. In Section 2.3.1 we discussed the use of MSE to assess prediction of the longitudinal outcome and IBS for the time-to-event.

Rizopoulos (2012) proposed a dynamic approach based on the Receiver Operating Characteristic Curve (ROC) analysis (Hanley & McNeil, 1982) to assess the ability of a joint model to classify the subjects based on their event status during a relevant time period. Commenges *et al.* (2012) proposed assessing predictions of joint models with the Expected Prospective-Observed Cross-Entropy (EPOCE) estimator using prognostic conditional log-likelihood.

Compare predictions of joint models and extended Cox model

In this section we stated the need for joint modelling when the longitudinal outcome is an internal time-dependent covariate in a Cox model: the regression coefficient es-

timate associated to the time-varying covariate is biased. However, if we set up a full likelihood for the Cox model (call it *marginal Cox model*) estimating the baseline hazard parameters along with the regression coefficients, the fitted values (estimated survival probabilities) of the marginal Cox model are almost as accurate as those obtained by joint modelling the longitudinal and the time-to-event outcomes. Here we refer to the in-sample survival probability estimates from baseline $(t = 0)$ to a relevant time point, say $t = t_i$.

To illustrate this point, consider the following example of simulated data of 500 subjects, simulated from a simple joint model with a random intercept, a random slope and baseline hazard from the Weibull$(\kappa = 4, \rho = 0.5)$ distribution:

$$\text{True}: \begin{cases} \text{Longitudinal} & y_i(t \mid b_{i0}, b_{i1}) = \underbrace{(0.5 + b_{i0}) + (1 + b_{i1})t}_{m_i(t)} + \varepsilon_i(t) \\ \text{Terminal} & h_i(t; \kappa, \rho \mid b_{i0}, b_{i1}) = h_0(t; 4, 0.5) \exp\left\{-0.2 m_i(t)\right\}, \end{cases}$$

where $(b_{i0}, b_{i1}) \sim \mathcal{N}_2(\mathbf{0}, B)$, $\text{vech}(B) = (2, 0, 0.5)$, and $\varepsilon_i(t) \sim \mathcal{N}(0, 5)$. The number of repeated measurements of the longitudinal outcome, n_i, ranges from 3 to 25. The dashed blue lines in the top (right and left) panels of Figure 3.5 show the true longitudinal profiles (assumed unobservable) of two subjects of the simulated data, and the black dots are the repeated measures, assumed to be measured with error. The middle (left and right) panels show the hazard rate based on the true longitudinal profiles (dashed blue line), and the observed repeated measures (black dots). The green line of the bottom row of Figure 3.5 represents the indicator functions whose value is 1 as long as the subject is event-free, and zero after the event time.

Fitting the correct joint model to these data we obtain the following estimates, which are close to the true parameters:

$$\text{JM}: \begin{cases} \text{Longitudinal} & \widehat{m}_i(t \mid \widehat{b}_{i0}, \widehat{b}_{i1}) = (0.59 + \widehat{b}_{i0}) + (1 + \widehat{b}_{i1})t + \varepsilon_i(t) \\ \text{Terminal} & h_i(t; \widehat{\kappa}^{(\text{JM})}, \widehat{\rho}^{(\text{JM})} \mid \widehat{b}_{i0}, \widehat{b}_{i1}) = h_0(t; 4.15, 0.49) \exp\left\{-0.18 m_i(t)\right\}. \end{cases}$$

From top to bottom, the solid red lines in Figure 3.5 represent the fitted longitudinal profiles with the joint model of the two selected subjects, their hazard rate and their estimated survival probabilities.

By fitting an marginal Cox model, which ignores the the fact that $m_i(t)$ is an internal covariate for the survival analysis model, we get biased estimates, in particular the regression coefficient of the effect of $m_i(t) \to h_i(t)$ and κ, the shape parameter of the baseline hazard:

$$\text{Marginal Cox} : \Big\{ \text{Terminal} \quad h_i(t_{ij}; \widehat{\kappa}^{(\text{Cox})}, \widehat{\rho}^{(\text{Cox})}) = h_0(t_{ij}; 3.74, 0.46) \exp\left\{-0.7 y_i(t_{ij})\right\}.$$

Full log-likelihood of the marginal Cox model with time-varying covariate $y(t)$ is given by

$$\begin{aligned}
\ell(\theta) &= \log \left(\prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[h_0(t_{ij}) \exp\{\eta y_{ij} + \mathbf{w}_i^\top \boldsymbol{\gamma}\} \right]^{\delta_{ij}} \right. \\
&\qquad \left. \exp\left\{ - \int_{t_{i(j-1)}}^{t_{ij}} h_0(s) \exp\{\eta y_{ij} + \mathbf{w}_i^\top \boldsymbol{\gamma}\} \mathrm{d}s \right\} \right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n_i} \delta_{ij} \left(\log h_0(t_{ij}) + \eta y_{ij} + \mathbf{w}_i^\top \boldsymbol{\gamma} \right) \\
&\qquad - \left[H_0(t_{ij}) - H_0(t_{i(j-1)}) \right] \exp\{\eta y_{ij} + \mathbf{w}_i^\top \boldsymbol{\gamma}\}.
\end{aligned}$$

The parameter estimation is done by maximizing this log-likelihood function and this can be done by any standard optimization routine.

Note that the cumulative hazard for subject i at time t is given by

$$\begin{aligned}
H_i(t) &= \sum_{j: t_{ij} \le t} \int_{t_{i(j-1)}}^{t_j} h_0(s) \exp\{\eta y_{ij} + \mathbf{w}_i^\top \boldsymbol{\gamma}\} \mathrm{d}s \\
&= \sum_{j: t_{ij} \le t} \left[H_0(t_{ij}) - H_0(t_{ij-1}) \right] \exp\{\eta y_{ij} + \mathbf{w}_i^\top \boldsymbol{\gamma}\} \\
&= H_i(t_{i(j-1)}) + \left[H_0(t) - H_0(t_{i(j-1)}) \right] \exp\{\eta y_{ij} + \mathbf{w}_i^\top \boldsymbol{\gamma}\},
\end{aligned}$$

with $H_i(0) = 0$.

Since we estimate as well the parameters of the baseline hazard, we can estimate $\widehat{H}_0(t)$ at any time t by interpolating $\widehat{H}_i(t)$ between the time intervals $[t_{i(j-1)}, t_{ij}),\ j = 1, \ldots, n_i, n_{i+1}$, with $t_{i0} = 0$ and $t_{in_{i+1}} = \tau$, where τ is an arbitrary evaluation time point, possibly an administrative censoring time. By being able to estimate $\widehat{H}_i(t)$ for

all t, the estimated survival probabilities are given by $\widehat{S}_i(t) = \exp\{-\widehat{H}_i(t)\}$.

If we compare the IBS computed for the model evaluated at the true parameter values, the fitted Cox model and the fitted joint model, we can see that their IBS are very close to each other (multiplied by 100 to facilitate the comparison):

- $100 \times \text{IBS}^{(\text{True})} = 34.68$,
- $100 \times \text{IBS}^{(\text{Cox})} = 34.05$,
- $100 \times \text{IBS}^{(\text{JM})} = 34.14$.

For a closer inspection of the IBS estimates, we analyze the contributions to the IBS of two subjects of the simulated data. Figure 3.5 shows the data and estimates of two subjects (left and right), with the longitudinal profiles in the top row, hazard rates in the middle and survival curves at the bottom. The bottom panel illustrates the contribution of each individual to the total IBS, calculated as the squared difference between the estimated survival probabilities $\widehat{S}_i(t)$ (joint model —, Cox model extrapolated between consecutive time points - -, and survival function evaluated at the true parameters - -), and the event indicator $I(T^* > t)$ (—), integrated over time. Comparing the estimated survival curves of each fitted model against the survival curve of the model evaluated at the true parameters, we see that the Cox model underestimates the survival probabilities of both subjects. In terms of prediction error, the joint model does better than the Cox model for the subject on the left since

$$10 \times \text{IBS}^{(\text{JM})} = 3.60 < 10 \times \text{IBS}^{(\text{Cox})} = 4.93$$

(as reference, the error or the model at the true parameters is $10 \times \text{IBS}^{(\text{True})} = 3.43$), but for the subject on the right, the Cox model does better than the joint model since

$$10 \times \text{IBS}^{(\text{JM})} = 2.12 > 10 \times \text{IBS}^{(\text{Cox})} = 1.65$$

($10 \times \text{IBS}^{(\text{True})} = 3.73$ for model evaluated at the true parameters).

Figure 3.6 summarizes the comparisons of the prediction error between the Cox and the joint models. Here we computed at a subject level the difference between the IBS of each fitted model (joint model and Cox model) and the IBS computed with the true model with red ○ indicating the subjects for which the joint model has smaller

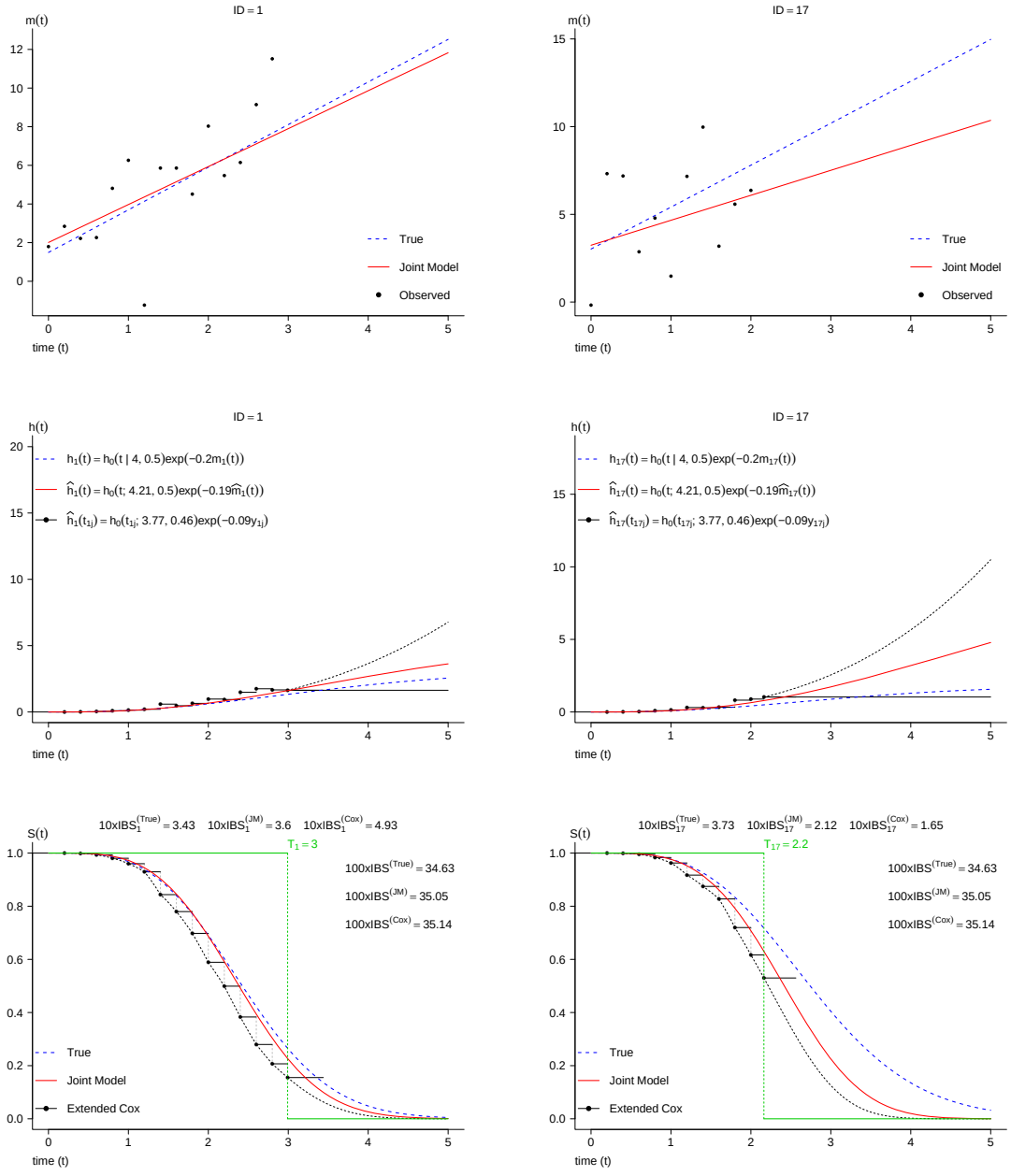
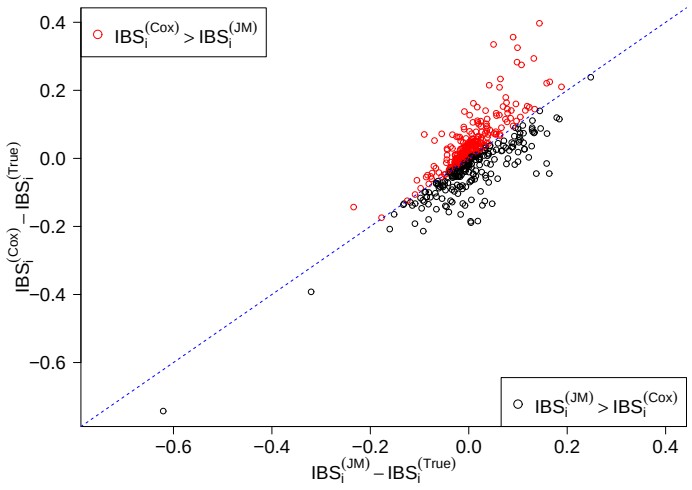


Figure 3.5: Longitudinal profiles, hazard and survival curves of two subjects from the simulated data. Top: true longitudinal profile $m_i(t)$ (- -), repeated measures y_{ij} (•), estimated profile with the joint model (—). Middle: hazard function evaluated at the true parameters - -, hazard estimated with the joint model (—), hazard estimated with the Cox model (–•–) with interpolation between consecutive time points (- -). Bottom: Indicator of the event status at time t $I(T^* > t)$ (—), survival function evaluated at the true parameters - -, survival curve estimated with the joint model (—), survival curve estimated with the Cox model (–•–) with interpolation between consecutive time points (- -).

prediction error than the Cox model and black $\circ$ the other way around (as reference a 45 degree line is drawn (- -). The table at the bottom of Figure 3.6 is a 2×2 table showing the frequency for each fitted model giving a smaller prediction error than the true model and the frequency of giving larger prediction error than the true model. From the plot and table we point out two results:

- The Cox model yields a smaller IBS than the true model in 230 (46%) subjects, almost as often as the frequency at which the joint model yields a smaller IBS than the true model which is in 258 (52%) subjects.
- The joint model outperforms the Cox model in terms of prediction error in 270 (54%) subjects, and the Cox outperforms the joint model in 230 (48%) subjects.



	Extended Cox		
Joint Model	$\text{IBS}_i^{(\text{Cox})} > \text{IBS}_i^{(\text{True})}$	$\text{IBS}_i^{(\text{Cox})} < \text{IBS}_i^{(\text{True})}$	Sum
$\text{IBS}_i^{(\text{JM})} > \text{IBS}_i^{(\text{True})}$	197	45	242
	(0.39)	(0.09)	(0.48)
$\text{IBS}_i^{(\text{JM})} < \text{IBS}_i^{(\text{True})}$	73	185	258
	(0.15)	(0.37)	(0.52)
Sum	270	230	500
	(0.54)	(0.46)	(1.00)

Figure 3.6: The plot shows the difference between the IBS of the true model and each fitted model (against the joint model on x-axis and against the Cox model on y-axis). The 45 line - - separates subjects with $\text{IBS}^{(\text{Cox})} > \text{IBS}^{(\text{JM})}$ (red ○) from subjects with $\text{IBS}^{(\text{Cox})} < \text{IBS}^{(\text{JM})}$ black ○. The table compares $\text{IBS}^{(\text{Cox})}$ and $\text{IBS}^{(\text{JM})}$ against $\text{IBS}^{(\text{True})}$ at individual level.

The bottom line is that the fitted values $\widehat{S}_i(t)$ produced with the marginal Cox model are on average almost as good, in terms of IBS, as the fitted values of the joint model. This suggests that we can still make accurate predictions with the marginal Cox model if the baseline hazard is also estimated. However, the estimate and interpretation of the regression coefficient of y_{ij} in the Cox mode as the "effect of $m_i(t)$ in the hazard of the event" will be wrong because this estimate is biased.

3.3 Shared Random Effects Joint Model for Longitudinal, Recurrent and Terminal Events Data

The joint modelling framework is not restricted to longitudinal and terminal event outcomes, and can be extended to model multiple outcomes simultaneously. For instance, the repeated measures of a continuous variable, and recurrences of an event both may be terminated by a major failure event such as death. The dynamics of this situation can be illustrated by the AIDS study in Abrams *et al.* (1994), where patients with AIDS were allocated to two treatment arms and followed-up for a median of 16 months. During the follow-up period, CD4 cell count was repeatedly taken on the patients and they were also monitored for the development of new or recurrent opportunistic disease, which means disease progression. The assumed relationship between these variables is depicted in Figure 3.7. It is conjectured that lower CD4 count is associated with higher risk of opportunistic disease, which in turn is associated with a higher risk of death, and lower CD4 count is associated with a higher risk of death from HIV-unrelated causes (Obel, 2012).

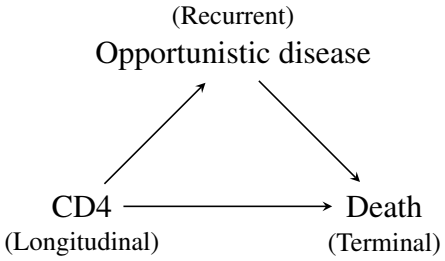


Figure 3.7: Hypothesized relationship between CD4 cell count, recurrences of opportunistic disease and mortality.

A usual assumption when analyzing data from longitudinal studies is that repeated measures of the outcome are collected at noninformative observation times, i.e. they are independent and thus the observation times carry no information about the repeated measures. This assumption is valid in clinical trials with fixed observation times, or in observational studies with random observation times. However, if observation times are somehow associated to the repeated measures, ignoring this feature may produce

selection bias due to possible dependence of the observation times and the marker. For example, Liu *et al.* (2008) study the cost accrual process described by recurrent hospital visits in the presence of potential death. Patients in advanced stages of a disease might visit the hospital more often, and their health status measured by biomarkers at each visit is worse. Therefore, abnormal values of the biomarkers are overrepresented and normal values underrepresented, resulting in a selection bias. By joint modelling of the cost accrual process (longitudinal outcome), the recurrent visits to the hospital and death it is possible to address both biases induced on the one hand by the dependence between the recurrent events process and the costs-accumulation process, and on the other hand the informative censoring. Finally, sometimes interest lies in exploring the joint distribution of a longitudinal outcome, a recurrent event a terminal event. To address this kind of problem in the joint modelling framework, we require to specify a submodel for 1) the longitudinal outcome, 2) the recurrent events and 3) the terminal-event.

Assume that the evolution in time of a quantity of interest, y, can be represented by a continuous-time Gaussian stochastic process $\{y(t), t > 0\}$. Denote by y_{ij} an observation of this process at time t_{ij} for subject $i, i = 1, \ldots, n$, and occasion $j, j = 1, \ldots, n_i$. Assume also a recurrent event process that is thought to be related to the continuous-time process, and a terminal event presumed to be associated to both the continuous-time and the recurrent event processes.

Let T_{ik}^* be the k^{th} recurrent event time for subject i, $(k = 1, \ldots, K_i)$, C_i a censoring time (different from the terminal event), T_i^* the time for the terminal event (for example the time to death).

Figure 3.8 illustrates how the outcome data would look like in the three outcome joint modelling setting. This diagram is similar to Figure 3.1, but with recurrent event times added and shown as R_k^R and identified with $+$ on the time axis. Note that the repeated measures time points need not be the same as the recurrent event times.

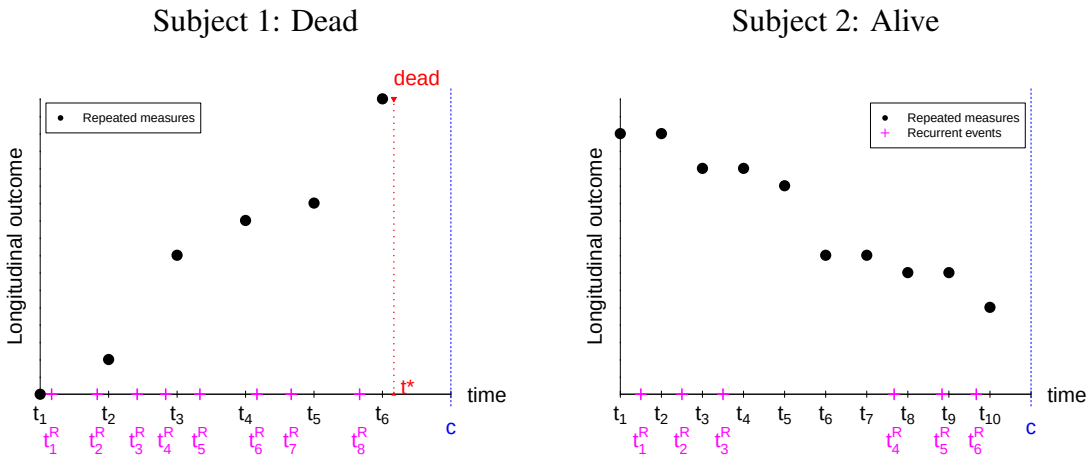


Figure 3.8: Subject 1 (left) has 6 repeated measures of the longitudinal outcome (•), 8 recurrent events at times $t_1^R, \ldots, t_8^R$, and died at time t^*; Subject 2 (right) has 10 repeated measures (•), 6 recurrent events at times $t_1^R, \ldots, t_6^R$ and is alive by the censoring time C.

The observed data of the recurrent events corresponds to

$$T_{ik} = \min(T_{ik}^*, C_i, T_i^*), \tag{3.22a}$$
$$\delta_{ik}^R = \mathbb{1}(T_{ik} = T_{ik}^*), \tag{3.22b}$$

where δ_{ik}^R is a binary indicator of censorship for the recurrent event process. That is,

$$\delta_{ik}^R = \begin{cases} 1 & \text{if time } T_{ik}^* \text{ corresponds to an observed recurrent event,} \\ 0 & \text{if time } T_{ik}^* \text{ corresponds to either censored or terminal event time} \end{cases}$$

The last follow-up time for subject i, T_i, is either a censoring time or a death time, corresponding to

$$T_i = \min(C_i, T_i^*), \tag{3.23a}$$
$$\delta_i = \mathbb{1}(T_i = T_i^*), \tag{3.23b}$$

where δ_i is the binary indicator of censorship for the terminal event process. That is,

$$\delta_i = \begin{cases} 1 & \text{if time } T_i^* \text{ corresponds to death time,} \\ 0 & \text{if time } T_i^* \text{ corresponds to censoring time.} \end{cases}$$

So the data we would observe in a sample of n subjects is

$$\mathcal{D} = \left\{T_{ik}, \delta_{ik}^{T}, T_i, \delta_i, y_{ij}, \text{ for } j = 1, \ldots, n_i, k = 1, \ldots, K_i \text{ and } i = 1, \ldots, n\right\}.$$

Similarly to the specification of a joint model of a longitudinal outcome and a terminal event, in the case where additionally a recurrent event is jointly modeled we specify a regression model for each outcome and connect them with functions of the fixed and random effects. The longitudinal outcome is expressed as a linear mixed model, and the recurrent events and the terminal event are expressed as proportional hazards models with a common random effect that acts multiplicatively on the hazard rate, as discussed in Section 2.2.1. The general specification of a joint model for a longitudinal outcome and recurrent and terminal events with random effects $\mathbf{v}_i = (\mathbf{b}_{0i}, u_i)$ is described by Equations (3.24a)–(3.24c).

The following set of Equations is the general specification of a joint model for longitudinal, recurrent events and a terminal event data with random effects $\mathbf{v}_i = (\mathbf{b}_{0i}, u_i)$

$$\begin{cases} \underset{\text{(Longitudinal)}}{y_i(t \mid \mathbf{b}_i)} = m_i(t) + \varepsilon_i(t) = \mathbf{x}_i(t)\boldsymbol{\beta} + \mathbf{z}_i(t)\mathbf{b}_i + \varepsilon_i(t) & \text{(3.24a)} \\ \underset{\text{(Recurrent)}}{r_i(t \mid \mathbf{v}_i)} = u_i r_0(t) \exp\left\{\mathbf{w}_{Ri}^{\top}\boldsymbol{\gamma}_R + g_{(\text{Rec})}\left(\mathbf{b}_i, t\right)^{\top}\boldsymbol{\eta}_R\right\} & \text{(3.24b)} \\ \underset{\text{(Terminal)}}{h_i(t \mid \mathbf{v}_i)} = u_i^{\alpha} h_0(t) \exp\left\{\mathbf{w}_{Ti}^{\top}\boldsymbol{\gamma}_T + g_{(\text{Ter})}\left(\mathbf{b}_i, t\right)^{\top}\boldsymbol{\eta}_T\right\}, & \text{(3.24c)} \end{cases}$$

where

$y_i(t) =$ the longitudinal outcome with measurement error,
$m_i(t) = \mathbf{x}_i(t)\boldsymbol{\beta} + \mathbf{z}_i(t)\mathbf{b}_i$ is the true and unobserved longitudinal outcome,
$r_i(t) =$ hazard function of the recurrent event,
$r_0(t) =$ baseline hazard of the recurrent event, a positive real-valued function,
$h_i(t) =$ hazard function of the terminal event,
$h_0(t) =$ baseline hazard of the terminal event, a positive real-valued function
$\mathbf{x}_i(t) = p$-vector of the fixed effects for subject i at time t,
$\mathbf{z}_i(t) = q$-vector of the random effects for subject i at time t,
$\mathbf{w}_{Ri} =$ baseline covariate vector of the recurrent event for subject i,
$\mathbf{w}_{Ti} =$ baseline covariate vector of the terminal event for subject i,
$\boldsymbol{\beta} = p$-vector of fixed effects regression coefficient,
$\boldsymbol{\gamma}_R =$ vector of regression coefficients of the recurrent event,
$\boldsymbol{\gamma}_T =$ vector of regression coefficients of the terminal event,
$\mathbf{b}_i = q$-vector of random effects of the longitudinal outcome for subject i,
$u_i =$ random effect shared by the recurrent and terminal events for subject i,
$g_{\text{(Rec)}} =$ link function between the longitudinal outcome and recurrent event,
$g_{\text{(Ter)}} =$ link function between the longitudinal outcome and terminal event,
$\boldsymbol{\eta}_R =$ vector of the association coefficients between $m_i(t)$ and $r_i(t)$,
$\boldsymbol{\eta}_T =$ vector of the association coefficients between $m_i(t)$ and $h_i(t)$,
$\alpha =$ association between recurrent and terminal events.

The measurement errors $\varepsilon_i(t) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_\varepsilon)$ represent the within-subject variability of the longitudinal outcome and are assumed independent and normally distributed random variables with constant variance σ^2_ε. The vectors $\mathbf{b}_i \sim \mathcal{N}_q(\mathbf{0}, B)$ are the subject-specific random effects assumed to follow a multivariate normal distribution centered at zero with covariance matrix B.

The random effects shared by the recurrent and terminal event $u_i \overset{\text{iid}}{\sim} \text{Gamma}(\phi^{-1}, \phi^{-1})$ are positive-valued random variables with variance parameter ϕ that must satisfy $\mathbb{E}(u_i) =$

1 and $\text{var}(u_i) = \phi$. Common choices for the distribution of u_i are Gamma and Log-normal. This random effect represents cluster-specific heterogeneity, possibly due to unobserved factors that modify multiplicatively the hazard rate of a group/cluster. It is presumed to act on both the recurrent events and the terminal event processes, and the effect is modulated by the parameter α. So when $\alpha = 1$ the effect of u_i is the same for the recurrent events and for the terminal event. When $\alpha > 0$, the risk of recurrence and death are positively associated with a different effect of the frailty on the two hazards.

By the standard assumption of joint modelling measurement errors and random effects are independent, $\varepsilon_i(t) \perp \mathbf{b}_i$, $\mathbf{b}_i \perp u_i$ and $u_i \perp \varepsilon_i(t)$. Hence the three outcomes are conditionally independent given the random effects. When independence is not reasonable, then it is necessary to model their covariance structure.

It is also assumed that:

- The recurrent, terminal and censoring processes are continuous.
- In a small interval $[t, t + d]$ the terminal event occurs first.
- The intensities or hazards of recurrent events and terminal event processes of patient i do not change after C_i.

Like the bivariate joint model case described in section 3.1 in this three-outcome joint model the link functions that connect the longitudinal outcome with the terminal event could be in principle any function of the random effects, $\mathbf{b}_i$. The regression coefficients associated to the link functions, $g_{(\text{Rec})}$ and $g_{(\text{Ter})}$: $\boldsymbol{\eta}, \boldsymbol{\varphi}$ and α, quantify the strength of the association between the outcomes.

The purpose of joint modelling include:

- Estimation of the model parameters.
- Hypotheses testing of all regression coefficients. The distinctive features of joint models relative to the marginal models are the association parameters $\boldsymbol{\eta_R}$, $\boldsymbol{\eta}_T$ and α.
- Prediction of the three outcomes. The joint model we have described considers the terminal event as outcome as the end point of a causal path between the three outcomes, so a natural prediction of interest would be the probability of

observing the terminal event in the future $(t+s)$, provided the history up the present, (t), i.e.

$$\Pr\left\{T_i \leq t+s \mid T_i \geq t, \mathcal{F}_i(t)\right\},$$

where $\mathcal{F}_i(t)$ denotes the full history up to t of the three outcomes and covariates). However, it might also be of interest to predict simultaneously the three outcomes.

- Goodness of fit to investigate the joint model that best describes data of the three outcomes taking into account their associations.
- Causal inference about hypothesized paths among the outcomes and covariates.

The parameters to estimate are

$$\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \boldsymbol{\gamma}_R, \boldsymbol{\gamma}_T, \boldsymbol{\eta}_R, \boldsymbol{\eta}_T, \alpha, r_0(t), h_0(t), \sigma_\varepsilon^2, \text{vech}(B), \phi\right).$$

Estimation of θ is based on maximum likelihood principles. The likelihood function is constructed by assuming conditional independence of the three process given the random effects $\mathbf{v}_i = (\mathbf{b}_i, u_i)$. Let $|\mathbf{v_i}|$ denote the cardinality of the vector of random effects. The contribution of subject i to the likelihood is given by Equation (3.25).

$$L_i(\boldsymbol{\theta} \mid \mathcal{D}) = \int_{|\mathbf{v}_i|} L_i^y L_i^R L_i^T f(\mathbf{v}_i) \mathrm{d}\mathbf{v}_i, \text{ where} \tag{3.25}$$

$$L_i^y = (2\pi\sigma_\varepsilon^2)^{-n_i/2} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2} \left\|\mathbf{y}_i - X_i\boldsymbol{\beta} - Z_i\mathbf{b}_i\right\|^2\right\}, \tag{3.25a}$$

$$L_i^R = \prod_{k=1}^{K_i} \left[r_i(t_{ik} \mid \mathbf{v}_i)\right]^{\delta_{ik}} \exp\left\{-\int_{t_{i(k-1)}}^{t_{ik}} r_i(t \mid \mathbf{v}_i)\mathrm{d}t\right\}, \tag{3.25b}$$

$$L_i^T = \left[h_i\left(t_i \mid \mathbf{v}_i\right)\right]^{\delta_i} \exp\left\{-\int_0^{t_i} h_i\left(t \mid \mathbf{v}_i\right) \mathrm{d}t\right\}, \tag{3.25c}$$

$$f(\mathbf{v}_i) = f(\mathbf{b}_i) f(u_i). \tag{3.25d}$$

The contribution of the longitudinal outcome has the form of the density of a n_i-dimensional normal distribution with mean m_{ij} (the linear predictor of the linear mixed model) and covariance matrix $\sigma^2 I_{n_i}$. The hazard rate of the recurrent and terminal events, $r_i(t|\cdot)$ and $h_i(t|\cdot)$ are as defined in Equations (3.24b) and (3.24c) respectively.

And $f(\mathbf{b}_i)$ and $f(u_i)$ are the densities of the random effects.

Estimation of this three outcome joint model is based on maximizing the likelihood function of Equation (3.25). Liu *et al.* (2008) and Liu & Huang (2009) proposed using the EM algorithm to estimate this joint model, specifying the model with piecewise constant baseline hazards. Król *et al.* (2016) followed a penalized maximum likelihood approach using the Marquardt algorithm (Marquardt, 1963) to optimize the likelihood (3.25), which combines the Newton–Raphson and steepest descent algorithms. Penalization of the likelihood is performed to obtain smooth estimates of the baseline hazards of the recurrent and terminal events which are approximated by M-splines[1]. Rondeau *et al.* (2012) follows the same approach as Król *et al.* (2016), but with parametric Weibull and piecewise constant baseline hazards.

The integrals of the random effects require numerical methods, with the Gauss(–Hermite) quadrature technique the most commonly used, see for instance, Król *et al.* (2016), Commenges & Jacqmin-Gadda (2015), Rizopoulos (2012), Liu *et al.* (2008), Liu & Huang (2009), among others.

Software packages developed for the analysis of longitudinal data in the presence of informative censoring: R packages JM, RJAGS and JSM; Matlab and Fortran; WinBUGS and Fortran90. The R function `trivPenal` of the R package frailtypack and SAS Proc NLMIXED both have capabilities for joint modelling of longitudinal data and recurrent events subject to a terminal event.

3.4 Other Approaches for Joint Modelling of Longitudinal and Time-to-Event Data

Other approaches to joint modelling longitudinal and time-to-event data are Latent Class Joint Models (Commenges & Jacqmin-Gadda, 2015; Proust-Lima & Taylor, 2009) and Bayesian analysis of joint models (Ibrahim *et al.*, 2001, 2004, 2010; Liu & Li, 2016; Rizopoulos & Ghosh, 2011).

[1] M-splines can be considered as a normalized version of B-splines with unit integral within boundary knots.

Alsefri *et al.* (2020) provides with a recent methodological review on Bayesian joint models of longitudinal and time-to-event data with focus on type of outcomes, model assumptions, association structure, estimation algorithm, dynamic prediction and software implementation.

Chapter 4

Joint model for frailty, recurrent falls and mortality with the CARE75+ data

4.1 Introduction

Frailty is a distinctive health state related to the ageing process that describes how the body gradually loses its built-in reserves, leaving it vulnerable to dramatic and sudden changes in health triggered by apparently minor illnesses, such as a chest infection, that otherwise the body could likely overcome. Frailty is associated with adverse outcomes such as frequent falls, disability, hospitalization, and mortality (Clegg *et al.*, 2013). Although the prevention and treatment of frailty is an aspiration of researchers in the ageing field that remains enigmatic (Nowak & Hubbard, 2009).

We dedicate this chapter to explore via joint modelling the relationship between frailty and mortality using data from the Community Ageing Research 75+ (CARE 75+) study, a population study of elderly people conducted in the Yorkshire and Humber region in England. This relationship has been studied before, but it has not been analyzed by joint modelling frailty and mortality, and this is the main relationship we are interested in. An additional element that we incorporate to the analysis is falls as a

time-to-event recurrent outcome, and we explore how this third element can be modeled to describe the frailty-falls-mortality relationship in the CARE75+ data set. We explore these relationships with different specifications of joint models, making different assumptions about the role of falls and about the functional form in which frailty and mortality are associated.

This chapter contains four sections. In the first section we succinctly describe the CARE75+ study and the three outcome variables of our analyses: frailty, falls and mortality, and the covariates we use to help understand the relationship between the outcomes. The second section describes how we specified a 3-outcome joint model for frailty, falls and mortality and the process we followed to select the variables for our final model. Model diagnostics of the fitted joint model is shown in the third section. Finally, in the last section of this chapter we give our conclusions of the fitted model discussing the most relevant aspects, and we propose an alternative joint model for frailty and mortality with falls being assumed an exogenous time-varying covariate.

4.2 The CARE75+ study

The CARE 75+ study is a longitudinal cohort of older people with frailty for observational research. This study aims to understand why some people remain fit and resilient in older age while others develop health problems and frailty and to determine what (treatable) problems have a major impact on the quality of life in older age.

To avoid confusion, recall that in Section 2.2.1 we contrasted the use of the concept of *frailty* in two different contexts: one in survival analysis models and the other in medicine and public health. As we just described, in medicine and geriatric studies *frailty* refers to a health state of the elderly and, as we explained in Section 2.2.1, in survival analysis and joint models, *frailty* stands for the notion of a random effect in statistics acting multiplicatively in the hazard rate of a time-to-event model. In order to prevent from mixing the two different uses of this concept, in the remaining part of this document we restrict *frailty* to its use in the medical context. In survival analysis

and joint models we will use instead the concept of *random effect*, just as is customary in linear mixed models and ANOVA.

Ageing is associated with functional decline and loss of autonomy, and age is a major risk factor for a wide spectrum of clinical conditions, including cardiovascular disease, cognitive impairment, and physical disability. As life expectancy continues to rise, preserving physical functioning in advanced age has emerged as a major clinical and public health priority (Sourdet *et al.*, 2012).

In epidemiological and intervention studies of falls it is important to consider carefully the definition used for *falls* as this may vary between studies (Masud & Morris, 2001). The World Health Organization defines a *fall* as an event which results in a person coming to rest inadvertently on the ground or floor or other lower level. Fall-related injuries may be fatal or non-fatal though most are non-fatal[1]. And the National Institute for Health and Care Excellence (NICE) defines a fall as an unintentional or unexpected loss of balance resulting in coming to rest on the floor, the ground, or an object below knee level[2].

Previous research of frailty, falls and their relationship with mortality suggests that frailty is associated with an increased risk of falls (Ensrud *et al.*, 2007; Samper-Ternent *et al.*, 2012) a greater risk of fracture, disability, and falls in women aged 55 and older in 10 countries, with similar patterns across age and geographic region Tom *et al.* (2013). The direct association between frailty and falls is also pointed out by Cheng & Chang (2017) resulting from meta-analysis of 10 studies indicating that compared to robust older adults, frail older adults are more likely to experience recurrent falls. Those who fell were more likely to be women, not married, had prior falls, more functional problems and poorer health (Samper-Ternent *et al.*, 2012). Additionally, frailty is associated with a higher risk of mortality (Chang & Lin, 2015).

On a different line, Nowak & Hubbard (2009) claim that falling should be recognized as a macrostate indicator of complex system failure rather than a specific disorder. If falling in the frail is truly a manifestation of complex system failure then searching for the cause of the incident is futile since this single cause does not exist. After all,

[1]World Health Organization: https://www.who.int/news-room/fact-sheets/detail/falls

[2]NICE: https://www.nice.org.uk/guidance/qs86

failure of a complex system is the cumulative effect of multiple faults and it is only the intricate linking of these detrimental processes that leads to the overt collapse of the system. Predictors of falls include muscle strength of lower extremities, postural competence/lateral balance, impaired vision, cognitive impairment and taking more than four medications or particular groups of drugs.

In the analyses of mortality rates from falls in the elderly (per 100,000 persons) using data from the national statistics systems of the USA, the Netherlands and Spain Hartholt *et al.* (2018, 2019); Padrón-Monedero *et al.* (2017) determined an increasing trend of mortality from falls in these three countries between the years 2000 and 2016.

Within the CARE75+ study researchers keep track of mortality and several potential risk factors for mortality and frailty, including gender, height, weight, ethnicity, marital status, education level, frequency of visits to a general practitioner, smoking and alcohol consumption habits, history of bone fractures and comorbidities.

The data to quantify frailty (in the Edmonton Frail Scale) for participants of the CARE75+ study are collected intermittently at set times, approximately every six months. Participants of the study have not been recruited all at the same time. By the administrative censoring time for this analysis (31st May 2017), some participants had been in the study for two months and some for over two years (Figure 4.1 shows the data collection time points for each participant). At each interview, participants answer a questionnaire from which the frailty score is computed. They are also asked ”How many times did you fall in the last 12 months?”. We are interested in analyzing falls as a recurrent event using the time-to-event methodology for recurrent events described in Section 2.2.2. However, the data set has no information related to the exact time at which falls occurred, so we estimated the time-to-fall under the assumption that falls are equally spaced in the relevant time interval.

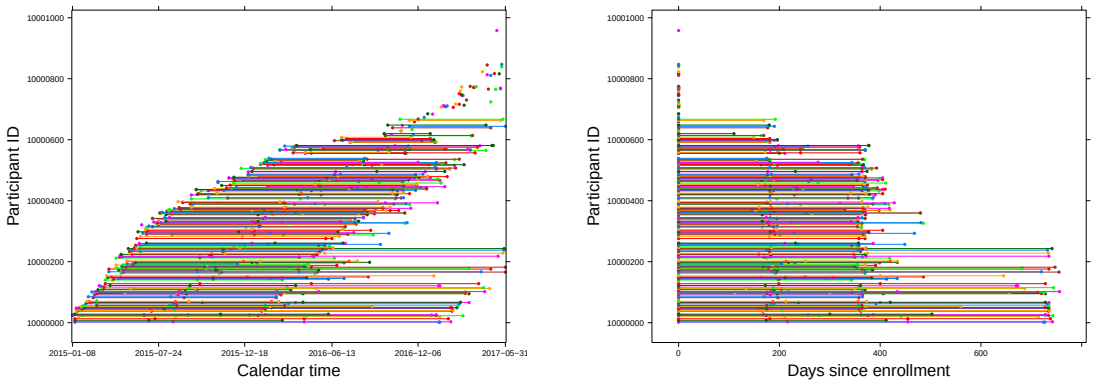


Figure 4.1: The plots show data collection time points of each participant of the CARE75+ study: (left side) interviewing dates and (right side) days since enrollment. The connecting lines between consecutive points is the time interval from one interview to the next.

4.2.1 Frailty, falls and mortality

The data set that we analyzed contains the records of 282 participants aged between 76 and 100 years old (mean age is 83). The mean frailty score of this group of people is 3.7 and the mean number of falls within a 6 months period is 1.6 (median is zero). Seventeen out of the 282 participants died before the censoring date (30-June-2017).

Frailty

The data set has the frailty scores measured in multiple scales. We decided to use the frailty score according to the Edmonton Frail Scale (EFS).

The EFS assesses each participant over nine domains (see Table 4.1): cognition, general health status, functional independence, social support, medication usage, nutrition, mood, continence and functional performance. The frailty score in the EFS ranges from 0 to 17 and participants are typically classified into five categories: [0-5] Not frail, [6-7] Vulnerable, [8-9] Mild frailty, [10-11] Moderate frailty, and [12-17] Severe frailty.

The histograms in Figure 4.2 of the frailty scores of participants in each interview show that the distribution of frailty does not vary much across time, with median of frailty

The Edmonton Frail Scale

NAME : _____

d.o.b. : _____ **DATE :** _____

Frailty domain	Item	0 point	1 point	2 points
Cognition	Please imagine that this pre-drawn circle is a clock. I would like you to place the numbers in the correct positions then place the hands to indicate a time of 'ten after eleven'	No errors	Minor spacing errors	Other errors
General health status	In the past year, how many times have you been admitted to a hospital?	0	1–2	≥2
	In general, how would you describe your health?	'Excellent', 'Very good', 'Good'	'Fair'	'Poor'
Functional independence	With how many of the following activities do you require help? (meal preparation, shopping, transportation, telephone, housekeeping, laundry, managing money, taking medications)	0–1	2–4	5–8
Social support	When you need help, can you count on someone who is willing and able to meet your needs?	Always	Sometimes	Never
Medication use	Do you use five or more different prescription medications on a regular basis?	No	Yes	
	At times, do you forget to take your prescription medications?	No	Yes	
Nutrition	Have you recently lost weight such that your clothing has become looser?	No	Yes	
Mood	Do you often feel sad or depressed?	No	Yes	
Continence	Do you have a problem with losing control of urine when you don't want to?	No	Yes	
Functional performance	I would like you to sit in this chair with your back and arms resting. Then, when I say 'GO', please stand up and walk at a safe and comfortable pace to the mark on the floor (approximately 3 m away), return to the chair and sit down'	0–10 s	11–20 s	One of : >20 s , or patient unwilling , or requires assistance
Totals	Final score is the sum of column totals			

Scoring :
0 - 5 = Not Frail
6 - 7 = Vulnerable
8 - 9 = Mild Frailty
10-11 = Moderate Frailty
12-17 = Severe Frailty

TOTAL **/17**

Administered by : _____

Table 4.1: Edmonton Frail Scale questionnaire assesses 10 domains including cognitive impairment, balance and mobility.

score 3 in the four interviews and average frailty between 3.5 and 4.43 (values shown in histograms). This group of people seems to be at most moderately frail since very few participants have frailty scores greater than 11.

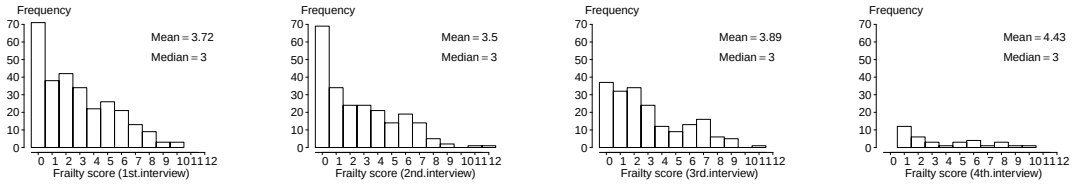


Figure 4.2: Histograms of the frailty scores (Edmonton Frailty Scale) of the participants enrolled in the study at each interview (1st to 4th).

Figure 4.3 (left) suggests that, on average, participants' frailty scores remain constant across time, but the older the participant at recruitment time, the more frail (right).

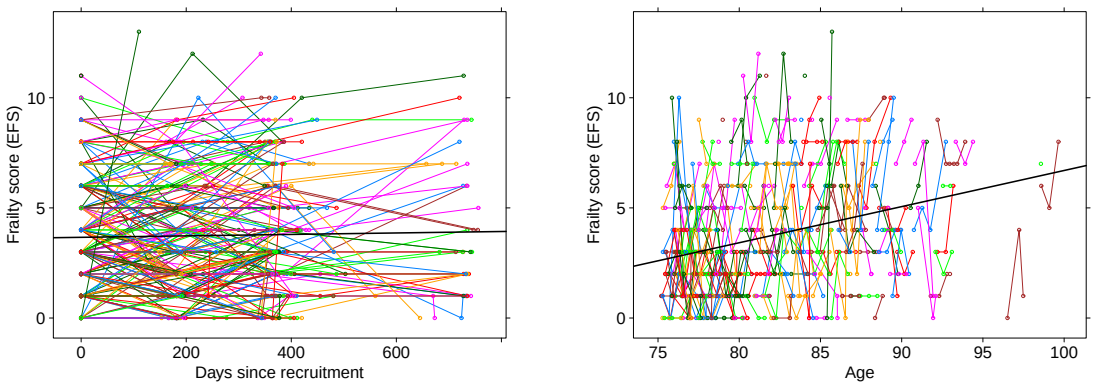


Figure 4.3: Left: Frailty scores of participants against days since recruitment; Right: Frailty scores of participants against age. The black superimposed lines (left and right) are regression lines.

There is an overall slightly increasing time trend of the frailty score. However, many participants are recruited in the study when they are older than 75, and we observe older people with higher frailty scores. As we can see in the left plot of Figure 4.3, the frailty scores at baseline are very different between participants, this suggests that a random

intercept for the linear mixed submodel might be appropriate. Additionally, since there is a lot of line crossing during the first 400 days since entering the study, a random slope of time might also be useful to model the individual time trends. However, a linear mixed model with a random slope for these data can be hard because of the relatively small sample size and number of repeated measures per participant: Among the 282 participants, only 35 have four repeated measures, 154 have three, 39 have two, and 54 have exactly one.

Falls

As we stated previously, participants of the CARE75+ study are interviewed every 6 months, and in each interview they are asked how many times they fell down in the past 12 months. The periodicity of the interviews is different from the time window of the occurrence of falls, so analyzing the data as it is would duplicate some falls. Moreover, there is no information in the data set about the exact time at which falls occurred and we were interested in analyzing falls as a time-to-event recurrent outcome. In order to prepare the falls data, we "adjusted" the number of falls reported by each participant to a 6 months window assuming that falls occurred at regular time intervals within the 12 months, and we set the end points of these time intervals as the time-to-fall. For instance, if a participant reported 12 falls in the last 12 months, the "adjusted" number of falls in 6 months is 6, one every month.

The bar plots in Figure 4.4 show that most participants reported zero falls. It is worth noting in the left plot of Figure 4.5 that some participants have persistently high number of falls in successive time periods, while others have significant number of falls. However, most of the participants reported zero falls since entering the study; 14.2% (40) reported to have fallen at least once, and only 10 participants have fallen a total of 5 times or more. The right plot of Figure 4.5 shows that people with the higher number of falls are between 80 and 90 years old.

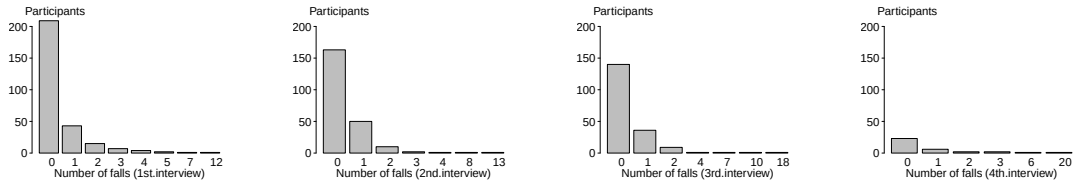


Figure 4.4: Bar plots for the number of falls reported by participants in each interview (1st to 4th). The height of each bar represents the number of participants, and the x-axis shows the number of falls reported.

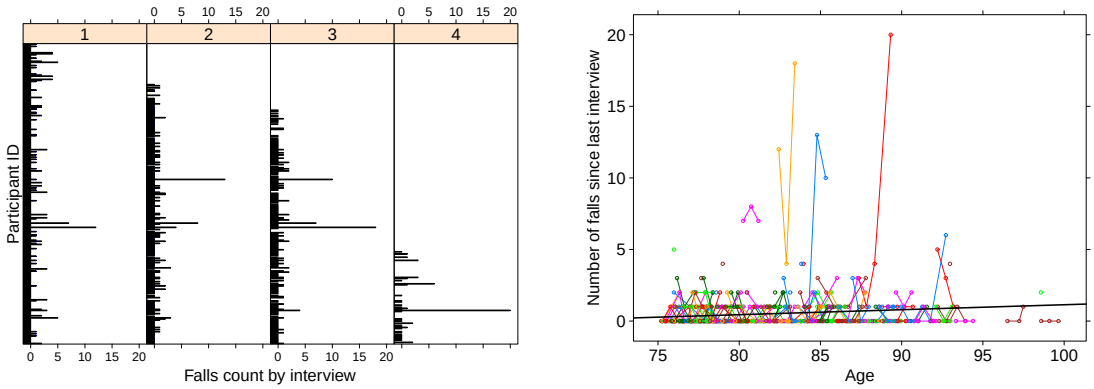


Figure 4.5: Left: Each horizontal bar represents the number of falls a participant reported in an interview. Right: Number of falls per participant at the age of enrollment in the study. The black superimposed line (right) is a regression line.

Mortality

The Kaplan–Meier estimate in Figure 4.6 shows that by the censoring time, mortality in the CARE75+ data set has been low. The survival probability beyond the administrative censoring time is 0.918, i.e. a large proportion of observations are censored, which might be due to the relatively short follow-up time and participants' age at the moment they enrolled in the study.

The top left plot in Figure 4.7 suggests that participants of the CARE75+ study who had died by the evaluation time (31-May-2017) were on average more frail than the censored participants (hazard ratio: 1.18). The top right plot in Figure 4.7 suggests an increasing risk of falls with increasing frailty (hazard ratio: 1.24).

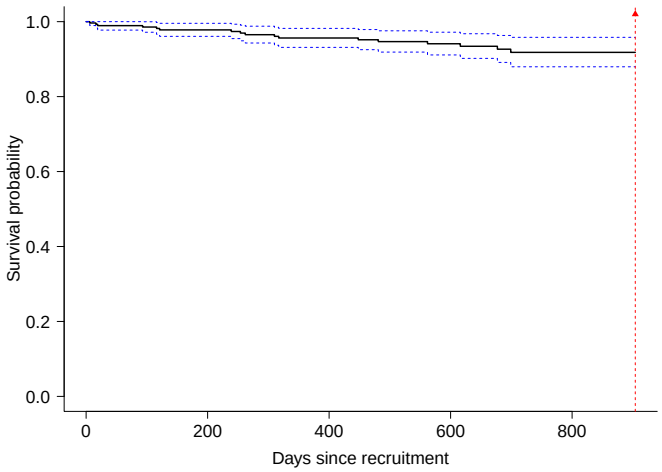


Figure 4.6: Kaplan–Meier estimate of death (—) and 95% confidence band (- -).

As mentioned before, very few participants have more than four falls, most of which are censored by 31-May-2017. All these participants are represented by the outliers in the bottom left plot of Figure 4.7. Although the plot might suggest decreasing hazard of death with increasing number of falls, the hazard ratio is ≈ 1.

4.2.2 Covariates

In addition to data about frailty, falls and mortality, the CARE75+ study records basic data about the participants, lifestyle habits and health conditions, some of which vary with time. We used some of these variables as covariates to construct a joint model for frailty, falls and mortality. Most of these covariates are categorical variables, and we collapsed the categories of each one of them into two groups. We briefly describe these covariates and how we created their dichotomous versions, showing the counts in each category in absolute values and percentages. Table 4.2 contains a summary of these variables and their associated hazard ratio of mortality.

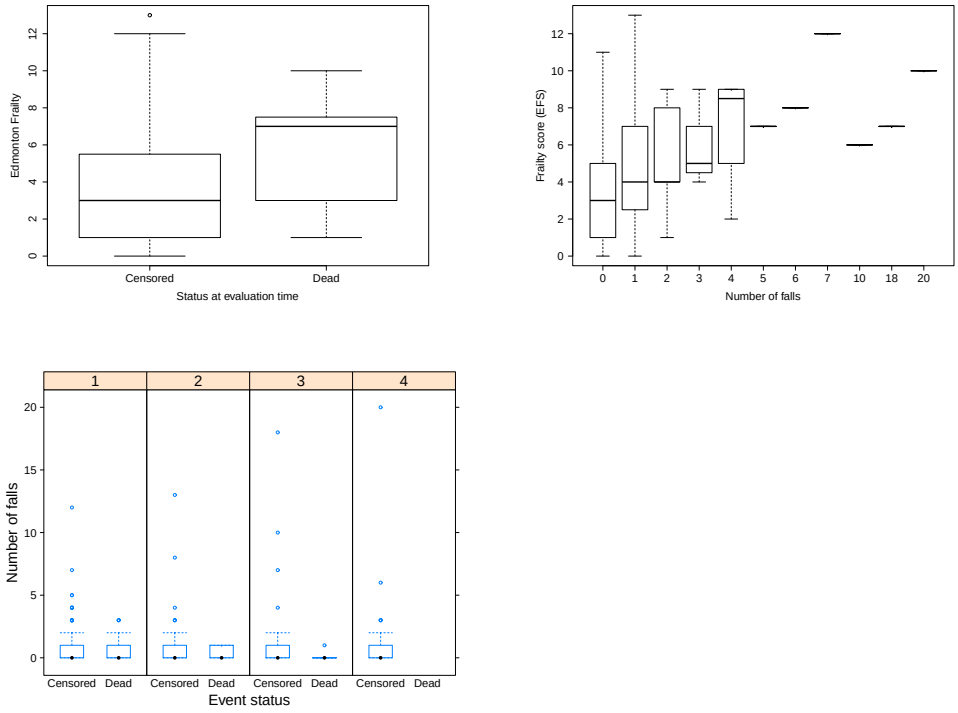


Figure 4.7: Top left: Frailty scores by terminal event status at 31/May/2017; top right: Frailty scores by falls count and bottom left: Falls count by terminal event status at 31/May/2017.

Time-fixed

Sex

The data set we analyzed with cut date 31-May-2017 contains data of 282 participants: 43.6% men and 56.8% women. In our analyses, we labeled this variable as `sex` with women being the baseline category:

$$\texttt{sex} = \begin{cases} 1 & \text{Male} \quad (123; 43.6\%) \\ 0 & \text{Female} \quad (159; 56.8\%). \end{cases}$$

Ethnicity

People in the study were asked about their ethnic origin according the 14 groups: White (85.5%), mixed White/Black Caribbean (0.4%), mixed White/Asian (0%), mixed White/Black African (0%), other mixed (0%), Black African (0%), Asian Indian (1.1%), Asian Blangladeshi (0.7%), Asian Pakistani (12.1%), other Asian (0%), Black Caribbean (0.4%), Other Black (0%), Chinese (0%), other (0%). Due to the large proportion of participant who identified themselves as White, we collapsed the other participants in a unique group as follows:

$$\texttt{ethnicity} = \begin{cases} 1 & \text{White} \quad (241; 85.5\%) \\ 0 & \text{Other} \quad (41; 14.5\%). \end{cases}$$

The data set is unbalanced with respect to ethnicity, being White most of the participants. However, the number of deaths are roughly proportionally spread across the two collapsed ethnicity groups, such that mortality is 0.058 among White and 0.07 among Other (hazard ratio in Table 4.2 is not significant at the 0.05 significance level). The imbalance of ethnicity in the data set maybe due to the fact that this is an ongoing study and it is possible that a considerable fraction of the target population has not being reached yet. With so few deaths in the data set and ethnicity being largely unbalanced, we suspect convergence difficulties in the optimization algorithm while fitting a joint model for frailty, falls and mortality.

Marital status

In recent studies, for instance Trevisan *et al.* (2016) and Trevisan *et al.* (2020), it has been suggested that marital status in the elderly can influence the development of frailty, and that this influence can vary between women and men (single or widowed men carry a higher risk of developing frailty, whereas widows have a lower odds of becoming frail than married women). Manzoli *et al.* (2007) conducted a meta-analysis to investigate the relationship between marital status and mortality in the elderly, concluding that persons marriage or support from the partner is associated with a reduction in all-cause mortality risk.

The martial status of participants of the CARE75+ is recorded as Single (never married) (2.5%), Married (39.7%), Remarried (3.5%), Separated but still legally married (0.4%), Divorced (8.2%), Widowed (45.7%). We created the dichotomous variable `marital` by collapsing marital status in two groups with the intention of keeping in the same group those who are likely to have the support of a spouse:

$$\texttt{marital} = \begin{cases} 1 & \text{Married or Remarried} \quad (122; 43.3\%) \\ 0 & \text{Other} \quad (160; 56.7\%). \end{cases}$$

Education

When participants are enrolled in the CARE75+ study they are asked "What was the highest educational qualification you attained?". In the UK education sector, there are several qualification types offered by the UK awarding bodies, which can be a professional body, school, college or university. Qualifications can be academic, vocational or skills-related, and are grouped together into different levels. Participants' information regarding qualifications is recorded as: No qualifications (62.4%), GCSE (General Certificate of Secondary Education) (14.9%), HNC/HND (Higher National Certificate / Higher National Diploma) (4.3%), Diploma (7.4%), AS and A level (3.2%), Bachelor's degree (5.3%), Postgraduate (2.5%),

We considered the data about qualifications as synonym of education level and collapsed the categories of this variable into two groups to distinguish those participants with no qualifications from those with any qualification and identified the latter as a

higher education group.

$$\texttt{education} = \begin{cases} 1 & \text{if} \begin{cases} \text{GCSE} \\ \text{HNC/HND} \\ \text{Diploma} \\ \text{AS and A level} \\ \text{Bachelor's degree} \\ \text{Postgraduate} \end{cases} (106; 37.6\%) \\ 0 & \text{No qualifications} \quad (176; \%62.4). \end{cases}$$

Lifestyle (alcohol, smoker)

The lifestyle habits recorded in the analyzed data set refer to alcohol consumption and smoking. Participants are asked if they ever drink alcohol, if they have ever smoked and if they smoke nowadays. We created the dichotomous analysis variables `alcohol` and `smoker` as follows:

$$\texttt{alcohol} = \begin{cases} 1 & \text{Participant drinks alcohol} \quad (195; 69.1\%) \\ 0 & \text{Otherwise} \quad (87; 30.9\%) \end{cases}$$

$$\texttt{smoker} = \begin{cases} 1 & \text{Participant is or was a smoker} \quad (136; 48.2\%) \\ 0 & \text{Participant has never been a smoker} \quad (146; 51.8\%) \end{cases}$$

Time-varying

Body Mass Index

The Body Mass Index (BMI) is an indicator to screen for weight categories that may lead to health problems. We computed the BMI with the records of weight and height:

$$\texttt{bmi} = \frac{\text{Weight (Kg.)}}{\text{Height}^2 \text{ (Meters)}^2}$$

Naturally, weight and height change in time, so `bmi` is a time-varying covariate in our analyses, even though dramatic changes of BMI in the elderly are unlikely. Figure 4.8 shows that the distribution of BMI is similar across the first three interviews differing

mainly by sample size (only a small number of participants have had a fourth interview). The mean and median BMI is about 27 (values shown in histograms) in each interview.

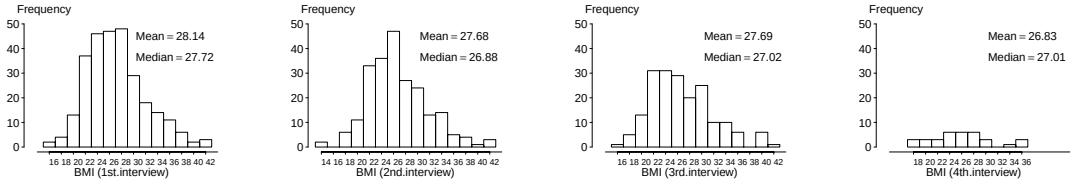


Figure 4.8: Body mass index by participants in each interview.

Comorbidities

Comorbidity is the presence of an additional conditions co-occurring with a primary condition, in this case frailty. Behavioral and mental conditions are also considered comorbidities. Figure 4.9 shows the number of comorbidities participants report in each interview. The two participants that reported the largest number of comorbidities in the first interview (13 and 19) died by the censoring date.

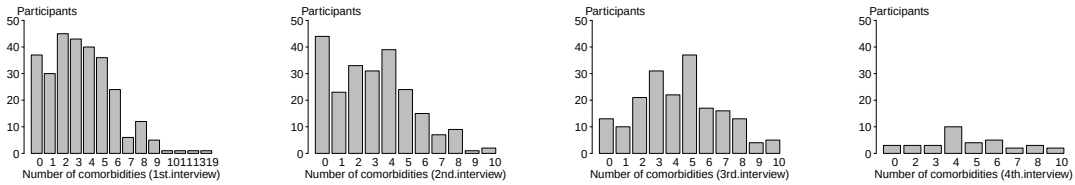


Figure 4.9: Number of comorbidities reported by participants in each interview.

Visits to a general practitioner (GP)

Figure 4.10 shows that most participants have between 1 and 3 visits to a general practitioner between interviews. Only 5 participants have more than 4 visits and they reported between 5 and 9 comorbidities.

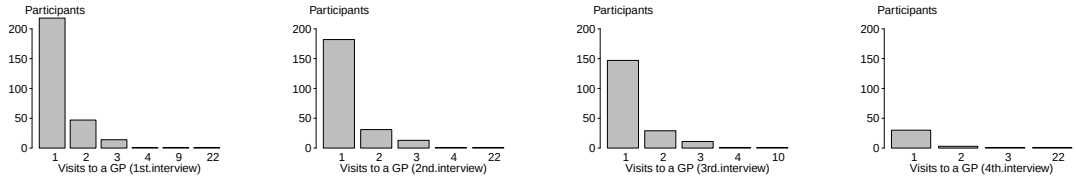


Figure 4.10: Frequency of visits to a general practitioner reported by participants in each interview.

Table 4.2 shows that each category of the binary versions of the time-fixed covariates has at least three events. Some of these covariates are roughly balanced in terms of the sample size, but this is not the case for `ethnicity` and `alcohol` where the sample size in the baseline categories is smaller. According to the interval estimates, the hazard ratio of mortality is different from 1 only for `comorbidities`, for the other covariates the interval estimates contain the value of 1.

We conclude this section by pointing out some features of the data set. Our interest is to fit a joint model for frailty, falls and mortality with the CARE75+ data set. This kind of model often has much more parameters to estimate than separate marginal models for each outcome, and optimizing the likelihood, which involves nonlinear functions of time and possibly several random effects, is more difficult. Hence usually larger sample sizes are required for joint modelling. In this analysis, the CARE75+ data set contains records of 282 participants, but only 189 have either three or four repeated measurements of frailty, which might represent a small sample size to fit a joint model for a longitudinal response and recurrent and terminal events. Furthermore, the data set does not contain data about the time-to-fall so we had to impute it based on the available data about the number of falls in a 12 months window, under the debatable assumption of falls occurring at regular time intervals.

Mortality is the terminal event. Frailty and recurrent falls are endogenous time-varying covariates in a model for mortality. The data set contains few deaths (17), and 252 participants had either 0 falls (201) or only 1 (51), this might not be sufficiently large for a joint modelling these relationships. Finally, ethnicity is a variable largely imbalanced in the data set. Although in covariates imbalance does not necessarily represent a problem on its own, in a marginal survival analysis model with the small number

Covariate	Total	Censored	Death	HR	LCL	UCL
Time-fixed						
Sex						
Male	123	117	6	1.435	0.531	3.881
Female	159	148	11	–	–	–
Ethnicity						
White	241	227	14	0.870	0.250	3.029
Other	41	38	3	–	–	–
Marital status						
Married/Remarried	160	151	9	1.156	0.446	2.998
Other	122	101	5	–	–	–
Education						
High level	106	101	5	0.663	0.234	1.884
Otherwise	176	164	12	–	–	–
Drinks alcohol						
Yes	195	185	10	0.680	0.258	1.788
No	87	80	7	–	–	–
Smoker						
Smoker	136	128	8	0.895	0.345	2.322
Non-smoker	146	137	9	–	–	–
Time-varying						
BMI	282	265	17	0.949	0.856	1.052
# comorbidities	282	265	17	1.178	1.009	1.375
# visits to a GP	282	265	17	0.930	0.527	1.643

Table 4.2: Counts of participants by covariates and status (censored or death), hazard ratio (HR) of mortality and 95% confidence interval (LCL, UCL).

of terminal events it might be more difficult for the likelihood optimization algorithm to converge in the joint modelling context. The fact that the CARE75+ is an ongoing population study the imbalance of the variable ethnicity in the data set might be due to the short follow up period, being a significant fraction of the target population still missing to be included in the study.

4.3 Joint modelling frailty, recurrent falls and mortality for the CARE75+ data

Our primary interest is to model the relationship between frailty and mortality with the CARE75+ data set. Since frailty would be an endogenous time-varying covariate in a time-to-event model for mortality we would need to jointly model their relationship. We suspect there are other possible relationships between frailty and falls and between falls and mortality, so we will explore all these relationships in a joint model.

As an initial step of our analysis, we fitted Cox models for mortality with frailty and falls as covariates to have an approximate idea of their possible effect on mortality. The time origin for all our survival and joint modelling analyses is "the moment participants enter the study", so the time scale is "time since recruitment or enrollment". Table 4.3 shows the log(hazard ratio) of mortality associated to frailty and falls, and the corresponding log(hazard ratio) after adding the other covariates of the data set in an extended Cox model, assuming both are external time-varying covariates as described in Section 2.2.1. Equations (4.1)–(4.6) describe these models, where γ_{frail} and γ_{falls} denote the log(hazard ratio) of mortality associated to the raw frailty score in the EFS scale $(y_i(t))$ and the number of falls $(N_i(t))$, respectively, at time t. The term "covariates" in these equations refers to the vector of variables containing number of comorbidities, number of visits to a GP, BMI, gender, ethnicity, marital status, education level and alcohol and smoking habits. At the moment we are interested in the hazard ratio of mortality associated to frailty and falls both unadjusted and confounding-adjusted, assuming that they are exogenous covariates in a model for mortality.

$$\text{PH}_y \quad h_i(t) = h_0(t)\exp\{\gamma_{\text{frail}} y_i(t)\} \tag{4.1}$$
$$\text{PH}_{y+\text{covs}} \quad h_i(t) = h_0(t)\exp\{\gamma_{\text{frail}} y_i(t) + \boldsymbol{\gamma}^\top \text{covariates}\} \tag{4.2}$$
$$\text{PH}_{yN} \quad h_i(t) = h_0(t)\exp\{\gamma_{\text{frail}} y_i(t) + \gamma_{\text{falls}} N_i(t)\} \tag{4.3}$$
$$\text{PH}_{yN+\text{covs}} \quad h_i(t) = h_0(t)\exp\{\gamma_{\text{frail}} y_i(t) + \gamma_{\text{falls}} N_i(t) + \boldsymbol{\gamma}^\top \text{covariates}\} \tag{4.4}$$
$$\text{PH}_N \quad h_i(t) = h_0(t)\exp\{\gamma_{\text{falls}} N_i(t)\} \tag{4.5}$$
$$\text{PH}_{N+\text{covs}} \quad h_i(t) = h_0(t)\exp\{\gamma_{\text{falls}} N_i(t) + \boldsymbol{\gamma}^\top \text{covariates}\} \tag{4.6}$$

	$\widehat{\gamma}_{\text{frail}}$		$\widehat{\gamma}_{\text{falls}}$	
Model	Estimate	p-value	Estimate	p-value
PH_y	0.169	0.039	–	–
$\text{PH}_{y+\text{covs}}$	0.152	0.125	–	–
PH_{yN}	0.180	0.034	−0.100	0.644
$\text{PH}_{yN+\text{covs}}$	0.170	0.100	−0.136	0.573
PH_N	–	–	−0.003	0.985
$\text{PH}_{N+\text{covs}}$	–	–	−0.056	0.792

Table 4.3: Hazard ratios of mortality associated to frailty and falls. "–" means that the model of the row does not have the covariate of the column.

The estimates in Table 4.3 were obtained for an unspecified baseline hazard using the `coxph()` function of the `survival` library in R. The log(hazard ratio) associated to frailty is 0.169 and significant at the level of 0.05, and its value and significance remain almost unchanged after adjusting for falls. However, when adding the other covariates to the model even though the log(hazard ratio) of frailty does not change much, its effect becomes not significant. The value of the log(hazard ratio) of falls is far from being significant regardless of whether frailty and other covariates are added to the model.

Our aim for this chapter is to construct a model that best describes the relationship between between frailty, falls and mortality in the CARE75+ data set, relaxing the assumption about frailty and falls being external time-varying covariates in a time-to-event model for mortality.

4.3.1 Assumptions and model formulation

This section explains a general formulation of a joint model for three outcomes: longitudinal, recurrent events and terminal time-to-event and show the results of fitting a joint model of this type to the CARE75+ data.

Following the general framework of a three-outcome joint model for longitudinal, and recurrent and terminal events data described in Section 3.3, we specify a joint model for frailty, falls and mortality for the CARE75+ data set, call it M_3^{CARE}. The link functions of Model M_3^{CARE} are $g_{(\text{Rec})}(b_{i0}, t) = g_{(\text{Ter})}(b_{i0}, t) = b_{i0}$ and it is described by Equations (4.7a)–(4.7c) and its corresponding likelihood function has the same general form as Equation 3.25.

$$M_3^{\text{CARE}} : \begin{cases} \text{Longitudinal} & y_i(t|b_{i0}) = (\beta_0 + b_{i0}) + \mathbf{x}_i^\top(t)\boldsymbol{\beta} + \varepsilon_i(t) \quad (4.7a) \\ \text{Recurrent} & r_i(t|\mathbf{v}_i) = u_i r_0(t) \exp\{\mathbf{w}_i^\top(t)\boldsymbol{\gamma}_R + \eta_R b_{i0}\} \quad (4.7b) \\ \text{Terminal} & h_i(t|\mathbf{v}_i) = u_i^\alpha h_0(t) \exp\{\mathbf{w}_i^\top(t)\boldsymbol{\gamma}_T + \eta_T b_{i0}\} \quad (4.7c) \end{cases}$$

where,

$\mathbf{w}_i(t)$: sex, ethnicity, highest education, marital status, smoker, alcohol, number of comorbidities, frequency of visits to a GP

$\mathbf{x}_i^\top(t)$: $[t, \mathbf{w}_i(t)^\top)]$

$y_i(t)$: Frailty score at time t

$r_i(t)$: Hazard rate of recurrent falls at time t

$h_i(t)$: Hazard rate of death at time t

$\varepsilon_i(t)$: measurement error of frailty score

b_{i0} : random intercept

u_i : random effect common to the two hazards

We assume a random intercept linear mixed model for the longitudinal outcome. We also assume that the longitudinal outcome is linked to the recurrent events and the terminal event only through the random intercept, $b_{i0} \sim \mathcal{N}(0, \sigma_b^2)$, and the recurrent events and the terminal event processes are linked through the random effect, $u_i \sim$

$\text{Gamma}(\phi^{-1}, \phi^{-1})$, with $\mathbf{v}_i^\top = (b_{0i}, u_i)$. The baseline hazards of the recurrent and terminal events are assumed parametric from a Weibull distribution with shape κ_R and κ_T and rate ρ_R and ρ_T, respectively.

4.3.2 Marginal models

Joint models are difficult to fit and require much longer computing times than the standard marginal models of each outcome. A major difficulty we experienced is the lack of convergence with some settings: the mean structure, the link function, whether or not a random slope is included, the type of recurrent event submodel, and inclusion of some covariates. For instance, when a random slope is included in the joint model, the optimization algorithm of the fitting function does not converge. This makes model fitting challenging. In particular, the variable selection process of the joint model of frailty, falls and mortality is a difficult task. In order to alleviate this process we fitted first marginal models for each outcome: frailty, falls and mortality, described by Equations (4.8)–(4.10). The differences between the marginal models we used and joint model M_3^{CARE} are: (1) all time-varying covariates are assumed exogenous (including frailty and falls which can be covariates of the other two marginal models) and (2) the marginal models do not have association parameters (η_R, η_T, α). Note that these marginal models are not good enough to represent the relationship between frailty, falls and mortality because they do not take into account the potential correlation among these three variables. Like model M_3^{CARE} the marginal models for frailty and falls have a random effect to account for the repeated measures of the frailty score and recurrent falls within subjects, but not the marginal model for mortality. So we carried out a variable selection process in each marginal model and used the chosen covariates as the initial covariate set to fit the joint model. Finally, we removed from the joint model all non-significant covariates.

$$\text{Longitudinal} \quad y_i(t|b_{i0}) = (\beta_0 + b_{i0}) + \mathbf{x}_i^\top(t)\boldsymbol{\beta} + \varepsilon_i(t) \tag{4.8}$$

$$\text{Recurrent} \quad r_i(t|u_{Ri}) = u_{Ri} r_0(t) \exp\{\mathbf{w}_i^\top(t)\boldsymbol{\gamma}_R + \gamma_{R.\text{frail}} y_i(t)\} \tag{4.9}$$

$$\text{Terminal} \quad h_i(t) = h_0(t) \exp\{\mathbf{w}_i^\top(t)\boldsymbol{\gamma}_T + \gamma_{T\text{falls}} N_i(t) + \gamma_{T.\text{frail}} y_i(t)\} \tag{4.10}$$

where $y_i(t)$ is the observed frailty score of subject i at time t, and $N_i(t)$ the number of falls reported by time t.

To avoid confusion, we refer to each regression subequation (6.1a)–(6.1c) as a **submodel** of joint model M_3^{CARE}, and to each Equation (4.8)–(4.10) as the **marginal model** of frailty, falls and mortality, respectively.

The variable selection mechanism in the marginal models was a combination of stepwise and backwards elimination process. We started by fitting the saturated model. Then, among the covariates with non-significant regression coefficient (significance level 0.05), we removed from the model the one with the largest p-value and re-fitted the model. During this backward elimination process, when removing a certain covariate affected the significance of the others we kept in the model the one that with the greater contribution to the log-likelihood function. We stopped when all the covariates in each marginal model were significant and used these covariates as the starting point for the variable process in the joint model M_3^{CARE}. An exception to this last criterion is `ethnicity` in the submodel of mortality, where this covariate remained in the model for convergence of the optimization algorithm. The first three columns of Table 4.4 show the estimates, their standard error and p-values corresponding to the marginal models containing only the significant covariates.

The parameters of the linear mixed model are estimated by maximizing the restricted log-likelihood function discussed in Section 2.1.2 and implemented in R in the `lme()` function of the `nlme` package. In this fitted model, the significant covariates are `ethnicity`, `married`, `education`, `alcohol`, `smoker`, `# comorbidities` and `falls`. The time slope estimate, $\widehat{\beta} = 0.223$, was significant suggesting an overall decreasing time trend of frailty of the people in the study, which although consistent with the intuitive idea of people becoming more frail as getting older, in this particular group of people the increase of frailty in time seems small (as seen in right side plot of Figure 4.3). This time effect becomes insignificant when the joint model is fitted, as discussed in Section 4.3.3. The estimates of the variance of the random intercept and the variance of the error are $\widehat{\sigma}_b^2 = 4.247$ and $\widehat{\sigma}_\varepsilon^2 = 1.105^2 = 1.221$, respectively.

We analyzed falls with the Andersen–Gill (AG) model (see Section 2.2.2) in the *total time* scale, assuming a parametric baseline hazard from the $\text{Weibull}(\kappa_R, \rho_R)$ distribu-

	Marginal model			Joint model ($\widehat{M}_3^{\text{CARE}}$)		
Parameter	Estimate	Std.Err	p-value	Estimate	Std.Err	p-value
Fixed effects						
Frailty						
$\widehat{\beta}_0$	6.229	0.404	< 0.001	6.166	0.343	< 0.001
$\widehat{\beta}_t$	0.223	0.078	0.004	-0.029	0.111	0.793
$\widehat{\beta}_{\text{eth}}$	-2.093	0.415	< 0.001	-2.026	0.354	< 0.001
$\widehat{\beta}_{\text{mar}}$	-0.814	0.250	0.001	-0.729	0.211	< 0.001
$\widehat{\beta}_{\text{edu}}$	-0.816	0.275	0.003	-0.856	0.223	< 0.001
$\widehat{\beta}_{\text{alc}}$	-1.094	0.235	0.001	-1.247	0.233	< 0.001
$\widehat{\beta}_{\text{smo}}$	-0.610	0.220	0.006	$+$	$+$	$+$
$\widehat{\beta}_{\text{com}}$	0.222	0.031	< 0.001	0.234	0.031	< 0.001
$\widehat{\beta}_{\text{falls}}$	0.330	0.137	0.016	$+$	$+$	$+$
Falls						
$\widehat{\gamma}_{R.\text{eth}}$	1.429	0.374	< 0.001	0.866	0.367	0.029
$\widehat{\gamma}_{R.\text{frail}}$	0.183	0.036	< 0.001	$+$	$+$	$+$
Mortality						
$\widehat{\gamma}_{T.\text{eth}}$	$+$	$+$	$+$	-0.077	0.703	0.912
$\widehat{\gamma}_{T.\text{frail}}$	0.171	0.082	0.036	$+$	$+$	$+$
Association						
$\widehat{\eta}_R$	$+$	$+$	$+$	0.439	0.081	< 0.001
$\widehat{\eta}_T$	$+$	$+$	$+$	0.430	0.206	0.037
$\widehat{\alpha}$	$+$	$+$	$+$	-0.962	0.703	0.171
Variance component						
Frailty						
$\widehat{\sigma}_{\varepsilon}$	1.105	$-$	$-$	2.152	$-$	$-$
$\widehat{\sigma}_b^2$	4.247	$-$	$-$	2.577	$-$	$-$
Falls						
$\widehat{\phi}_R$	1.864	0.379	< 0.001	$+$	$+$	$+$
Falls–Mortality						
$\widehat{\phi}$	$+$	$+$	$+$	1.143	0.849	0.0891
Baseline hazard						
Falls						
$\widehat{\kappa}_R$	0.978	0.066	$-$	1.023	$-$	$-$
$\widehat{\rho}_R$	0.056	0.003	$-$	0.322	$-$	$-$
Mortality						
$\widehat{\kappa}_T$	0.792	0.078	$-$	1.040	$-$	$-$
$\widehat{\rho}_T$	0.004	0.000	$-$	0.170	$-$	$-$

Table 4.4: Marginal and joint model fit. "$-$" not directly available from software output. "$+$" not a parameter of model on the column. "$\bullet$" correspond to ϕ in M_3^{CARE}.

tion and a random effect, $u_{Ri} \sim \text{Gamma}(\phi_R^{-1}, \phi_R^{-1})$, trying to keep consistency with the specifications of the joint model M_3^{CARE}. The parameters of the AG model were estimated by maximizing the likelihood function of Equation (2.40) by the Marquardt algorithm (Marquardt, 1963; Rondeau *et al.*, 2003), which combines the Newton–Raphson and steepest descent algorithms and implemented in the `frailtyPenal()` function of the R library `frailtypack`, as discussed in Section 2.2.2. The relationship between frailty and falls is estimated by $\widehat{\beta}_{\text{falls}} = 0.330$ in the marginal model of frailty (see Table 4.4), suggesting a direct association between falls and frailty. This is consistent with the significant coefficient, $\widehat{\gamma}_{R.\text{frailty}}$, in the marginal model of falls. Note that `ethnicity` is a significant covariate in both frailty and falls marginal models.

In the proportional hazards marginal model of mortality we also assumed a baseline hazard governed by the $\text{Weibull}(\kappa_T, \rho_T)$ distribution. This marginal model does not include a random effect[1]. To fit this model we used also the R function `frailtyPenal()` mentioned above, which can also accommodate a random effects if necessary. An alternative R model fitting tool for this specification of the proportional hazards model is the function `parfm()` of the library `parfm`, for which several distributions are available for parametric estimation of the baseline hazard and accommodates random effects in for the proportional hazards model. Not surprisingly, the rate parameter of the baseline hazard, $\widehat{\rho}_T = 0.004$, is small. Note that, according to the marginal model of mortality, `falls` has no significant relationship with mortality (this covariate is not in the model of mortality because it was eliminated during the variable selection process).

As summary, the main conclusions regarding the relationship between frailty, falls and mortality derived from the marginal models fit are: direct associations between (1) frailty and falls and (2) frailty and mortality, and (3) no association between falls and mortality. Ethnicity is strongly associated with frailty and falls, but not with mortality. Apparently, there is a positive and significant time trend of the average frailty.

[1] We also ran the model with a random effect, $u_{Ti} \sim \text{Gamma}(\phi_T^{-1}, \phi_T^{-1})$. The fixed effects estimates were the same, with an estimate of random effect's parameter $\widehat{\phi}_{T_i} = 0.001$, standard error $\widehat{\text{se}}(\widehat{\phi}_{Ti}) = 0.0001$, and p-value = 0.5

4.3.3 Joint model fit

We started the variable selection process for the joint model for frailty, falls and mortality, M_3^{CARE}, by including in each submodel the covariates of the marginal models listed in Table 4.4. Just as we did with the marginal models, we removed sequentially from the model the non-significant covariates (significance level 0.05) from the joint model, one at a time. According to the formulation of the joint model M_3^{CARE}, the association between frailty and falls is modelled by the random intercept, b_{i0}, and the strength of this association is quantified by η_R. In the variable selection process of the joint model, `smoker` and `falls` were removed from the linear mixed submodel because they were not significant (p-value > 0.05), the rest of the covariates remained in this submodel. In the submodel of falls, `ethnicity` was still significant so it was retained in the final model. An exception to the exclusion criterion was the `ethnicity` in the submodel of mortality, which we kept even though it was not significant for the joint model to converge as the `trivPenal()` does not converge when the terminal event submodel has no covariates in it.

In Section 3.3 we discussed parameter estimation of joint models similar to M_3^{CARE}. To estimate the parameters of model M_3^{CARE} we followed the approach of Król *et al.* (2016) and Rondeau *et al.* (2012) using the Marquardt algorithm to optimize the likelihood function of this model, using the Gauss-Hermite technique for the numerical integration with respect to the random effects and assuming that the baseline hazards of falls and mortality are governed by the Weibull(κ, ρ) distribution. This estimation procedure is implemented in the `trivPenal()` function of the R library `frailtypack`, calling other functions programmed in `FORTRAN`. Recall that in M_3^{CARE} there is a shared random effect between falls and mortality, u_i with ϕ being the parameter of the distribution of u_i that determines its variance. We denote by $\widehat{M}_3^{\text{CARE}}$ the model that results from the variable selection process in the joint model, i.e. only the significant covariates are kept in this model. The last three columns of Table 4.4 show the parameter estimates of model $\widehat{M}_3^{\text{CARE}}$, their standard error and p-values.

Note that it is not possible to make a direct comparison of $\widehat{\gamma}_{T.\text{frail}}$ and $\widehat{\eta}_T$ because they are related to different variables with different meanings. On the one hand, $\widehat{\gamma}_{R.\text{frail}}$ is related to $y_i(t)$, i.e. the observed value of the frailty score at t which is assumed

to contain measurement error. As discussed in Section 3.1, when measurement error is ignored and time-varying covariates are endogenous in time-to-event models, the regression coefficients are biased. On the other hand, $\widehat{\eta}_T$ is related to the deviation of the true and unobserved frailty with respect to the population average.

By joint modelling frailty, falls and mortality we noticed that `ethnicity`, `married`, `education`, `alcohol` and `number of comorbidities` are significant covariates of frailty. The values of these estimates along with the intercept do not seem far off from the estimates of the marginal model estimates. However, the time trend of the average frailty of the CARE75+ data set, $\widehat{\beta} = -0.029$, becomes not significant in the joint model. We suspect this behaviour is due to the relatively small number of repeated measures per subject. This is explored further with simulation studies in Chapter 6.

The variable `ethnicity` is a significant fixed effect covariate of falls but not of mortality, which is consistent with the marginal models of these two outcomes.

The estimate of the variance of the random intercept of the submodel of frailty, $\widehat{\sigma}_b^2 = 2.577$ is smaller compared to this estimate in the marginal model of frailty.

The estimate of the parameter that accounts for the strength of the association between falls and mortality, $\widehat{\alpha} = -0.962$, it is not significant, so we cannot rule out the possibility of no association between these two outcomes. This is in line with the conclusion of no association between falls and mortality in the marginal models. Although at first glance the lack of association between falls an mortality seems counterintuitive, we must recall the short follow up period (2 years), the few terminal events in the data set (17 deaths) and the relatively small number of falls per subject (83% of participants have a total of zero or one fall). Hence getting a correct estimate of the random effect might be difficult in this data set, for which the mean structure needs to be correctly estimated. Additionally, it is possible that the falls reported by the CARE75+ participants are not severe enough to pose important deterioration in their general health condition or mobility, and they implement precautionary measures, for instance using an assistive device to prevent further falls or to make them even less severe although more frequent. The analyzed data set does not contain information regarding the severity of falls.

The estimate of the association parameter, ($\widehat{\eta}_R = 0.439$), supports the direct association between frailty and falls, suggested by the descriptive analysis of Section 4.2.1 and the marginal models. Recall the interpretation of the coefficients of a proportional hazards model discussed in Section 2.2.1 and the idea of the random intercept b_{i0} as a latent variable that models the correlation of repeated measures of frailty within the same individual, representing the deviation of an individual's frailty at baseline with respect the population average. So $\exp(\widehat{\eta}_R \times \widehat{b}_{i0})$ estimates the relative risk of falls at time of an individual whose frailty differs by $\widehat{b}_{i0}$ units with respect to the average frailty.

We derive a similar conclusion about the association parameter $\widehat{\eta}_T = 0.430$, indicating a higher relative risk of mortality the more frail the individuals are. This agrees with the analysis of Section 4.2.1 and the marginal model estimate of the association between frailty and mortality.

The shape parameters estimates of the baseline hazards of both the recurrent and terminal events are close to 1, meaning that they are likely governed by the Exponential distribution, provided the family we assumed (Weibull) is the correct one. The rate parameter estimates of both hazards, $\widehat{\rho}_R = 0.322$ and $\widehat{\rho}_T = 0.170$, are greater compared to their estimates in the marginal models.

In the next section we verify the assumptions of joint model $\widehat{M}_3^{\text{CARE}}$.

4.4 Model diagnostics

In statistical modelling, the goal of model diagnostics is to verify that assumptions of the model are reasonable in the fitted model. The standard tools to assess a model's assumptions are residual plots. In the standard linear mixed effects model, two types of residuals are often used: *conditional* (subject-specific) residuals and *marginal* (population averaged) residuals (Molenberghs & Verbeke, 2000). The conditional residuals aim to validate the assumptions of the hierarchical version of the model described in Equation (2.3), and are defined as

$$\widehat{\varepsilon}_i^c(t) = y_i(t) - \mathbf{x}_i^\top(t)\widehat{\boldsymbol{\beta}} - \mathbf{z}_i^\top(t)\widehat{\mathbf{b}}_i, \tag{4.11}$$

with corresponding standardized version for the conditional independence model (see Equation (2.5))

$$\widehat{\varepsilon}_i^{*c}(t) = y_i(t) - \left(\mathbf{x}_i^\top(t)\widehat{\boldsymbol{\beta}} - \mathbf{z}_i^\top(t)\widehat{\mathbf{b}}_i\right)/\widehat{\sigma}_\varepsilon, \tag{4.12}$$

where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}_\varepsilon$ are the maximum likelihood estimates and $\widehat{\mathbf{b}}_i$ the empirical Bayes estimates of the random effects. These residuals predict the conditional errors, $\varepsilon_i(t)$, and can be used for checking the homoscedasticity and normality assumptions.

On the other hand, the marginal residuals focus on the marginal model for the longitudinal outcome implied by the hierarchical representation, i.e. $\mathbf{y}_i = X_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, V_i)$ and $V_i = Z_i B Z_i^\top + \sigma_\varepsilon^2 I_{n_i}$. The marginal residuals are defined as

$$\widehat{\boldsymbol{\epsilon}}_i^m = \mathbf{y}_i - X_i\widehat{\boldsymbol{\beta}}, \tag{4.13}$$

with corresponding standardized version

$$\widehat{\boldsymbol{\epsilon}}_i^{*m} = \widehat{V}_i^{-1/2}\left(\mathbf{y}_i - X_i\widehat{\boldsymbol{\beta}}\right), \tag{4.14}$$

where $\widehat{V}_i = Z_i\widehat{B}Z_i^\top + \widehat{\sigma}_\varepsilon^2 I_{n_i}$ denotes the estimated marginal covariance matrix of $\mathbf{y}_i$. The marginal residuals predict the marginal errors $\mathbf{y}_i - X_i\boldsymbol{\beta} = Z_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i$, and can be used to investigate misspecification of the mean structure $X_i\beta$ as well as to validate the assumptions for the within-subjects covariance structure, V_i. We used the conditional residuals to check the assumptions of the longitudinal part of the model.

A standard type of residuals for the relative risk submodel of the joint model is the martingale residuals. These are based on the counting process notation of the time-to-event data, and in particular on the subject-specific counting process martingale, which is defined for the i^{th} subject as

$$e_i^m(t_i) = N_i(t) - \int_0^t \Delta_i(s)\widehat{h}_0(s)\exp\{\mathbf{w}_i\widehat{\boldsymbol{\gamma}} + \eta\widehat{m}_i(s)\mathrm{d}s\}, \tag{4.15}$$

where $N_i(t)$ is the counting process denoting the number of events for subject i by time t, $\Delta_i(t)$ is the left continuous at risk process with $\Delta_i(t) = 1$ if subject i is at risk at time t and 0 otherwise, $\widehat{m}_i(t) = \mathbf{x}_i^\top(t)\widehat{\beta} - \mathbf{z}_i^\top(t)\widehat{\mathbf{b}}_i$, and $\widehat{h}_0(t)$ denotes the estimated baseline hazard function. The martingale residuals can be viewed as the difference between the

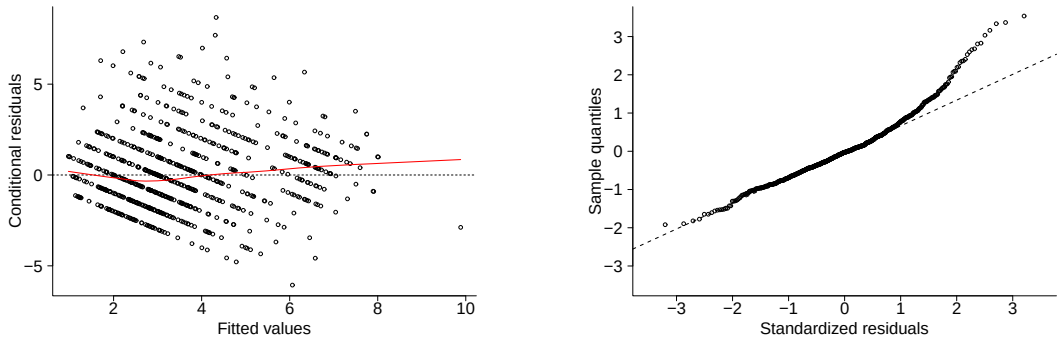


Figure 4.11: Left: Conditional residuals against fitted values (overlaid red curve is the loess smoother); right: Q-Q plot of standardized residuals.

observed number of events for the i^{th} subject by time t, and the expected number of events by the same time based on the fitted model (Barlow & Prentice, 1988; Therneau *et al.*, 1990).

The top left plot of Figure 4.11 shows that the standardized residuals are roughly evenly spread about zero along the fitted values of the longitudinal outcome, $\widehat{y}_i(t)$, with the loess smoother depicting almost a horizontal line. In general the formation of stripes in this kind of diagnostics plots occurs when individuals with the same observed values have very different fitted values. The longitudinal outcome of model M_3^{CARE} is is the frailty score measured in the EFS, so the stripes on the plot are due the fact that frailty in the EFS takes integer values between 0 and 17. The Q-Q plot at the top right of Figure 4.11 shows signs of right skewness because of too many large values of frailty.

Figure 4.12 contains plots of residuals against covariates, showing in all them that residuals are centered around zero. There seems to be a slight sign of heterogeneity in the plot of `etnicity`, which might be due to the two groups being highly unbalanced (White: 241 vs Other: 41).

Dobson & Henderson (2003) showed that residuals between observed and expected longitudinal outcome responses after fitting a joint model can be affected by informative dropout. In order to take into account the effects of informative dropout in the assessment of model adequacy, they proposed a residual analysis conditioning upon

dropout time and type (T_i, δ_i^D). That is, their proposal is based on the conditional expectation of the residuals given (T_i, δ_i^D). Our residual analysis of the longitudinal outcome can be complemented by testing for random dropouts caused by mortality and analyzing the expected residuals given the time-to-death, following the method proposed by Dobson & Henderson (2003).

As we saw in the previous section, the available data set for the CARE75+ study suggest that frailty and is associated to several covariates: ethnicity, marital status, education level, drinking alcohol and having more comorbidities. However, due to the relatively short follow up time, and the relatively low number of terminal events and falls there might be less information to detect significant associations of mortality and falls with covariates. The residual analysis of falls and mortality submodels shown in Figure 4.13 indicate that both falls and mortality are overestimated (estimates are often greater than observed values). It might also be worth investigating with different specifications of the joint model for instance, an alternative setting for the baseline hazard (different distribution of a parametric baseline or a B-splines) or a different distribution of the random effects u_i.

4.5 Conclusion

We used joint modelling to explore the relationship between frailty, falls and mortality in the CARE75+ study. In fitting model $\widehat{M}_3^{\text{CARE}}$ to the data we assumed that these three outcomes are associated via latent variables that represent unobserved features of the subjects in the study. The main findings are:

- Several covariates are associated to frailty. Being white, married or remarried, having higher education and drinking alcohol are associated to lower frailty. In the opposite direction, the greater the number of comorbidities, the more frail people are.
- There is a higher risk of falls among white people. The relative risk of falls with respect to the other ethnicities is $\exp(1.429) = 4.175$.

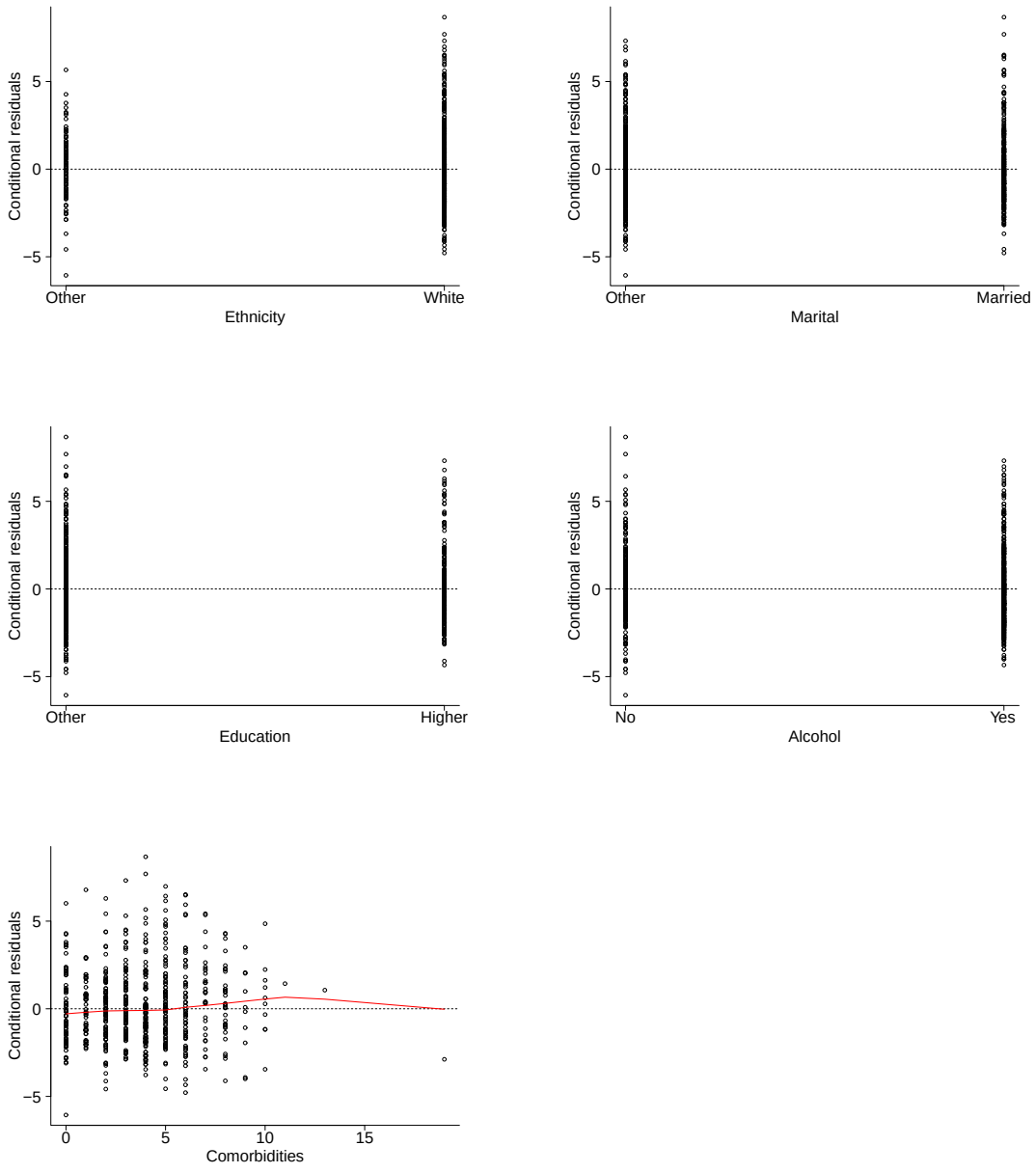


Figure 4.12: Conditional residuals against covariates (○) with overlaid loess smoothers —.

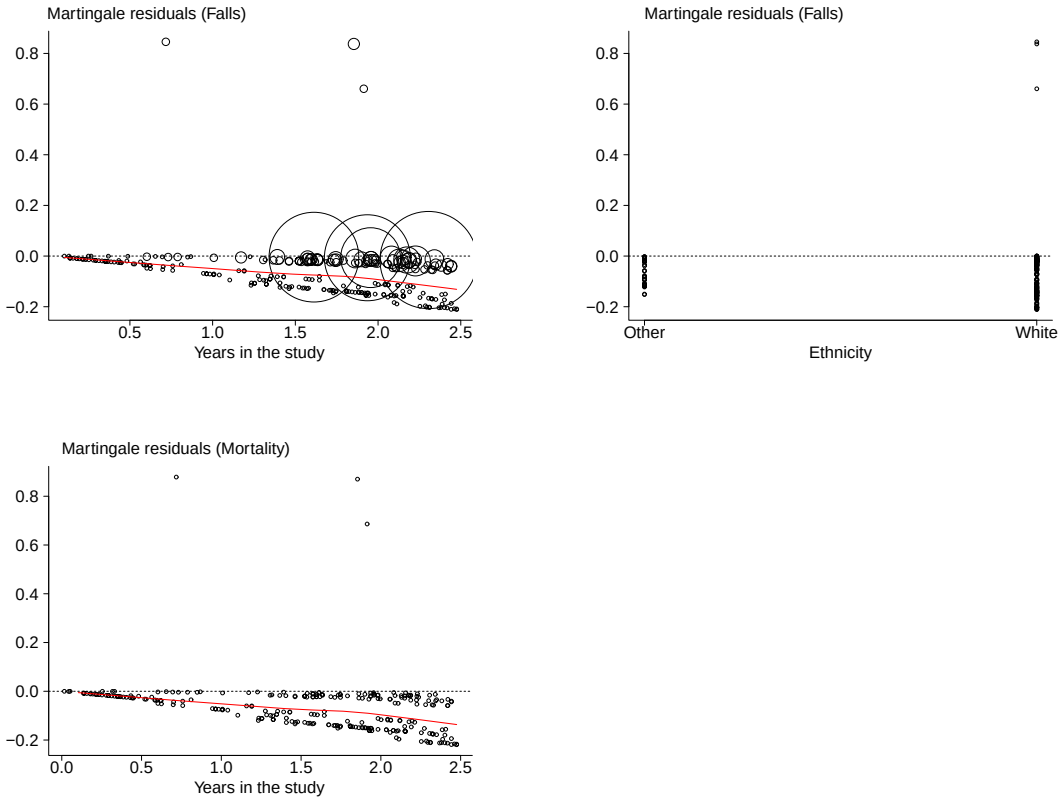


Figure 4.13: Top left: Martingale residuals of the falls submodel against time, (the size of the points is relative to the number of falls by each participant); top right: Martingale residuals against ethnicity; bottom: Martingale residuals of the mortality submodel against time. The red curves are loess smoothers.

- None of the covariates we explored with in the CARE75+ data set is significantly associated to mortality.
- Frailty is strongly associated with both falls and mortality, an association that is characterized by a random intercept in the linear mixed submodel of frailty. The relative risk of falls is $\exp(0.439) = 1.551$ for subjects whose frailty is one unit above the population average, *ceteris paribus*, and the relative risk of mortality is $\exp(0.430) = 1.537$.
- Apparently, falls have no effect on mortality as their association parameter was not significant at a 0.05 level. Although this result might seem counterintuitive, this result might be due to the short follow-up period. As we mentioned previously, the epidemiology of falls is complex, and when falls are non-fatal its relationship with death is mediated by its multiple adverse outcomes. Falls related accidents lead to several adverse outcomes in older people: long-term institutional care, hospital admissions, injuries, fear of falling after an accidental fall, reduced activities of daily living and lower quality of life Masud & Morris (2001). Even though these falls related outcomes increase the risk of death, it might require a longer follow-up period than is available in the CARE75+ study.
- The model diagnostics indicate some signs of skewness due to many large values of frailty. Further refinements and considerations might be required since, as Dobson & Henderson (2003) have shown, residuals of the longitudinal outcome might be affected when dropout is informative, which we still need to investigate in our analysis. Additionally, in both falls and mortality submodels the fitted model $\widehat{M}_3^{\text{CARE}}$ seems to overestimate the the risk of falls and mortality. Perhaps this is due to (1) the relatively small sample size for joint modelling of longitudinal data and recurrent and terminal events, (2) the small number terminal events, and (3) the lack of precise times in which falls occur.
- Regarding this last point we estimated the time-to-falls based on the count data provided on the data set coming from participants being asked about the number of times they fell within a 12 months window. The time-to-fall series was estimated assuming that falls occurred uniformly in the relevant time period.

Alternative specifications of a joint model for the relationship between frailty, falls and mortality are possible. Given that falls and mortality are not significantly associated, it is worth exploring frailty, falls and mortality with a joint model assuming that falls is an exogenous time-varying covariate for mortality. Additionally, in model M_3^{CARE} we assumed that frailty and mortality are related through the random intercept (b_{i0}) representing the deviation of the subject-specific frailty with respect to the population average at baseline, and with model M_2^{CARE} we would like to explore with a different specification of the frailty-mortality relationship by assuming that frailty free of measurement error $(m_i(t))$ is an endogenous time-varying covariate of mortality. Additionally, the relationship frailty-falls can be analyzed by assuming that this relationship can be expressed as an effect of falls on frailty, this is the number of falls is a time-varying covariate for in the frailty submodel. This model would have two regression equations so we denote it by M_2^{CARE} and it is described by Equations (4.16a)–(4.16b) under the following assumptions:

- Falls is an exogenous time-varying covariate of mortality.
- The link between frailty and mortality is the current level of frailty free of measurement error, i.e. the linear predictor of frailty in the linear mixed submodel.
- Falls is a time-varying covariate of frailty.
- No random effect in the submodel of mortality, i.e. assume no heterogeneity of the hazards due to unobserved covariates.
- The baseline hazard of the mortality submodel is approximated with B-splines instead of assuming a parametric form.

$$M_2^{\text{CARE}}: \begin{cases} y_i(t \mid b_{i0}) = \underbrace{(\beta_0 + b_{i0}) + \mathbf{x}_i^\top(t)\boldsymbol{\beta} + \beta_{\text{falls}} N_i(t)}_{m_i(t)} + \varepsilon_i(t) & \text{(4.16a)} \\ h_i(t \mid b_{i0}) = h_0(t)\exp\{\gamma_{T.\text{eth}}\texttt{eth}_i + \gamma_{T.\text{falls}} N_i(t) + \eta_L m_i(t)\}, & \text{(4.16b)} \end{cases}$$

Note that in model M_2^{CARE} there is no regression submodel for falls, which now is a covariate in the mortality submodel (Equation 4.16b). Also, the whole linear predictor of the frailty submodel, which includes the random intercept, is a time-varying covariate in the mortality submodel with η_L quantifying the strength of the association between frailty and mortality. The fitted model is $\widehat{M}_2^{\text{CARE}}$, and its estimates are shown in Table 4.5 alongside with the estimates of $\widehat{M}_3^{\text{CARE}}$.

	$\widehat{M}_3^{\text{CARE}}$			$\widehat{M}_2^{\text{CARE}}$		
Parameter	Estimate	Std.Err	p-value	Estimate	Std.Err	p-value
Fixed effects						
Frailty						
$\widehat{\beta}_0$	6.166	0.343	< 0.001	6.096	0.328	< 0.001
$\widehat{\beta}_t$	-0.029	0.111	0.793	-0.094	0.111	0.396
$\widehat{\beta}_{\text{eth}}$	-2.026	0.354	< 0.001	-2.154	0.341	< 0.001
$\widehat{\beta}_{\text{mar}}$	-0.729	0.211	< 0.001	-0.823	0.210	< 0.001
$\widehat{\beta}_{\text{edu}}$	-0.856	0.223	< 0.001	-0.815	0.223	< 0.001
$\widehat{\beta}_{\text{alc}}$	-1.247	0.233	< 0.001	-1.232	0.234	< 0.001
$\widehat{\beta}_{\text{com}}$	0.234	0.031	< 0.001	0.251	0.030	< 0.001
$\widehat{\beta}_{\text{falls}}$	$+$	$+$	$+$	0.329	0.051	< 0.001
Falls						
$\widehat{\gamma}_{R.\text{eth}}$	0.866	0.367	0.029	$+$	$+$	$+$
Mortality						
$\widehat{\gamma}_{T.\text{eth}}$	-0.077	0.703	0.912	1.566	0.796	0.049
$\widehat{\gamma}_{T.\text{falls}}$	$+$	$+$	$+$	-0.259	0.235	0.269
Association						
$\widehat{\eta}_R$	0.439	0.081	< 0.001	$+$	$+$	$+$
$\widehat{\eta}_T$	0.430	0.206	0.037	$+$	$+$	$+$
$\widehat{\alpha}$	-0.962	0.703	0.171	$+$	$+$	$+$
$\widehat{\eta}_L$	$+$	$+$	$+$	0.511	0.150	< 0.001
Variance component						
Frailty						
$\widehat{\sigma}_\varepsilon$	2.152	$-$	$-$	1.459	$-$	$-$
$\widehat{\sigma}_b^2$	2.577	$-$	$-$	2.188	$-$	$-$
$\widehat{\phi}$	1.143	0.849	0.0891	$+$	$+$	$+$
Baseline hazard						
Falls						
$\widehat{\kappa}_R$	1.023	$-$	$-$	$+$	$+$	$+$
$\widehat{\rho}_R$	0.322	$-$	$-$	$+$	$+$	$+$
Mortality						
$\widehat{\kappa}_T$	1.040	$-$	$-$	$+$	$+$	$+$
$\widehat{\rho}_T$	0.170	$-$	$-$	$+$	$+$	$+$

Table 4.5: M_3^{CARE} and M_2^{CARE} estimates. "$-$" not directly available from the software output. "$+$" not a parameter to be estimated by the model on the column.

The estimates of $\widehat{M}_2^{\text{CARE}}$ and $\widehat{M}_3^{\text{CARE}}$ are consistent with each other and, fundamentally, the conclusion about association between frailty and mortality is the same. However, in model $\widehat{M}_2^{\text{CARE}}$ the association parameter of frailty–mortality relationship, $\widehat{\eta}_T$, represents the estimate of the effect of frailty on the relative hazard of mortality of subject i, both at time t.

4.5.1 Discussion

As we saw in this chapter, variable selection in joint models is complicated for several reasons.

1. In general, fitting a joint model is complicated because the corresponding likelihood is difficult to optimize. Due to the conditional independence assumption of the outcomes given the random effects, the likelihood function of joint models contains integrals with respect to the random effects that have to be done numerically, so the processing times required to optimize this function are much longer compared to the standard marginal models. The processing time increases quickly with the number of regression equations and the number of random effects, especially in models like M_3^{CARE} containing a recurrent event submodel. Depending on the specification of the baseline hazards, the likelihood might also contain integrals of non-linear functions of time for which numerical methods are needed, adding to the complexity of the optimization task, especially if a recurrent event submodel is included in the joint model formulation. The number of observations is one more element that has a material effect in the processing time. In our experience, fitting a joint model for for the CARE75+ can take between 13 minutes to 6 hours, and it is not unusual for the optimization algorithm to fail to converge. Non-parametric spline-based approximations are commonly considered for the baseline hazards, with the advantage that no assumptions are required about their distribution, but it makes the output less straightforward to interpret and the estimated model less friendly to be used with other data sets.

2. The total number of all possible covariate combinations increases with the number of submodels and it is not straightforward to understand the effect of having the same covariates in two or more submodels.

3. Choosing the link to characterize the association between submodels involves a large number of options since in principle any function of the random effects is possible, and it is still unclear if different links affect in different ways the significance of covariates in a joint model.

In statistical modelling, variable selection is carried out in different ways depending on the intended use of the fitted model: *description*, *causal inference* or *prediction* (Shmueli *et al.*, 2010). In this chapter, we explored with joint models the relationship between frailty, falls and mortality in the CARE75+ data set, for which we specified and fitted two joint models: $\widehat{M}_3^{\text{CARE}}$ and $\widehat{M}_2^{\text{CARE}}$. The covariates in these two models were selected based on their statistical significance, a strategy that is not in line with causal inference. For causal inference we would need to state upfront our hypotheses about the relationships among all the available variables, identifying all possible confounders to the frailty-falls-mortality relationship, to finally fit the joint model adjusting for confounding, which implies keeping in each submodel model all confounders even when their associated regression coefficient were not significant. DAGs have proven to be a useful tool for stating the hypothesized relationships among all the variables of a model. Furthermore, it is not evident which model between $\widehat{M}_2^{\text{CARE}}$ and $\widehat{M}_3^{\text{CARE}}$ is better to describe the frailty-falls-mortality relationship. So with the help of DAGs, in Chapter 6 we conduct a simulation study by first specifying two causal models similar to $\widehat{M}_2^{\text{CARE}}$ and $\widehat{M}_3^{\text{CARE}}$, simulating a series of data sets from these models and finally analyzing each data set with the two models. This allows us the to investigate the consequences of model misspecification.

On the other hand, models $\widehat{M}_2^{\text{CARE}}$ and $\widehat{M}_3^{\text{CARE}}$ fitted in this chapter would not be appropriate for out-of-sample predictions since the parameters were estimated by optimizing the likelihood function and that would not necessarily lead to optimized out-of-sample predictions. In $\widehat{M}_3^{\text{CARE}}$ and $\widehat{M}_2^{\text{CARE}}$ `ethnicity` is a covariate in all the submodels, but in general it might be the case that several covariates appear in more than one submodel, and it is not clear if this overfits the model. Additionally, we noticed that some variables of the CARE75+ data are correlated, as shown in Table 5.1 that contains the correlations and p-values for the null hypothesis that the correlations are equal to zero. The highest correlation is between ethnicity and alcohol (0.51), followed by the correlation between gender and marital status (0.36). Other correlations statistically

	Gender	BMI	Ethnicity	Marital	Education	Smoke	Alcohol
BMI	−0.03 (0.63)						
Ethnicity	−0.10 (0.68)	0.07 (0.22)					
Marital	0.36 (0.00)	−0.09 (0.14)	−0.09 (0.15)				
Education	0.06 (0.29)	0.07 (0.23)	0.20 (0.00)	0.01 (0.86)			
Smoke	0.10 (0.08)	−0.06 (0.29)	0.26 (0.00)	0.04 (0.51)	0.06 (0.33)		
Alcohol	−0.02 (0.74)	0.10 (0.11)	0.51 (0.00)	−0.09 (0.15)	0.18 (0.00)	0.19 (0.00)	
Falls	−0.01 (0.88)	−0.05 (0.39)	0.09 (0.14)	0.06 (0.32)	−0.04 (0.51)	0.00 (0.96)	−0.06 (0.32)

Table 4.6: Covariates correlation matrix (and p-values associated with the null hypothesis test of zero-correlation.)

different from zero are between ethnicity and education and smoking, and between alcohol and education and smoking.

In Chapters 5 and 6 we address variable selection in joint modelling when the aim of the model is prediction and causal inference. In Chapter 5 we propose a strategy to select variables in a joint model of longitudinal and time-to-event outcomes with the goal of optimizing prediction of both outcomes. Even in the simplest case, joint modelling involves estimating a large number of parameters compared to the standard marginal models, so larger sample sizes are required. We explore the performance of our strategy in simulation studies of small sample sizes datasets. The simulation model is similar to model $\widehat{M}_2^{\text{CARE}}$ and then we apply it to the CARE75+ data set to optimize prediction of frailty and mortality as joint model outcomes.

In Chapter 6 we explore with the two different settings of joint models, similar to $\widehat{M}_2^{\text{CARE}}$ and $\widehat{M}_3^{\text{CARE}}$, but with the difference that we pay special attention to the confounding structure with the help of DAGs. We simulate a series of data sets from these two models and analyze each data set with the two models in order to understand the consequences of model misspecification.

4.5.2 Future work

As mentioned before, further research is needed with respect to model diagnostics. We can complement the residual analysis of frailty with the method proposed by Dobson & Henderson (2003). We are assuming proportional hazards in the falls and mortality submodels, an assumption that we still need to verify.

We want to plot the fitted model in some way that would be useful for interpretations. Plotting the average frailty and fitted hazards for some fictional / representative subject might be helpful.

We would like to conduct some sensitivity analysis to explore with model uncertainty. In particular, if different covariates were in the final model, we are interested in knowing how much do the fitted curves (frailty profiles and cumulative hazards) change.

Chapter 5

Prediction accuracy and variable selection for joint models of longitudinal and time-to-event data

5.1 Introduction

The presence of highly correlated covariates affects the test statistics of the parameters' estimates by inflating the estimated variance (Rawlings *et al.*, 2001; Seber & Lee, 2012), producing less stable results. In Chapter 4 the joint analysis of frailty, recurrent falls and mortality using the CARE75+ data suggested that mortality is associated with frailty and correlated covariates. In Table 5.1 we see statistically significant correlations ($\alpha < 0.05$) between gender and marital status, between ethnicity and education, smoking and alcohol consumption, and between education and alcohol consumption. Whether correlation is high enough to give problems with model fitting depends also on the sample size. The CARE75+ data set has 282 subjects, which is a relatively small sample size for joint modelling frailty, falls and mortality, and there are only 17 deaths among them. Our aim for this chapter is to find the joint model for frailty and mortality that optimizes out-of-sample predictions of both outcomes. For this, we explore with a combined strategy of penalized likelihood (Section 2.4) and cross-validation to select

the fixed effects covariates that optimize prediction of a joint model of longitudinal and time-to-event data.

	Gender	BMI	Ethnicity	Marital	Education	Smoke	Alcohol
BMI	−0.03 (0.63)						
Ethnicity	−0.10 (0.68)	0.07 (0.22)					
Marital	0.36 (0.00)	−0.09 (0.14)	−0.09 (0.15)				
Education	0.06 (0.29)	0.07 (0.23)	0.20 (0.00)	0.01 (0.86)			
Smoke	0.10 (0.08)	−0.06 (0.29)	0.26 (0.00)	0.04 (0.51)	0.06 (0.33)		
Alcohol	−0.02 (0.74)	0.10 (0.11)	0.51 (0.00)	−0.09 (0.15)	0.18 (0.00)	0.19 (0.00)	
Falls	−0.01 (0.88)	−0.05 (0.39)	0.09 (0.14)	0.06 (0.32)	−0.04 (0.51)	0.00 (0.96)	−0.06 (0.32)

Table 5.1: Covariates correlation matrix (and p-values associated with the null hypothesis test of zero-correlation.)

By using penalized likelihood methods, like ridge regression (Hoerl & Kennard, 1970) and the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), we can get around the problem caused by highly correlated covariates in regression analysis, as discussed in Section 2.4.

Variable selection in linear mixed and survival analysis models has been an active research topic. For instance, in the context of mixed effects models Fan & Li (2001) and Peng & Lu (2012) discussed asymptotic "oracle" properties of the the smoothly clipped absolute deviation (SCAD) penalty function to select groups of the random effects and Zou (2006) proposed the adaptive LASSO (ALASSO). In the survival analysis context, Fan & Li (2002) and Zhang & Lu (2007) discussed variable selection of the fixed effects, and Benner *et al.* (2010) used high-dimensional data simulations to compare the oracle properties of various penalties and the predictive performance of the resulting model with the integrated Brier score (IBS). On this last point, He *et al.* (2015) compared the results of using LASSO, elastic net, SCAD and ALASSO for variable selection, recommending the use of LASSO and elastic net.

Variable selection for joint models of longitudinal and time-to-event data via penalized likelihood methods is a topic that has been studied with the aim to maximize goodness of fit. Chen & Wang (2017) followed a penalized likelihood approach to select the fixed and random effects in joint modelling longitudinal and time-to-event outcomes, by imposing four penalty functions (fixed & random effects of each submodel) of the adaptive least absolute shrinkage and selection operator (ALASSO) with the

goal to minimize the Bayesian Information Criterion (BIC). He *et al.* (2015) imposed L_1-norm penalized likelihood for selecting the random effects that minimized the BIC. We are interested in optimizing the accuracy of predictions of a joint model, and optimizing goodness of fit only provides accuracy within the data used for modelling (Anderson, 2007; Burnham & Anderson, 2003, 2004). In order to assess the accuracy of a model in out-of-sample data we require other techniques, such as bootstrap or cross-validation.

Our interest is to find the subset of covariates that are most relevant for out-of-sample predictions. We aim for assessing the effect of frailty on mortality and for making predictions for both frailty and mortality using the CARE75+ data. Since frailty and mortality are not independent they need to be analyzed as joint outcomes, for which joint modelling is an appropriate framework. However, there are some features of the CARE75+ data set that impose challenges for joint modelling and that must be addressed: (a) the CARE75+ data set is relatively small $(n = 282)$, (b) the number of deaths is too small, and (c) the data set contains correlated covariates. Therefore, we need a strategy for variable selection to account for all these issues. We propose using a penalized likelihood approach and cross-validation to select the set of covariates that maximize the accuracy of predictions of both outcomes.

Shrinkage methods yield parameter estimates with smaller variance sacrificing a small amount of bias. Intuitively, the decrease in variance results naturally from restricting the sampling space of regression coefficients when applying shrinkage methods (Hastie *et al.*, 2009; Heinze *et al.*, 2018, p.225). Therefore, shrinkage methods are useful in statistical modelling where the focus is on obtaining accurate predictions, that is predictions with a small mean-squared error (MSE), and where models are hard to estimate due to approximate colinearity. Regression coefficients for which selection is less stable are shrunken more strongly than coefficients for which selection bias is more stable.

Our proposed strategy has a K-fold cross-validation design. A joint model is "trained" with each fold, where the objective function is the penalized log-likelihood, imposing separate L_1-norm penalties to the fixed-effects regression coefficients of each submodel. The estimated parameters are then used to score the $K - 1$ remaining folds of data. By imposing L_1-norm penalties some of the regression coefficients will shrink to

zero so this stage of the the strategy will perform variable selection. By scoring each test set with the trained model, we are exploring the ability of the model to predict the outcomes in out-of-sample data. In practice, when joint modelling longitudinal and time-to-event data the focus can be on one outcome: (a) the longitudinal outcome if there are informative dropouts represented by the survival analysis submodel, and (b) the time-to-event outcome when the quantitative one is an internal time-varying covariate of a survival analysis model, or in both if the aim is to explore their joint distribution. We are interested in exploring the joint distribution of both outcomes. By being able to get accurate predictions of frailty and mortality for the CARE75+ data it might be possible to adapt the management plans for care homes. The accuracy of predictions of the longitudinal outcome is assessed by the MSE, and the time-to-event outcome by the integrated Brier score (IBS). Ideally, we would like to minimize simultaneously the MSE and the IBS.

We applied this method in simulation experiments and in the CARE75+ data to study its performance in predicting mortality and frailty. The results of the experiments suggest that it is not always possible to simultaneously optimize the accuracy of predictions of the two outcomes; however it was useful to identify a subset of covariates for the CARE75+ data where the prediction accuracy of both outcomes are better.

5.2 Methodology

5.2.1 Model specification

As discussed in Section 3.1, the joint modelling of longitudinal and time-to-event outcomes requires the specification of a regression submodel for each outcome, a covariance structure for random effects which are assumed to act in both submodels, and a link function to connect both submodels. The joint model we explore in this chapter considers 1) a random intercept assumed to be normally distributed with mean zero and variance σ_b^2, and 2) that both outcomes are linked by this random intercept. Equations

(5.1a)–(5.1b) describe the general form of this joint model:

$$M^{(\mathrm{S})}: \begin{cases} \text{Longitudinal} & y_i(t \mid b_{i0}) = (\beta_0 + b_{i0}) + \mathbf{x}_i^\top(t)\boldsymbol{\beta} + \varepsilon_i(t) \quad (5.1a) \\ \text{Terminal} & h_i(t \mid b_{i0}) = h_0(t)\exp\{\mathbf{w}_i^\top\boldsymbol{\gamma} + \eta b_{i0}\}, \quad (5.1b) \end{cases}$$

where

$\mathbf{w}_i^\top = p$-vector of time-fixed covariates,
$\boldsymbol{\gamma}^\top = (\gamma_1, \ldots, \gamma_p)$,
$\mathbf{x}_i^\top(t) = (t, \mathbf{w}_i^\top)$ is a $(p+1)$-vector,
$\boldsymbol{\beta}^\top = (\beta_t, \beta_1, \ldots, \beta_p)$.

In the context of the CARE75+ data, the random intercept in $M^{(\mathrm{S})}$ models the correlation of frailty observed at time-points as well as the correlation between frailty and mortality.

Note that for the moment there are no time-varying covariates in the survival analysis submodel, and the only time-varying covariantes in the linear mixed submodel is time, t.

Equation (5.1a) is the linear mixed model formulation of the longitudinal outcome of subject i, where $m_i(t) = (\beta_0 + b_{i0}) + \mathbf{x}_i^\top(t)\boldsymbol{\beta}$ represents the true value of the longitudinal outcome at time t. Here, $y_i(t), m_i(t)$ and $\varepsilon_i(t)$ are scalars, and $\mathbf{x}_i(t)$ is a $(p+1)$-vector of baseline and possibly time-dependent covariates whose values are recorded at time t. The $(p+1)$-vector $\boldsymbol{\beta}$ denotes the fixed effects regression coefficients. The random intercept, b_{i0}, represents a shift of the longitudinal outcome profile of subject i with respect to the population profile, $\beta_0 + \beta_t t$.

Equation (5.1b) is the regression submodel for the hazard rate of of a time-to-event outcome, where $\mathbf{w}_i$ is the p-vector of time-fixed covariates for subject i. The p-vector $\boldsymbol{\gamma}$ contains the regression coefficients for these covariates. The link between the longitudinal and the time-to-event outcomes is $g(b_{i0}, t) = b_{i0}$, and η quantifies the strength of the association between the two outcomes.

We make the usual assumptions of conditional independence between (1) the longitudinal and the time-to-event outcomes and (2) frailty measurements at different

time points of the same person given the random effect. We also assume a normal distribution for the measurement error and the random intercept, $\varepsilon_i(t) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$, $b_{i0} \sim \mathcal{N}(0, \sigma_b^2)$, with $h_0(t)$ the baseline hazard of the survival analysis submodel.

5.2.2 Variable selection strategy

Our proposed strategy for variable selection combines a penalized likelihood approach and 5-fold cross-validation. We impose separate L_1-norm penalties (Chen *et al.*, 2001; Tibshirani, 1996) for the fixed-effects coefficients of each submodel to perform variable selection. The penalized log-likelihood for a sample with data $\mathcal{D} = \{t_i, \delta_i, \mathbf{y}_i;\ i = 1, \ldots, n\}$ is

$$\ell_\lambda(\boldsymbol{\theta} \mid \mathcal{D}) = \ell(\boldsymbol{\theta} \mid \mathcal{D}) + \lambda_L \|\boldsymbol{\beta}\|_1 + \lambda_S \|\boldsymbol{\gamma}\|_1, \tag{5.2}$$

where $\ell(\boldsymbol{\theta} \mid \mathcal{D})$ is the unpenalized likelihood described by Equation (3.2), $\|\boldsymbol{\beta}\|_r := \left(\sum_{k=1}^{p} |\boldsymbol{\beta}_i|^r\right)^{1/r}$ is the r-norm of the p-vector $\boldsymbol{\beta}$, and $\lambda = (\lambda_L, \lambda_S)$ are hyperparameters.

Our proposed strategy is summarized in Algorithm 5.1. The cross-validation procedure of Algorithm 5.1 starts by splitting the data set $(\mathcal{D})$ at random, with respect to the subject identifier into K roughly equal size and non-overlapping groups. In a data set of n subjects each fold should contain the data of approximately n/K different subjects. For each fold $(k = 1, \ldots, K)$ the test set $(\mathcal{D}_{(k)}^{\texttt{test}})$, is comprised by the data of subjects in the k^{th} group, and the training set $(\mathcal{D}_{(k)}^{\texttt{train}})$ by the data of subjects in the other $K - 1$ groups (Step 1). The parameters are estimated with $\mathcal{D}_{(k)}^{\texttt{train}}$, and the estimates are used to score $\mathcal{D}_{(k)}^{\texttt{test}}$. Once this procedure is repeated for the K folds, the data of each subject in $\mathcal{D}$ should have been used to estimate the parameters $K - 1$ times and scored once (Step (b)).

Notice that Step 4 of Algorithm 5.1 assumes there is a unique pair (λ_L, λ_S) that optimizes simultaneously the accuracy of predictions of both outcomes, which might not be the case. The solution might be an overlapping region of the grid of (λ_L, λ_S) values that minimizes the MSE and the grid of (λ_L, λ_S) that minimizes the IBS. Even more, these two (λ_L, λ_S) regions might very well have no values in common, in which

Algorithm 5.1 Variable selection algorithm to maximize the accuracy of predictions.

0: Choose the number of folds K.
1: Split the data, $\mathcal{D}$, into $\{\mathcal{D}_{(k)}^{\texttt{train}}, \mathcal{D}_{(k)}^{\texttt{test}}\}$, $k = 1, \ldots, K$.
2: For a grid of values for λ_L and λ_S
   (a) Estimate the parameters of the joint model:

$$\widehat{\boldsymbol{\theta}}_{(k)} = \arg\max_{\boldsymbol{\theta}} \ell_\lambda \left(\boldsymbol{\theta} \mid \mathcal{D}_{(k)}^{\texttt{train}} \right)$$

   (b) Evaluate the performance of $\widehat{\boldsymbol{\theta}}_{(k)}$ on the test set:
      (i) For each subject $i = 1, \ldots, n^{\text{test}}$
         - Predict the random effects, $\widehat{b}_{i0}$, as described by Equations (3.6). Here we are assume that each subject has only one longitudinal outcome measure at baseline, y_{i1}, and that the last time the subject was known to be alive was also at baseline, i.e. $\tau_i = t_{i1}$. So the densities of the posterior (3.5) are as described by Equations (3.19a)–(3.20c).
         - Compute the expected value of the longitudinal outcome $\widehat{\omega}(t \mid t_{i1})$ as described by Equation (3.8).
         - Compute the residual survival probabilities $\widehat{\pi}_i(u \mid t_{i1})$ for $u \in (0, t^*)$, as described by Equation (3.11). Since $t_{i1} = 0$, $\widehat{\pi}_i(u \mid t_{i1}) = \widehat{S}_i(u \mid t_{i1})$.
      (ii) Compute the MSE as in Equation (2.46) and the IBS as in Equation (2.54).
3: Compute cross-validated MSE (MSE_{CV}) and IBS (IBS_{CV}) by averaging over the K folds.
4: Desired solution: $(\lambda_L^*, \lambda_S^*)$ for which (MSE_{CV}) and (IBS_{CV}) are minimal, and parameter estimates.

case the solution will be a compromise between the accuracy of prediction of one outcome.

It is important to emphasize that in order to evaluate our strategy, predictions of the random effect on the test set are carried out ignoring all subsequent repeated measures after baseline and assume that the last time these subjects were known to be event-free was exactly at baseline.

5.3 Simulation studies

5.3.1 Design

We are interested in the results of the variable selection strategy when there are correlated variables. In particular, we would like to know it makes any difference if correlated covariates are in the same submodel when the goal is to optimize prediction.

To test the proposed algorithm we simulated six data sets from the joint model of a longitudinal and a time-to-event outcomes $M^{(\mathrm{S})}$, with seven fixed effects and time fixed covariates $\mathbf{w}^\top = (w_1, \ldots, w_7)$, so $p = 7$. The measurement error of the longitudinal outcome of model $M^{(\mathrm{S})}$ has constant variance, $\mathrm{var}(\varepsilon_i(t)) = \sigma^2_\varepsilon = 4$. The variance of the random intercept of the linear mixed submodel in Equation (5.1a) is $\sigma^2_b = 4$. The baseline hazard of the time-to-event submodel in Equation (5.1b) is governed by the $\mathrm{Weibull}(\kappa, \rho)$ distribution, so $h_0(t) = h_0(t; \kappa, \rho) = \kappa\rho(\rho t)^{\kappa-1}$, where $\kappa = 2$ and $\rho = 1.5$ are the shape and rate parameters, respectively. For each individual of the simulated data, we simulated seven covariates: $w_{i1}, w_{i2}, w_{i3}, w_{i4} \sim \mathrm{Bernoulli}(0.5)$, and $(w_{i5}, w_{i6}, w_{i7}) \sim \mathcal{N}_3(\mathbf{0}, \Sigma_w)$, where the covariance matrix Σ_w indicates that w_5 is uncorrelated with w_6 and w_7, and $\mathrm{cor}(w_6, w_7) = 0.95$ as follows:

$$\Sigma_w = \begin{bmatrix} 1 & & \\ 0 & 2 & \\ 0 & 1.9 & 2 \end{bmatrix}.$$

Table 6.3 summarizes the values of the parameters and regression coefficients of this simulation experiment.

Our simulation study has only one simulation for each scenario because the processing time is too long. Our design is 5-fold cross validation with 64 combinations of the two penalties, so in each scenario the training-test procedure is repeated 320 times. On average, with a sample size $n = 50$ (scenarios 1 and 3) each training-test procedure takes 26 minutes (5 days and 19 hours to complete the 320 training-test routine). With the larger sample size $n = 250$ (scenarios 2, 4, 5 and 6), on average each training-test procedure takes 2 hours and 6 minutes, so completing the 320 training-test routine in each scenario takes 28 days and 2 hours. In addition, computing the Hessian at each estimation step takes as much as the training-test procedure. We parallelized the task by splitting the work in 8 parts and processing it simultaneously in 8 computers, each scenario 1 and 3 $(n = 50)$ taking 1.5 days to complete, and scenarios 2, 4, 5 and 6 taking 7 days each one.

All six simulated data sets were generated the joint model $M^{(\mathrm{S})}$ described by Equations (5.1a)–(5.1b) but, as Table 6.3 points out, they differ in either the sample size or the non-zero regression coefficients. We considered four sets of non-zero regression coefficients and labeled these scenarios as $M_{12}^{(\mathrm{S})}, M_{34}^{(\mathrm{S})}, M_{5}^{(\mathrm{S})}$ and $M_{6}^{(\mathrm{S})}$:

- Simulations 1 and 2 are generated from model $M(\mathrm{S})$ with non-zero fixed effects regression coefficients $(\beta_t, \beta_1, \beta_4, \gamma_1, \gamma_2)$, and we identify this scenario as $M_{12}^{(\mathrm{S})}$. The sample size of simulations 1 and 2 are 50 and 250, respectively.
- Simulations 3 and 4 are generated from model $M^{(\mathrm{S})}$ with non-zero fixed effects regression coefficients $(\beta_t, \beta_1, \beta_4, \gamma_1, \gamma_2, \gamma_7)$ and we identify this scenario as $M_{34}^{(\mathrm{S})}$. The sample size of simulations 3 and 4 are 50 and 250 respectively. The difference between $M_{12}^{(\mathrm{S})}$ and $M_{34}^{(\mathrm{S})}$ is that γ_7 in the latter is non-zero, whose associated covariate, w_7, is highly correlated with w_6. This is, among the two highly correlated covariates, w_7 has a non-zero regression coefficient only in the survival analysis submodel.
- Simulation 5 has a sample size of 250, and the set non-zero regression coefficients in this scenario, labeled as $M_{5}^{(\mathrm{S})}$, is $(\beta_t, \beta_1, \beta_4, \beta_7, \gamma_1, \gamma_2)$. In contrast to

$M_{34}^{(S)}$, the covariate w_7 in $M_5^{(S)}$ has non-zero regression coefficient only in the linear-mixed submodel.

- Simulation 6, with sample size 250, has a set of non-zero regression coefficients $(\beta_t, \beta_1, \beta_4, \beta_7, \gamma_1, \gamma_2, \gamma_7)$ labeled as $M_6^{(S)}$. In this scenario, w_7 has non-zero regression coefficient in both linear-mixed and survival analysis submodels.

The repeated measures of the longitudinal outcome for each subject, $y_i(t_{ij})$, were simulated every 0.2 time units (so $t_{ij} = 0, 0.2, \ldots$) with a maximum of 20 repeated measures. The censoring time was fixed at 3 to have less than 10% of censored survival times. And the accuracy of predictions was assessed from $u \in (0, t^*)$ with $t^* = 3$. This means that there is no need to adjust the IBS for censoring.

The penalized log-likelihood function for the parameter vector $\boldsymbol{\theta}$ given data $\mathcal{D}$ corresponding to model $M^{(S)}$ is as described by Equation (5.2), with the following factors:

- $f(\mathbf{y}_i \mid b_{i0}) = (2\pi\sigma_\varepsilon^2)^{-n_i/2} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2} \|\mathbf{y}_i - \mathbf{1}_{n_i}(\beta_0 + b_{i0}) - X_i(t)\boldsymbol{\beta}\|^2\right\}$
- $f(t_i, \delta_i \mid b_{i0}) = \left[\rho\kappa(\rho t_i)^{\kappa-1} \exp\left\{\mathbf{w}_i^\top\boldsymbol{\gamma} + \eta b_{i0}\right\}\right]^{\delta_i} e^{-(\rho t_i)^\kappa \exp\left\{\mathbf{w}_i^\top\boldsymbol{\gamma} + \eta b_{i0}\right\}}$
- $f(b_{i0}) = (2\pi\sigma_b^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma_b^2} b_{i0}^2\right\}$

The algorithm was tested for a 5-fold cross-validation design with values of the penalties between 10^{-4} and 10^3, in increasing in powers of 10, so $\log_{10}(\lambda_L), \log_{10}(\lambda_S) = \{3, 2, 1, 0, -1, -2, -3, -4\}$

5.3.2 Optimization of prediction accuracy

Let $\boldsymbol{\theta}$ be the set of parameters of a joint model for longituidnal and time-to-event data. The optimization of $\ell(\boldsymbol{\theta}|\mathcal{D})$ was done using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, an iterative quasi-Newton method. The BFGS algorithm is implemented in the general optimization R function `optim()`.

The estimates of the expected value of the longitudinal outcome and the residual survival probabilities in step 2(b)(i) of Algorithm 5.1 are evaluated at $\widehat{b}_{i0}$, and the integral

		Simulation					
	Related	$M_{12}^{(S)}$		$M_{34}^{(S)}$		$M_5^{(S)}$	$M_6^{(S)}$
	variable	1	2	3	4	5	6
Sample size							
n	–	50	250	50	250	250	250
Fixed effects							
Longitudinal							
β_0	1	3	3	3	3	3	3
β_t	time	0.5	0.5	0.5	0.5	0.5	0.5
β_1	w_1	0.5	0.5	0.5	0.5	0.5	0.5
β_2	w_2	0	0	0	0	0	0
β_3	w_3	0	0	0	0	0	0
β_4	w_4	0.5	0.5	0.5	0.5	0.5	0.5
β_5	w_5	0	0	0	0	0	0
β_6	w_6	0	0	0	0	0	0
β_7	w_7	0	0	0	0	1	1
Terminal							
γ_1	w_1	0.1	0.1	0.1	0.1	0.1	0.1
γ_2	w_2	0.1	0.1	0.1	0.1	0.1	0.1
γ_3	w_3	0	0	0	0	0	0
γ_4	w_4	0	0	0	0	0	0
γ_5	w_5	0	0	0	0	0	0
γ_6	w_6	0	0	0	0	0	0
γ_7	w_7	0	0	0.5	0.5	0	0.5
Association							
η	b_{i0}	0.5	0.5	0.5	0.5	0.5	0.5
Variance							
σ_ε^2	$\varepsilon_i(t)$	4	4	4	4	4	4
σ_b^2	b_{i0}	4	4	4	4	4	4
$h_0(t;\kappa,\rho)$							
κ	–	2	2	2	2	2	2
ρ	–	1.5	1.5	1.5	1.5	1.5	1.5
Link function							
$g(b_{i0})$	–	b_{i0}	b_{i0}	b_{i0}	b_{i0}	b_{i0}	b_{i0}

Table 5.2: Parameter values of the simulation study: variable selection to optimize prediction.

to predict the random effect as the mean of the posterior distribution was approximated with the Gauss–Kronrod method (Ziegel, 1987) implemented in the R function `integrate()`.

The results of the simulation experiment are summarized in Figure 5.1, containing the cross-validated MSE and IBS of the six simulated data sets. In the first and third rows of Figure 5.1 we see that, roughly, the MSE decreases monotonically with increasing values of the two tuning parameters, λ_S and λ_S, this is when the constraint on the parameter space is less stringent, although the MSE is clearly more sensitive to λ_L than to λ_S. This suggest that the accuracy of prediction of the longitudinal outcome could be optimized at larger values of the tuning parameters. An exception to this is simulation 2 (the second plot at the top row), where λ_S seems to have little effect and the MSE is minimized as $\lambda_L \to 10$. We should recall at this point that simulations 1 and 2 have the same set of non-zero regression coefficients and they differ only in their sample size. This suggests that with larger samples it might be easier to find the combination of (λ_L, λ_S) at which the MSE is minimized.

The plots corresponding to the IBS in the second and fourth rows of Figure 5.1 also show clearly the accuracy of predictions of the time-to-event outcome are more sensitive to λ_S (on the horizontal axis), being $\lambda_S = 10$ the value of this tuning parameter where the IBS can be minimized. Only in simulation 1 is the IBS smaller at $\lambda_S = 0$.

It is worth noting that there are some regions of the plots where the MSE or the IBS seem to increase suddenly (these regions are identified by the white pixels between colored regions). This might be something that requires further exploration since the likelihood function of joint models can be complicated to optimize due to the large number of parameters to estimate, the integrals of the random effects and the (possibly) non-linear functions of time of the baseline hazards. Our analyses indicate that the optimization algorithm completely converged for all combinations of the two penalties.

Our goal is to optimize simultaneously the accuracy of predictions of both outcomes. It is not possible to give a unique point of (λ_L, λ_S) where both MSE and IBS are optimized, but there is a region of the grid of the two penalties where this occurs. However,

Simulation	$M_s^{(S)}$	$\log_{10}(\lambda_L^*)$	$\log_{10}(\lambda_S^*)$
1	$M_{12}^{(S)}$	3	2
2	$M_{12}^{(S)}$	1	$(2, 3)$
3	$M_{34}^{(S)}$	$(1, 2, 3)$	$(0, 1)$
4	$M_{34}^{(S)}$	$(1, 2, 3)$	1
5	$M_5^{(S)}$	$(2, 3)$	1
6	$M_6^{(S)}$	$(2, 3)$	1

Table 5.3: Solution of the cross-validation experiment. Values of $(\log_{10}(\lambda_L), \log_{10}(\lambda_L))$ that minimize simultaneously the MSE and the IBS in each of the six simulated data sets.

this region is not the same for all simulations. The best $(\log_{10}(\lambda_L), \log_{10}(\lambda_S))$ regions for each simulation are shown in Table 5.3.

Recall that the models of simulations $4, 5$ and 6 the covariate w_7 has non-zero regression coefficient in at least one of the submodels, and that w_7 and w_6 are highly correlated. It seems to be more complicated to simultaneously optimize the accuracy of predictions of the two outcomes in the presence of highly correlated covariates since the regions of (λ_L, λ_S) that minimizes the MSE does not overlap the region (λ_L, λ_S) that minimizes the IBS. This is more evident in simulations 5 and 6, where the linear mixed submodel has a non-zero regression coefficient for w_7. The region $(\log_{10}(\lambda_L^*) > 1, \log_{10}(\lambda_S^*) = 1)$ that we identified as the solution for simulations 5 and 6 in Table 5.3 is prioritizing the accuracy of prediction of the longitudinal outcome since this region compromises a little accuracy of the prediction of the time-to-event outcome in order to include the region where the MSE is minimal. If on the other hand we want to prioritize the accuracy of prediction of the time-to-event outcome, then the solution would be $(\log_{10}(\lambda_L^*) = 0, \log_{10}(\lambda_S^*) > 1)$ for simulation 5 and $(\log_{10}(\lambda_L^*) = 0, \log_{10}(\lambda_S^*) = 1)$ for simulation 6. It might be important to investigate further this behavior with a larger set of correlated covariates.

In Section 5.4 we apply this strategy to the CARE75+ data set.

In 5 out of our 6 simulation scenarios, the solution, $(\lambda_L^*, \lambda_S^*)$, that optimizes prediction is not a unique pair, but rather a region comprised by a set of values of these two penalties. As a secondary criterion to help in identifying the optimal pair (λ_L, λ_S) for each simulation, we looked into the fixed effects regression coefficients estimated from optimizing $\ell_\lambda(\boldsymbol{\theta} \mid \mathcal{D})$ for the specific $(\lambda_L^*, \lambda_S^*)$ combinations, and compared them against their corresponding values chosen for the simulation experiment (shown in Table 6.3). More specifically, we assessed whether the regression coefficients estimates are non-zero whenever their true value in the simulation experiment are non-zero, and zero otherwise. We follow this approach as a means to untie the $(\lambda_L^*, \lambda_S^*)$ and it is discussed in Section 5.3.3, i.e. optimizing prediction is the leading criterion, and variable selection accuracy the secondary one. Note that we could have done it the other way around by prioritizing variable selection accuracy and rendering prediction optimization as the secondary criterion.

5.3.3 Variable selection assessment

As mentioned in the Introduction of this chapter, our aim is to find the subset of covariates that optimizes predictions of a joint model for longitudinal and time-to-event outcomes, while getting an interpretable model. In this section we explore the extent in which our proposed strategy to optimize predictions estimates as non-zeroes the fixed effects regression coefficients whose values in the simulation design are actually non-zero, and estimates as zero those coefficients whose value is actually zero.

We refer to the values of the regression coefficients in the simulation design in Table 6.3 as the *true* values, and their estimates, $\widehat{\beta}_t$, $\widehat{\beta}_k$ and $\widehat{\gamma}_k$, $k = 1, \ldots, 7$ as the *estimated* values. Additionally, we considered the joint model as a whole, so in this section we denote by $\boldsymbol{\theta}$ the q-vector of regression coefficients of the two submodels.

$$\boldsymbol{\theta}^\top = (\theta_1, \theta_2, \ldots, \theta_q) = (\beta_t, \beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_p). \tag{5.3}$$

In our simulation experiment, $q = 2p + 1$ and $p = 7$.

We assessed the agreement between the true and estimated values by constructing a binary classifier for the regression coefficients and compare the output of this classifier

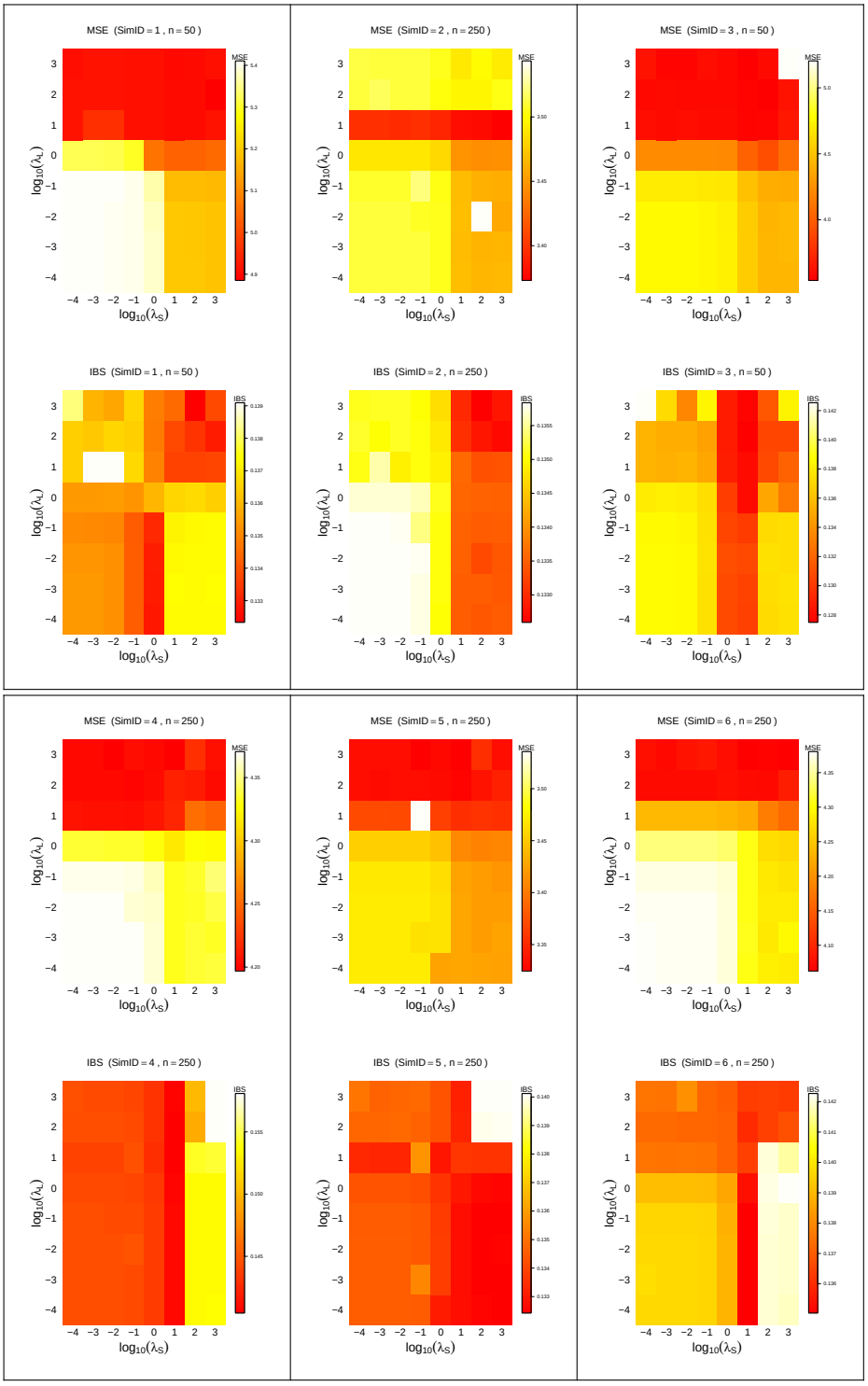


Figure 5.1: Cross-validated MSE (odd rows) and IBS (even rows) of the of the six simulated data sets plotted against the two penalties: λ_S on the horizontal axis and λ_L on the vertical axis ($\log_{10}$ scale).

applied to the true and the estimated values of the regression coefficients. This comparison is done in a *confusion matrix*, a 2×2 table whose entries are the counts of the four possible combinations of the output of a binary classifier applied to the true and estimated values as shown on the left side of Table 5.4, where

$$\text{T0} := \#\ (\text{True} = 0\ \&\ \text{Estimate} = 0)$$
$$\text{F0} := \#\ (\text{True} = 1\ \&\ \text{Estimate} = 0)$$
$$\text{F1} := \#\ (\text{True} = 0\ \&\ \text{Estimate} = 1)$$
$$\text{T1} := \#\ (\text{True} = 1\ \&\ \text{Estimate} = 1)$$

The diagonal elements of the confusion matrix indicate the agreement between the true and estimated values, while the off-diagonals the discordance. Typically the four cells of a confusion matrix are known as *True Positive* (correct positive prediction), *False Positive* (incorrect positive prediction), *False Negative* (incorrect negative prediction) and *True Negative* (correct negative prediction) (Fawcett, 2006) .

There are several metrics that intend to assess the performance of a binary classifier, the right side of Table 5.4 show the most common. In the context of our simulation study these metrics must be read as follows:

- Error: The proportion of discordance between true and estimated values among the total number of estimated coefficients.
- Accuracy: The proportion of agreement between true and estimated values among the total number of estimated coefficients.
- Sensitivity: The proportion of actual positives that are correctly estimated as such. It is also known as the *True Positive Rate*.
- Specificity: The proportion of actual negatives that are correctly estimated as such. It is also known as the *True Negative Rate*.
- Precision: The proportion of estimated positives that are actually positive.
- False Positive Rate: The proportion of actual negatives that are incorrectly estimated as positive.

True	Estimate 0	1	Sum
0	T0	F1	N_0^T
1	F0	T1	N_1^T
Sum	N_0^E	N_1^E	N

Measure	Definition
Error	$(\text{F0+F1})/N$
Accuracy	$(\text{T0+T1})/N$
Sensitivity (TPR)	$\text{T1}/N_1^T$
Specificity (TNR)	$\text{T0}/N_0^T$
Precision	$\text{T1}/N_1^E$
False Positive Rate	$1 - \text{Specificity}$
False Negative Rate	$1 - \text{Sensitivity}$

Table 5.4: Confusion matrix and common performance metrics calculated from it.

- False Negative Rate: The proportion of actual positives that are incorrectly estimated as negative.

The range of values the metrics in Table 5.4 is $[0, 1]$. Error and Accuracy are complements of each other $(\text{Error} = 1 - \text{Accuracy})$, and the False Positive (Negative) Rate is the complement of Specificity (Sensitivity). Accuracy can also be expressed as a weighted average of Sensitivity and Specificity with weights equal to the sample prevalence (i.e. $\Pr(\text{True} = 1)$) and its complement $(\Pr(\text{True} = 0))$,

$$\text{Accuracy} = \text{Prevalence} \times \text{Sensitivity} + (1 - \text{Prevalence}) \times \text{Specificity},$$

where $\text{Prevalence} = N_1^T/N$.

When one of the dimensions of a confusion matrix represents a hypothesis system (null and alternative) and the other one the prediction based on some test, the quantities $1 -$ Sensitivity and $1 -$ Specificiy are analogous to the Type I and Type II errors, respectively, in the context hypothesis testing (Zhou *et al.*, 2009).

In order to assess the performance of our variable selection strategy described in Algorithm 5.1 in terms of the agreement between estimates and their true values, we defined a binary classifier. For each (λ_L, λ_S) combination and the K folds of the cross-validation design (a total of 320 combinations per simulation) we applied the binary classifier to each fixed effect regression coefficient estimate that result from optimizing the penalized log-likelihood defined in Equation (5.2) and its true value and produced a confusion matrix. We assessed the level of agreement between estimates and their

true values with the metrics described in Table 5.4. We averaged each performance metric over the K folds, so we analyzed the cross-validated performance metrics. We repeated this process in each simulated data set.

The binary classifier we used is the indicator function $\mathbb{1}\left(|x| > 0\right)$ that returns 1 if $x \neq 0$ and 0 otherwise. Let

$$\mathrm{T}_j = \mathbb{1}\left(|\theta_j| > 0\right) \quad \text{and} \quad \mathrm{P}_j = \mathbb{1}(|\widehat{\theta}_j| > 0)$$

denote the value of the binary classifier applied to the true value of the regression coefficient θ_j and its estimate, $\widehat{\theta}_j$, respectively. Since some estimates are very small but not exactly equal to zero, we chose $\epsilon = 10^{-5}$ as the threshold to consider $\widehat{\theta}_j = 0$.

Define the True Negatives, False Negatives, False Positives and True Positives as in Equations (5.4a)–(5.4a).

$$\mathrm{T0} := \sum_{j=1}^{q} \mathbb{1}(\mathrm{T}_j = 0 \;\&\; \mathrm{P}_j = 0), \tag{5.4a}$$

$$\mathrm{F0} := \sum_{j=1}^{q} \mathbb{1}(\mathrm{T}_j = 1 \;\&\; \mathrm{P}_j = 0), \tag{5.4b}$$

$$\mathrm{F1} := \sum_{j=1}^{q} \mathbb{1}(\mathrm{T}_j = 0 \;\&\; \mathrm{P}_j = 1), \tag{5.4c}$$

$$\mathrm{T1} := \sum_{j=1}^{q} \mathbb{1}(\mathrm{T}_j = 1 \;\&\; \mathrm{P}_j = 1). \tag{5.4d}$$

Since our interest is to explore the extent in which our proposed strategy developed to optimize predictions estimates as non-zeroes the fixed effects regression coefficients that are actually non-zeroes in the simulation design, and as zeroes those that are actually zeroes, we based our analyses of this section primarily on Sensitivity and Specificity, and on Accuracy because it summarizes these two metrics. Figures 5.2 and 5.3 show the Accuracy, Sensitivity and Specificity of applying the binary classifier to the six simulations. Appendix B.2 contains plots of the complete set of six metrics for the binary classifier analyzed in this section and other two classifiers discussed in Section 5.5. These plots show a roughly monotonic behavior of the three metrics with respect

		Prediction Accuracy		Variable Selection				
Simulation	$M_s^{(S)}$	$\log_{10}(\lambda_L^*)$	$\log_{10}(\lambda_S^*)$	$\log_{10}(\widetilde{\lambda}_L^*)$	$\log_{10}(\widetilde{\lambda}_S^*)$	Acc	Sen	Spe
1	$M_{12}^{(S)}$	3	2	3	2	0.48	0.28	0.58
2	$M_{12}^{(S)}$	1	(2,3)	1	3	0.57	0.72	0.50
3	$M_{34}^{(S)}$	(1,2,3)	(0,1)	3	1	0.55	0.50	0.58
4	$M_{34}^{(S)}$	(1,2,3)	1	3	1	0.64	0.77	0.56
5	$M_{5}^{(S)}$	(2,3)	1	2	1	0.44	0.87	0.13
6	$M_{6}^{(S)}$	(2,3)	(1,2,3)	3	3	0.64	0.46	0.80

Table 5.5: Classifier 1. Values of $(\log_{10}(\lambda_L), \log_{10}(\lambda_L))$ that optimize prediction accuracy, Sensitivity and Specificity in the variable selection process. Acc $=$ Accuracy, Sen $=$ Sensitivity, Spe $=$ Specificity.

to the two penalties (shown in $\log_{10}$ scale), with bigger changes occurring in $\lambda_L \geq 10$ and $\lambda_S \geq 10$, which is the region of more accurate predictions, as discussed in Section 5.3.2. Table 5.5 complements Table 5.3 by including additional columns for Accuracy, Sensitivity and Specificity corresponding to $(\lambda_L^*, \lambda_S^*)$, i.e. the combination of the two penalties where predictions are closer to being optimal. In all the simulations where $(\lambda_L^*, \lambda_S^*)$ is not a unique combination, we identified as $(\widetilde{\lambda}_L^*, \widetilde{\lambda}_S^*)$ the combination within $(\lambda_L^*, \lambda_S^*)$ with the highest Accuracy and considered it as the solution.

Finally, in order to obtain a definitive model that optimizes prediction, we looked into the estimates K folds of the cross validation design at $(\widetilde{\lambda}_L^*, \widetilde{\lambda}_S^*)$, computed $\mathrm{P}_j,\ j = 1, \ldots, q$ and determined that θ_j should be in the final model if $\mathrm{P}_j = 1$ in at least 50% of the K folds. Table 5.6 identifies the fixed effects covariates that should be included in the final joint model in order to optimize prediction of both outcomes and shows the estimates of all the parameters and regression coefficients of the final model of each of the six simulations.

In all six simulations, the model that optimizes prediction has relatively low Accuracy, ranging from 0.44 to 0.57. Simulations 1 and 3, which have the smaller sample sizes, have lower Sensitivity and higher Specificity. In both simulations 1 and 3, the definitive model that optimizes prediction accuracy should include no fixed effects in the

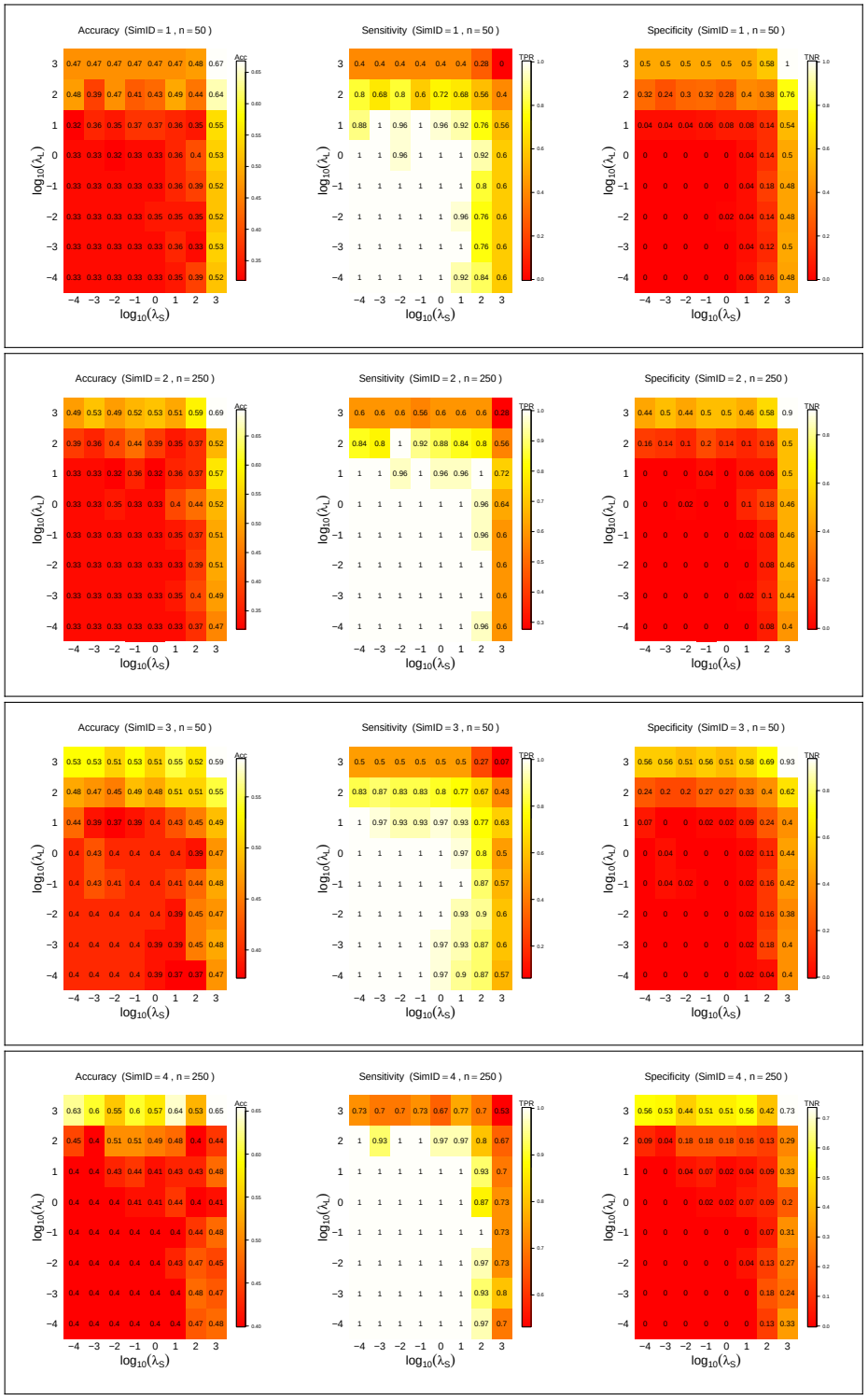


Figure 5.2: Classifier 1. Performance metrics of the variable selection process. Top to bottom: Simulations 1 to 4. Left to right: Accuracy, Sensitivity and Specificity.

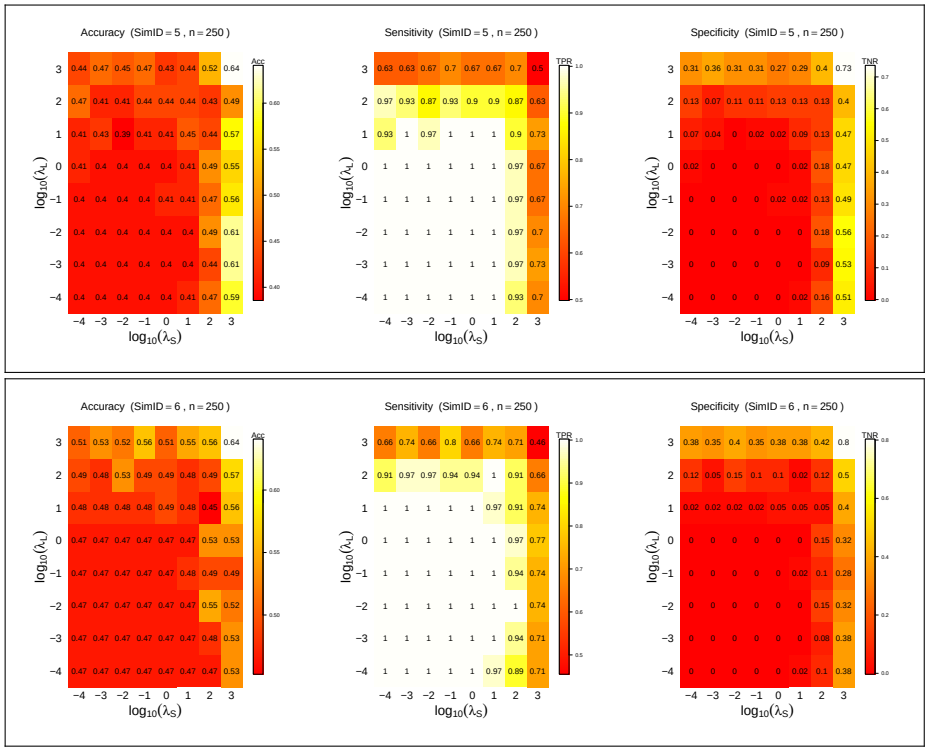


Figure 5.3: Classifier 1. Performance metrics of the variable selection process. Top to bottom: Simulations 5 and 6. Left to right: Accuracy, Sensitivity and Specificity.

		Simulation					
	Related	$M_{12}^{(S)}$		$M_{34}^{(S)}$		$M_5^{(S)}$	$M_6^{(S)}$
	variable	1	2	3	4	5	6
Sample size							
n	–	50	250	50	250	250	250
Fixed effects							
Longitudinal							
β_0	1	2.012	3.383	1.936	3.031	3.127	3.601
β_t	time	0.159	0.582	0.045	0.426	0.589	0.526
β_1	w_1	–	0.090	–	–	0.401	–
β_2	w_2	–	0.229	–	–	0.247	–
β_3	w_3	–	−0.280	–	–	−0.242	–
β_4	w_4	–	0.262	–	0.877	0.428	–
β_5	w_5	–	0.049	–	–	0.056	–
β_6	w_6	–	0.031	–	–	0.294	0.048
β_7	w_7	–	−0.093	–	–	0.611	0.373
Terminal							
γ_1	w_1	0.463	–	−0.218	0.046	0.335	–
γ_2	w_2	−0.565	0.403	0.177	0.338	0.415	–
γ_3	w_3	−1.193	–	0.076	0.096	0.021	–
γ_4	w_4	−0.180	–	−0.409	−0.063	0.170	–
γ_5	w_5	−0.289	–	−0.166	0.036	0.004	–
γ_6	w_6	0.122	–	−0.016	−0.263	0.280	–
γ_7	w_7	−0.145	–	0.353	0.796	−0.310	–
Association							
η	b_{i0}	0.550	0.555	0.522	0.660	0.560	0.574
Variance							
σ_ε^2	$\varepsilon_i(t)$	4.439	3.782	3.680	4.312	3.781	3.970
σ_b^2	b_{i0}	5.656	4.100	2.331	3.099	4.088	4.830
$h_0(t;\kappa,\rho)$							
κ	shape	1.076	0.994	0.968	1.029	1.000	0.976
ρ	rate	1.876	1.220	1.504	1.223	1.074	1.291

Table 5.6: Final model that optimizes prediction for each simulation. Parameter estimates of the joint model that includes only the covariates indicated by our variable selection strategy in order to optimize prediction. "–" indicates that the covariate it is not included in the final model.

submodel of the quantitative outcome submodel, not even time, and all $w_1, \ldots, w_7$ in the time-to-event submodel. It is worth noting that even with w_6 and w_7 being highly correlated, the definite model for simulation 3 should include both of these covariates.

The final model of some of the simulations with larger sample size (2,4,and 5) have higher Sensitivity relative to the small sample simulations. The exception is simulation 6, whose simulation model has regression coefficients $\beta_7 = 1$ and $\gamma_7 = 0.5$ for the covariate w_7 and this covariate is highly correlated with w_6. As in the smaller size simulation 3, it is interesting that the definite model of the larger size simulations should include both highly correlated covariates, w_6 and w_7, in the quantitative outcome submodel in order to optimize prediction.

The Sensitivity of the model for simulation 5 is notably high and its Specificity notably low (0.13), which means that very few covariates whose regression coefficient is actually non-zero and, as shown in Table 5.7, the definitive model should include all covariates in both submodels. In the simulation model $M_5^{(S)}$ the regression coefficients associated to w_7 are $\beta_7 = 1$ and $\gamma_7 = 0.5$, and $\text{cor}(w_6, w_7) = 0.95$.

A consistent result across the six simulations is that the definite model that optimizes prediction would always include the two correlated covariates, w_6 and w_7 at least in one of the submodels. This suggests that a submodel with correlated covariates improves prediction, regardless of whether or not these covariates have an effect on the outcome. In Section 5.4 we analyze the results of applying our variable selection strategy to the CARE75+ data.

5.4 Applications to the CARE75+ data

Figure 5.4 shows the cross-validated MSE and IBS that result from applying the Algorithm 5.1 to the CARE75+ data set. Here we can notice that the region of (λ_L, λ_S) that minimizes the MSE does not overlap the region that minimizes the IBS. For this real data set we are not able to compute the performance metrics of the confusion matrix as we did for in the simulation experiments because we obviously don't know the true

	Covariate	Estimate	SE	z-value	p-value
Fixed effects					
Longitudinal					
β_0	1	7.268	0.338	21.477	< 0.001
$\beta_{\texttt{time}}$	`years`	0.104	0.112	0.923	0.356
$\beta_{\texttt{sex}}$	`sex`	-0.160	0.253	-0.630	0.529
$\beta_{\texttt{eth}}$	`ethinicity`	-2.328	0.380	-6.125	< 0.001
$\beta_{\texttt{mar}}$	`marital`	-0.907	0.245	-3.707	< 0.001
$\beta_{\texttt{edu}}$	`education`	-1.173	0.246	-4.765	< 0.001
$\beta_{\texttt{smo}}$	`smoke`	-0.175	0.210	-0.834	0.405
$\beta_{\texttt{alc}}$	`alcohol`	-0.971	0.250	-3.885	< 0.001
$\beta_{\texttt{fal}}$	`falls`	0.207	0.038	5.491	< 0.001
Terminal					
$-$	$-$	$-$	$-$	$-$	$-$
Association					
η	b_{i0}	0.276	0.110	2.513	0.012
Variance					
σ^2_ε	$\varepsilon_i(t)$	2.243	$-$	$-$	$-$
σ^2_b	b_{i0}	2.735	$-$	$-$	$-$
$h_0(t;\kappa,\rho)$					
κ	shape	0.864	$-$	$-$	$-$
ρ	rate	0.006	$-$	$-$	$-$

Table 5.7: Joint model that optimizes prediction of frailty and mortality in the CARE75+ data set.

values of the regression coefficients and parameters of the model. Thus, we need a compromise between MSE and IBS.

A reasonable solution would be $(\lambda_L^*, \lambda_S^*) = (0.0001, 1000)$, resulting in $\text{MSE} = 6.26$ and $\text{IBS} = 42.15$. This solution optimizes MSE and the level of IBS is very close to the optimal level, so it satisfies our interest of optimizing simultaneously the predictions for frailty and for mortality. The final model for the CARE75+ data should include `sex`, `ethnicity`, `marital`, `education`, `smoke`, `alcohol`, `fallscount`, and `years` in the submodel for frailty and no covariates in the submodel for mortality. The parameter estimates of this model are in Table 5.7.

It is important to note in Figure 5.4 the negligible changes in the IBS even on very large differences in λ_L and λ_S. Choodari-Oskooei *et al.* (2012b) show some evidence that measures of predictive accuracy, like the Brier score, are generally lower that explained variation and explained randomness measures, because they capture the uncertainty in a binary outcome (event and non-events) accounted by a model rather than capturing the uncertainty about the survival time itself. This behavior is similar in logistic regression. Additionally, it has been reported, for instance by Benedetti (2010), that the Bier score becomes inadequate for rare (infrequent) events because it does not correctly assess the ability of the model to discriminate events from non-events.

In the hypothetical scenario that our priority was to optimize predictions for mortality, then the solution would be $(\lambda_L^*, \lambda_S^*) = (100, 10)$ resulting in $\text{MSE} = 10.81$ and $\text{IBS} = 41.84$.

We would like to display graphically the regression coefficient estimates as function of the two penalties to see which covariates switch in or out of a submodel (i.e. associated regression coefficient shrinking towards zero) as the values of the two penalties vary. The fact that we deal with two penalties makes it difficult to visualize simultaneously the surface of the 15 regression coefficients estimates. As an example, we include in Appendix B.2.2 three different graphical representations (surface, heat map and scatter plot) of the regression coefficients behavior with respect to the values of (λ_L, λ_S) for simulation 6, where we observe that most of the coefficients decrease monotonically with respect to both penalties but some of them respond in this way only to the penalty of the submodel they belong to. This would require further investigation,

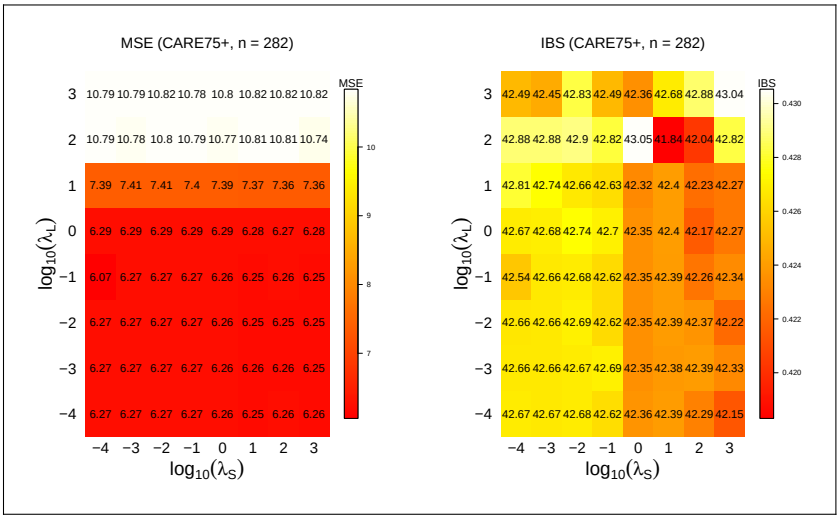


Figure 5.4: Cross-validated MSE and IBS of the of the CARE75+ data set plotted against the two penalties: λ_S on the horizontal axis and λ_L on the vertical axis ($\log_{10}$ scale).

in particular, finding a better tool to assess graphically all the regression coefficients simultaneously.

5.5 Discussion

Our strategy is useful to select fixed effects covariates for a joint model for longitudinal and time-to-event data that optimize the accuracy of prediction of both the quantitative and the time-to-event outcomes. This is done by penalizing the log-likelihood function with separate penalties for the fixed effects regression coefficients of each submodel and cross-validation. However, it is not always possible to optimize simultaneously the two outcomes. The simulation study suggests that with highly correlated covariates, the region of the hyperpameters (the two penalties) where the MSE is optimal might not overlap with the region of optimal IBS. In such a case, our strategy allows to choose among values within a small region of the two penalties that require a compromise between MSE and IBS depending on which outcome is the priority. According to our simulation study it is possible to find a solution with very small compromise between the prediction accuracy of the two outcomes. The fact that highly correlated covariates

can be in the final model that optimizes prediction raises the question if our strategy is overfitting. In principle, our strategy should not yield an overfitted model because this problem is being addressed by regularizing the model via shrinkage methods and cross-validation. This might require further investigation with more simulations and larger sample sizes and more correlated covariates.

We assessed the accuracy of prediction with a separate metric for each outcome: MSE for the longitudinal and IBS for the time-to-event. In the future we would like to investigate with the possibility of defining an overall measure of prediction accuracy for the two outcomes of the form $\varphi = \varphi(\text{MSE,IBS})$, and assess if this kind of measures can be extended to an arbitrary number of outcomes in a joint model.

We assessed the variable selection abilities of our strategy in simulation studies in terms of Accuracy, Sensitivity and Specificity that result from comparing the regression coefficient estimates against their true values under a binary classifier. The values of these metrics are not as high as we have seen in simpler cases of single-outcome regression models or multivariate models for normally distributed outcomes. Nonetheless, this is of secondary interest for our proposed strategy, the principal criterion being to optimize prediction. We constructed two additional binary classifiers that impose additional restrictions to the binary classifier used in our simulation study. Even though we obtained similar results they seem too ambitious for the joint modelling context and we would rather explore how they perform in simpler contexts, like single-outcome models. These two binary classifiers are discussed briefly in the Section 5.5.1.

Our main interest and primary criterion for model selection is optimizing prediction, and we used as secondary criterion the Accuracy of variable selection relative to the true model when it was difficult to determine which (λ_L, λ_S) combination gave the best prediction. We should emphasize that this can very well be done the other way around: by considering variable selection accuracy as the primary criterion and prediction optimization as the secondary one. To illustrate this point, consider simulation 4 in the fourth row of Figure 5.2. Even though we find in $(\log_{10}(\lambda_L), \log_{10}(\lambda_S)) = (3, 1)$ the highest Accuracy for variable selection with highest Specificity, it also has the lowest Sensitivity. This point minimizes the MSE, but it renders the largest IBS, which is a different solution to the one found earlier. We get the same solution by Sacrificing a little Accuracy to gain some Sensitivity, where the points $\log_{10}(\lambda_L) = 3$ and

$\log_{10}(\lambda_S) = (-4, 1, 3)$ become very good candidates. We can confirm at the bottom left of Figure 5.1 that the point $(\log_{10}(\widetilde{\lambda}_L^*), \log_{10}(\widetilde{\lambda}_S^*)) = (3, 1)$ minimize both MSE and IBS, resulting in optimal prediction of both outcomes.

Su *et al.* (2016) proposed a sparse estimation method for Cox models by optimizing an approximated information criterion. The method approximates the ℓ_0-norm with a continuous function; it mimics the best subset selection using a penalized likelihood approach yet with no need of a tuning parameter. Han *et al.* (2020) studied the use of this selection method in joint models of recurrent and terminal events. We can explore using this penalty for joint models of longitudinal and recurrent and terminal events.

5.5.1 Extensions and future work

A very important, but often ignored problem of data-driven variable selection is model stability, that is the robustness of the selected model to small perturbations of the data set (Heinze *et al.*, 2018). Bootstrap resampling with replacement or subsampling without replacement are valuable tools to investigate and quantify model stability of selected models. The basic idea is to draw B resamples from the original data set and to repeat variable selection in each of the resmaples. Important types of quantities that this approach can provide are:

- bootstrap inclusion frequencies to quantify how likely a covariate is selected,
- sampling distributions of regression coefficients,
- model selection frequencies to quantify how likely a particular set of covariates is to be in the model.

Classifier 2

In the second classifier, we add the condition to classifier 1 that the estimate, say $\widehat{\theta}_j$, and is its true value, θ_j, must agree also in their sign, i.e. if both are positive or both negative. Let

$$\mathrm{T}_j = \mathrm{sign}(\theta_j) \qquad \text{and} \qquad \mathrm{P}_j = \mathrm{sign}(\widehat{\theta}_j),$$

where the sign function is defined as

$$\text{sign}(x) := \begin{cases} 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0, \\ +1 & \text{if } x > 0. \end{cases}$$

Here we take $\epsilon = 10^{-5}$ as threshold to consider $\widehat{\theta}_j = 0$, so $\text{sign}(\widehat{\theta}_j) = 0$ if $|\widehat{\theta}_j| \leq \epsilon$.

With the functions T_j and P_j, we define the True Negatives, False Negatives, False Positives and True Positives as

$$\begin{aligned}
\text{T0} &:= \sum_{j=1}^{q} \mathbb{1}(\text{T}_j = 0 \;\&\; \text{P}_j = 0), \\
\text{F0} &:= \sum_{j=1}^{q} \mathbb{1}\left([\text{T}_j = +1 \;\&\; \text{P}_j \leq 0] \text{ or } [\text{T}_j = -1 \;\&\; \text{P}_j \geq 0]\right), \\
\text{F1} &:= \sum_{j=1}^{q} \mathbb{1}\left([\text{T}_j = 0 \;\&\; \text{P}_j = +1] \text{ or } [\text{T}_j = 0 \;\&\; \text{P}_j = -1]\right), \\
\text{T1} &:= \sum_{j=1}^{q} \mathbb{1}\left([\text{T}_j = +1 \;\&\; \text{P}_j = +1] \text{ or } [\text{T}_j = -1 \;\&\; \text{P}_j = -1]\right).
\end{aligned}$$

Just as in the binary classifier defined in Section 5.3.3, T0 denotes the number of coincidences of true values and estimates being equal to zero. In contrast, here T1 denotes the number of coincidences in the sign (negative or positive) between true values and estimates. F0 is the count of coefficients whose value is actually non-zero that are estimated either as zero or as non-zero but with the opposite sign. Finally, F1 is the count of coefficients whose value is actually zero but estimated as non-zero (either positive or negative). Table 5.8 shows the results of classifier 2.

Classifier 3

The third classifier is based on the interval estimate of the regression coefficients, call it $\widehat{\theta}_{jI} = (\widehat{\theta}_{jL}, \widehat{\theta}_{jU}) = \widehat{\theta}_j \pm \alpha\widehat{\text{se}}(\widehat{\theta}_j)$, $\alpha \in \mathbb{R}^+$. Here we assessed whether or not $\theta_j \in \widehat{\theta}_{jI}$

Simulation	$M_s^{(S)}$	Prediction Accuracy		Variable Selection				
		$\log_{10}(\lambda_L^*)$	$\log_{10}(\lambda_S^*)$	$\log_{10}(\lambda_L)$	$\log_{10}(\lambda_S)$	Acc	Sen	Spe
1	$M_{12}^{(S)}$	3	2	3	2	0.45	0.20	0.58
2	$M_{12}^{(S)}$	1	(2,3)	1	3	0.55	0.64	0.50
3	$M_{34}^{(S)}$	(1,2,3)	(0,1)	3	1	0.49	0.37	0.58
4	$M_{34}^{(S)}$	(1,2,3)	1	3	1	0.64	0.77	0.56
5	$M_5^{(S)}$	(2,3)	1	2	1	0.41	0.83	0.13
6	$M_6^{(S)}$	(2,3)	(1,2,3)	2	3	0.57	0.66	0.50

Table 5.8: Classifier 2. Values of $(\log_{10}(\lambda_L), \log_{10}(\lambda_L))$ that optimize prediction accuracy, Sensitivity and Specificity in the variable selection process.

and if both ends of $\widehat{\theta}_{jI}$ are positive (or negative) and agree with the the sign of its true value, θ_j.

We don't have reliable standard error estimates because some of the variances based on the Hessian matrix that results from `optim()` are negative. Here, we used instead the standard deviation of the K folds of the cross-validation design. This is $\widehat{\text{var}}(\widehat{\theta}) = K^{-1}\sum_{k=1}^{K}\left(\widehat{\theta}_k - \overline{\widehat{\theta}}\right)^2$, with $\widehat{\text{se}}(\widehat{\theta}) = (\widehat{\text{var}}(\widehat{\theta}))^{1/2}$ Due to the asymptotic normal distribution of Maximum Likelihood Estimates, by choosing $\alpha = 3$ we would expect the parameters' interval estimates to have approximate nominal coverage $> 99\%$.

Define

$$\text{T}_j = \text{sign}(\theta_j) \qquad \text{and} \qquad \text{P}_j = \begin{cases} 0 & \text{if } 0 \in \widehat{\theta}_{jI}, \\ -1 & \text{if } \text{sign}(\widehat{\theta}_{jL}) = \text{sign}(\widehat{\theta}_{jU}) = -1, \\ +1 & \text{if } \text{sign}(\widehat{\theta}_{jL}) = \text{sign}(\widehat{\theta}_{jU}) = +1. \end{cases}$$

Here T0, F0, F1 and T1 are defined exactly the same way as in classifier 2. Table 5.9 shows the results of classifier 3. In five out of the six scenarios the best combination of penalties yield Sensitivity > 0.5, and in all six scenarios Specificity ≤ 0.4. However, we consider that this classifier still needs to be investigated further in simpler models, for instance the marginal LMM and Cox models, because even when it is too strin-

gent for an automatic variable selection method it is the classifier that results in larger Sensitivity.

		Prediction Accuracy		Variable Selection				
Simulation	$M_s^{(S)}$	$\log_{10}(\lambda_L^*)$	$\log_{10}(\lambda_S^*)$	$\log_{10}(\lambda_L^*)$	$\log_{10}(\lambda_S^*)$	Acc	Sen	Spe
1	$M_{12}^{(S)}$	3	2	3	2	0.29	0.60	0.14
2	$M_{12}^{(S)}$	1	(2,3)	1	3	0.48	0.88	0.32
3	$M_{34}^{(S)}$	(1,2,3)	(0,1)	3	0	0.39	0.36	0.40
4	$M_{34}^{(S)}$	(1,2,3)	1	1	1	0.52	0.84	0.36
5	$M_5^{(S)}$	(2,3)	1	3	1	0.41	0.60	0.32
6	$M_6^{(S)}$	(2,3)	1	3	1	0.37	0.64	0.24

Table 5.9: Classifier 3. Values of $(\log_{10}(\lambda_L), \log_{10}(\lambda_L))$ that optimize prediction accuracy, Sensitivity and Specificity in the variable selection process.

Chapter 6

Causal inference and joint modelling specification

6.1 Introduction

In this chapter, we explore the relationship between frailty, falls and mortality with the CARE75+ data set addressing confounding. We use DAGs to state upfront our hypotheses about the relationships among all the variables in the CARE75+ data set, identifying all possible confounders to the frailty-falls-mortality relationship based on the causal effect rule (Section 2.5.6), and we fit the joint model that corresponds to the relationships stated on the DAGs.

In Chapter 4 we analyzed the CARE75+ dataset by joint modelling the relationship between frailty, falls and mortality. We fitted, analyzed and discussed the 3-outcome joint model, $\widehat{M}_3^{\text{CARE}}$, and explored with an alternative joint model for frailty and mortality, $\widehat{M}_2^{\text{CARE}}$, assuming falls as exogenous time-varying covariate. The variable selection process was done in two steps. First, stepwise selection was carried out on separate submodels for each outcome using a p-value cutoff of 0.05 to form a *provisional covariate set*. The second step consisted of removing from the provisional covariate set all covariates whose regression coefficient estimate had p-value < 0.05 when attempting to include them the joint model. The exception to the 0.05 significance criterion are

the regression coefficient of `time` in the linear mixed submodel since it is of interest in its own right to know the time effect on frailty, and `ethnicity` in the mortality submodel that was not removed for convergence of the optimization algorithm. Tables 4.4 and 4.5 show the covariates retained in $\widehat{M}_3^{\text{CARE}}$ and $\widehat{M}_2^{\text{CARE}}$ and the parameter estimates.

By fitting $\widehat{M}_3^{\text{CARE}}$ we aimed at describing the relationship between frailty, falls and mortality in the CARE75+ data set by joint modelling these three outcomes of the participants of the study. We made no attempt to draw any causal conclusions. In this chapter we revisit model $\widehat{M}_3^{\text{CARE}}$, analyze it in light of the causal inference framework described in Section 2.5, explain why this model is incorrect for causal inference and reformulate the joint model in line with the causal inference framework using DAGs to estimate the effects of frailty and falls on mortality.

A second aim of this chapter is to study the consequences of model misspecification in joint modelling. In Chapter 4 we fitted model $\widehat{M}_2^{\text{CARE}}$ as an alternative formulation of the frailty-falls-mortality relationship. In Chapter 6 we analyze $\widehat{M}_2^{\text{CARE}}$ under the causal inference framework in parallel to model $\widehat{M}_3^{\text{CARE}}$. We regard model $\widehat{M}_3^{\text{CARE}}$, fitted in Chapter 3, as the leading model all along our analyses of this chapter, and consider the alternative $\widehat{M}_2^{\text{CARE}}$ to understand the consequences of model misspecification. We complement the analysis by conducting a simulation study in which we simulate a series of data sets from two joint models similar to $\widehat{M}_3^{\text{CARE}}$ and $\widehat{M}_2^{\text{CARE}}$ and analyze each data set with both joint models. Our aim by doing so is to assess the extent in which the model parameters and regression coefficients are correctly/wrongly estimated when the data is analyzed with the wrong model.

This chapter has five sections. The first section is this introduction. In the second section we revisit models $\widehat{M}_3^{\text{CARE}}$ and $\widehat{M}_2^{\text{CARE}}$, but this time conducting the variable selection process under the causal inference framework focusing on the effects of frailty and falls on mortality adjusting for confounding. The third section is a simulation study to assess the consequences of misspecifying a joint model on the parameters and regression coefficients estimates. We discuss the main results of the simulation study in the fourth section. Finally, in the fifth section we outline future work and possible extensions to causal inference in the context of joint modelling.

6.2 Joint model for frailty, falls and mortality with the CARE75+ data (Revisited)

The complexity of joint models conveys challenges to statistical modelling. Some are pointed out by Hickey *et al.* (2016, 2018), which can be grouped in methodological, computational and study design issues. Choosing the link to characterize the associations between outcomes is nontrivial, as is choosing the type of submodel for each outcome, the distributional assumptions of the random effects and the form of baseline hazards. Each of these choices require careful consideration, thus fitting a joint model can be a difficult task. Additionally, due to the fact that fitting a joint model implies dealing with multiple regression equations, variable selection becomes more challenging compared to fitting separate marginal models.

As we pointed out in Section 4.5.1, variable selection in statistical modelling is done in different ways depending on the intended use of the fitted model. In order to build a joint model for estimating the effects of frailty and falls on mortality we should first state our causal assumptions about the relationships between the variables in the data set, and we do so by using Directed Acyclic Graphs (DAGs) as explained in Section 2.5.1. In the frailty-falls-mortality relationship, mortality is the endpoint and we are interested in estimating the effects of frailty and falls on mortality, but the possible relationship between frailty and falls is not clear (frailty $\rightarrow$ falls, frailty $\leftarrow$ falls). We need to be explicit about how the other variables of the data set relate to frailty, falls and mortality in order to identify and adjust for all possible confounders of the effects of frailty and falls on mortality. The importance of correctly adjusting for confounding is to avoid (1) inducing spurious associations between frailty, falls and mortality, and (2) blocking the effects of frailty and falls on mortality. These problems are illustrated with simple examples of the basic structures in DAGs in Section 2.5.2.

The criterion to identify the confounders to adjust for is stated in **the causal effect rule** explained in Section 2.5.6. According to the causal effect rule, once a DAG is proposed we should include in each submodel only the *parents* of frailty, falls and mortality respectively, which implies leaving out all variables lying in the paths from frailty and falls to mortality. In this context, the statistical significance of the regression

coefficients of confounders becomes irrelevant, and we should not decide whether or not to keep them in the model based on their p-value. The reason for adjusting for confounders regardless of their p-value is because due to their relationship with frailty, falls and mortality, leaving them out the model could affect the main effects of interest: frailty and falls on mortality.

Equations (6.1a)–(6.1c) describe the model $\widehat{M}_3^{\text{CARE}}$, discussed and fitted in Chapter 4, and the DAG G_3^{CARE} of Figure 6.1 (left) depicts the relationships among variables according to $\widehat{M}_3^{\text{CARE}}$. We omit "$\widehat{\ }$" from the $\widehat{M}_3^{\text{CARE}}$ estimates, but it must be clear that we refer to the fitted model, and for the moment we focus on the mean structure, ignoring the parameters of the baseline hazards and variance components. Recall that model $\widehat{M}_3^{\text{CARE}}$ was built under the assumption that the relationship between frailty, falls and mortality is completely characterized by the random effects b_{i0} and u_i: frailty, falls and mortality are linked by a random intercept (b_{i0}), and additionally, falls and mortality have a common random effect (u_i) acting multiplicatively on the hazards of falls and mortality (refer to Section 4.2 for a rationale of this joint model specification). We refer to G_3^{CARE} as the DAG implied by $\widehat{M}_3^{\text{CARE}}$. Observed variables are represented by a solid rectangle and all unobserved by a dashed circle. The arrows represent our assumptions about the direction of the paths between a pair of variables, with red arrows emphasizing the paths between frailty, falls and mortality. The letters besides the arrows are the regression coefficients associated to the covariates of the joint model, emphasizing in blue the association parameters of the joint model. For instance, `ethnicity` is a covariate in all three submodels, so `eth` is the initial node of three arrows with their corresponding regression coefficient: $\texttt{eth} \xrightarrow[\gamma_{R.\text{eth}}]{} r(t)$, $\texttt{eth} \xrightarrow[\boldsymbol{\beta}]{} m(t)$ and $\texttt{eth} \xrightarrow[\gamma_{T.\text{eth}}]{} h(t)$.

$$
\widehat{M}_3^{\text{CARE}} : \begin{cases} \text{L:} & y_i(t \mid b_{i0}) = (\beta_0 + b_{i0}) + \beta_t t + \mathbf{x}_i^\top(t)\boldsymbol{\beta} + \varepsilon_i(t) & \text{(6.1a)} \\ \text{R:} & r_i(t \mid \mathbf{v}_i) = u_i r_0(t) \exp\{\gamma_{R.\text{eth}}\texttt{eth}_i + \eta_R b_{i0}\} & \text{(6.1b)} \\ \text{T:} & h_i(t \mid \mathbf{v}_i) = u_i^{\alpha} h_0(t) \exp\{\gamma_{T.\text{eth}}\texttt{eth}_i + \eta_T b_{i0}\}, & \text{(6.1c)} \end{cases}
$$

$$\widehat{M}_2^{\text{CARE}}: \begin{cases} \text{L: } y_i(t \mid b_{i0}) = \underbrace{(\beta_0 + b_{i0}) + \mathbf{x}_i^\top(t)\boldsymbol{\beta} + \beta_{\text{falls}} N_i(t)}_{m_i(t)} + \varepsilon_i(t) & \text{(6.2a)} \\ \text{T: } h_i(t \mid b_{i0}) = h_0(t)\exp\{\gamma_{T.\text{eth}}\texttt{eth}_i + \gamma_{T.\text{falls}} N_i(t) + \eta_L m_i(t)\}, & \text{(6.2b)} \end{cases}$$

where

$$\mathbf{x}_i^\top(t) = (\texttt{ethnicity}_i, \texttt{marital}_i, \texttt{education}_i, \texttt{alcohol}_i, \texttt{comorbidities}_i(t))$$
$$\boldsymbol{\beta}^\top = (\beta_{\texttt{eth}}, \beta_{\texttt{mar}}, \beta_{\texttt{edu}}, \beta_{\texttt{alc}}, \beta_{\texttt{com}}).$$
$$\mathbf{v}_i^\top = (b_{i0}, u_i).$$

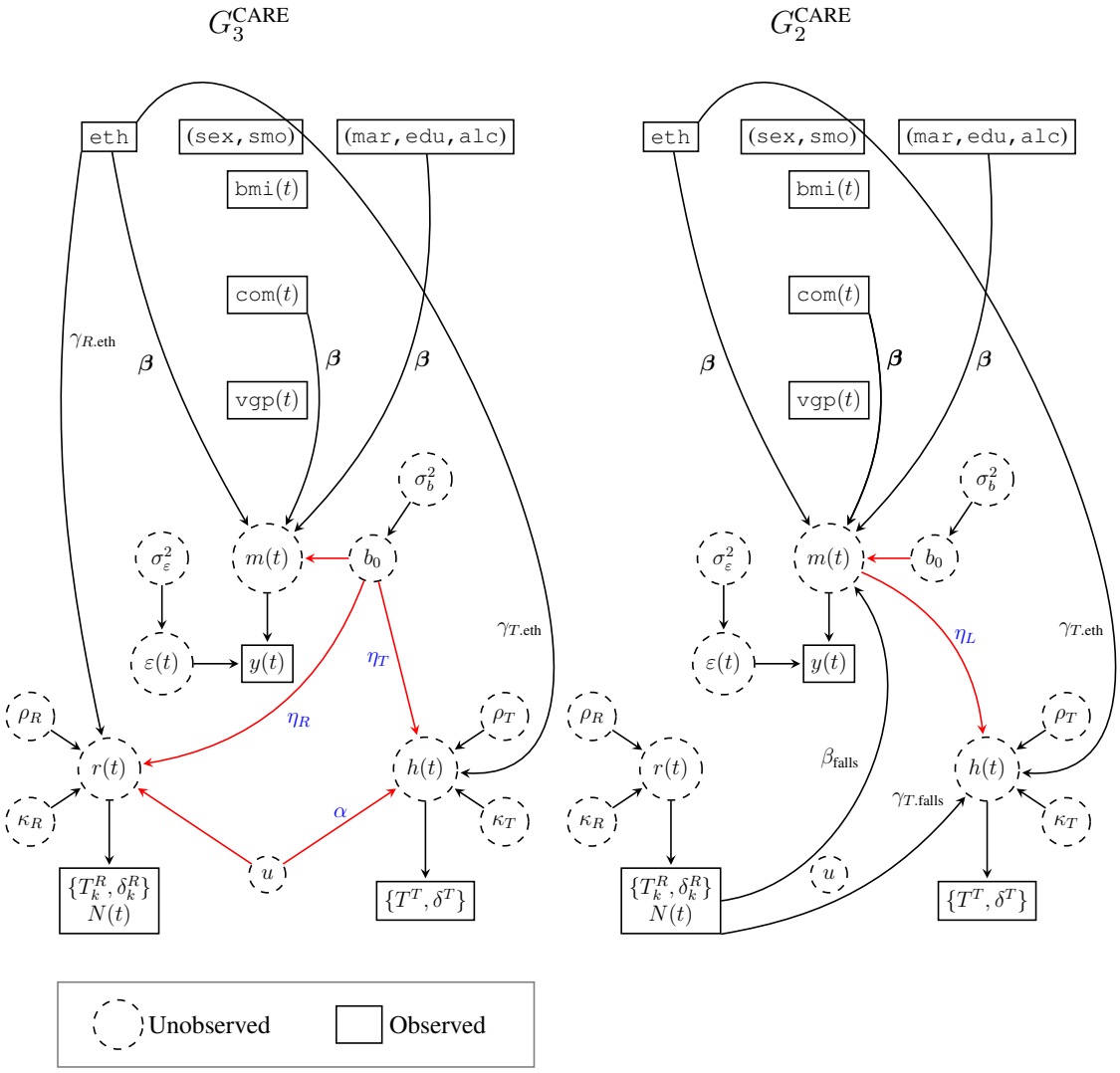
In model $\widehat{M}_3^{\text{CARE}}$, the vector $\mathbf{x}_i(t)$ represents the covariates of subject i whose regression coefficients, denoted by the vector $\boldsymbol{\beta}$, were significant at the level of 0.05 in the submodel of frailty. `ethnicity` was significant for frailty and falls, and no covariate was significant for mortality although we kept `ethnicity` in the mortality submodel. Frailty, falls and mortality are linked by the random effects b_{i0} and u_i, and the strength of these links is quantified by the association parameters η_R, η_T and α.

The problem with $\widehat{M}_3^{\text{CARE}}$ as a model for estimating the effects of frailty and falls on mortality is that most of the covariates can be considered as common causes of frailty, falls and mortality and only some of them are included in the joint model and, as we discussed in Section 2.5.2, failing to adjust for confounding and adjusting for mediating variables might bias the causal effects of interest.

Empirical evidence reported in the literature supports hypotheses about factors like sex, ethnicity, education level, marital status, obesity, and lifestyle habits being correlated with frailty, the risk of falls and the risk of mortality. To mention some examples, (Hao *et al.*, 2019) suggest that age, sex, education levels, BMI, marital status and alcohol intake are confounders of the relationships frailty-mortality and frailty-hospital readmission; while estimating the effect of frailty on the risk of falls and hip/nonspine fractures, Ensrud *et al.* (2007) adjusted for age, health status, medical conditions, functional status, depressive symptoms and BMI as confounders.

Model $\widehat{M}_2^{\text{CARE}}$, described by Equations (6.2a)–(6.2b) and explained in more detail in Section 4.5, would have the same limitation as $\widehat{M}_3^{\text{CARE}}$ with respect to the estimated

6.2 Joint model for frailty, falls and mortality with the CARE75+ data (Revisited)



sex : Gender	vgp(t) : # visits to a GP by t	$\varepsilon(t)$: Error of frailty score
eth : Ethnicity	$N(t)$: # falls at t	b_0 : Random intercept
mar : Married or remarried	$m(t)$: Frailty at t free of measurement error	u : Random effect shared by the two hazards
edu : Highest education	$y(t)$: Frailty score at t with measurement error	σ_ε^2 : Variance of $\varepsilon(t)$
alc : Drinks alcohol	$r(t)$: Hazard risk of falls	σ_b^2 : Variance of b_0
smo : Smoker	$h(t)$: Hazard risk of death	$\rho.$: Shape of Weibull
bmi(t) : BMI at t		$\kappa.$: Rate of Weibull
com(t) : # comorbidities by t		

Figure 6.1: DAGs G_3^{CARE} (left) and G_2^{CARE} (right) for frailty, recurrent falls and mortality in the CARE75+ data set, corresponding to models $\widehat{M}_3^{\text{CARE}}$ and $\widehat{M}_2^{\text{CARE}}$, respectively.

effects of frailty and falls on mortality since we included the same covariates in the submodel of frailty and mortality as those in model $\widehat{M}_3^{\text{CARE}}$. As discussed in Section 4.5, model $\widehat{M}_2^{\text{CARE}}$ differs from model $\widehat{M}_3^{\text{CARE}}$ in three main aspects: in model $\widehat{M}_2^{\text{CARE}}$ falls is exogenous time-varying covariate of mortality and of frailty, and the link between frailty and mortality is the current level of frailty free of measurement error ($m_i(t)$). The right panel of Figure 6.1 depicts G_2^{CARE}, the DAG implied by $\widehat{M}_2^{\text{CARE}}$, showing the relationships among all the variables in the model, in particular the covariates included in each submodel.

6.2.1 A causal joint model for frailty, falls and mortality for the CARE75+ data set

In order to construct a model for causal inference we would need to state upfront our causal assumptions, this is the hypotheses about the relationships among all the variables, identifying all possible confounders to the frailty-falls-mortality relationship, to finally fit the joint model adjusting for confounding, i.e. keeping in each submodel model all confounders even when their associated regression coefficients were not significant. We adapted DAGs G_3^{CARE} and G_2^{CARE} in order to make them consistent with our assumptions about confounding of frailty, falls and mortality, taking into account the empirical evidence in the literature. We named G_{3C}^{CARE} and G_{2C}^{CARE} the new DAGs and their corresponding joint models $\widehat{M}_{3C}^{\text{CARE}}$ and $\widehat{M}_{2C}^{\text{CARE}}$ respectively, to emphasize that they are based on our causal assumptions about confounding as follows:

- `ethnicity`, `sex`, `marital`, `education`, `alcohol`, `smoker`, `bmi` are common causes of frailty, falls and mortality
- `comorbidities` is a common cause of frailty and falls, but it acts on mortality through frailty and falls.
- `visits to GP` follows from people's frailty and comorbidities.

Figure 6.2 depicts DAGs G_{3C}^{CARE} and G_{2C}^{CARE} and Equations (6.3a)–(6.3c) and (6.4a)–(6.4b) describe their associated joint model, which we named M_{3C}^{CARE} and M_2^{CARE}. We tried to keep the nodes of the DAGs in Figure 6.2 in the same position as those in Figure 6.1, only the common causes of frailty, falls and mortality are in a slightly different

position being grouped a the top of the DAG. The DAGs of Figure 6.2 differ from those in Figure 6.1 only in the arrows pointing from the covariates to $m(t), r(t)$, $h(t)$ and $\texttt{gpv}(t)$, while the arrows connecting frailty, falls and mortality are kept exactly the same.

$$M_{3C}^{\text{CARE}}: \begin{cases} \text{L:} & y_i(t \mid b_{i0}) = (\beta_0 + b_{i0}) + \beta_t t + \mathbf{x}_i^\top(t)\boldsymbol{\beta} + \varepsilon_i(t) & \text{(6.3a)} \\ \text{R:} & r_i(t \mid \mathbf{v}_i) = u_i r_0(t) \exp\{\mathbf{x}_i^\top(t)\boldsymbol{\gamma}_R + \eta_R b_{i0}\} & \text{(6.3b)} \\ \text{T:} & h_i(t \mid \mathbf{v}_i) = u_i^\alpha h_0(t) \exp\{\mathbf{w}_i^\top(t)\boldsymbol{\gamma}_T + \eta_T b_{i0}\}, & \text{(6.3c)} \end{cases}$$

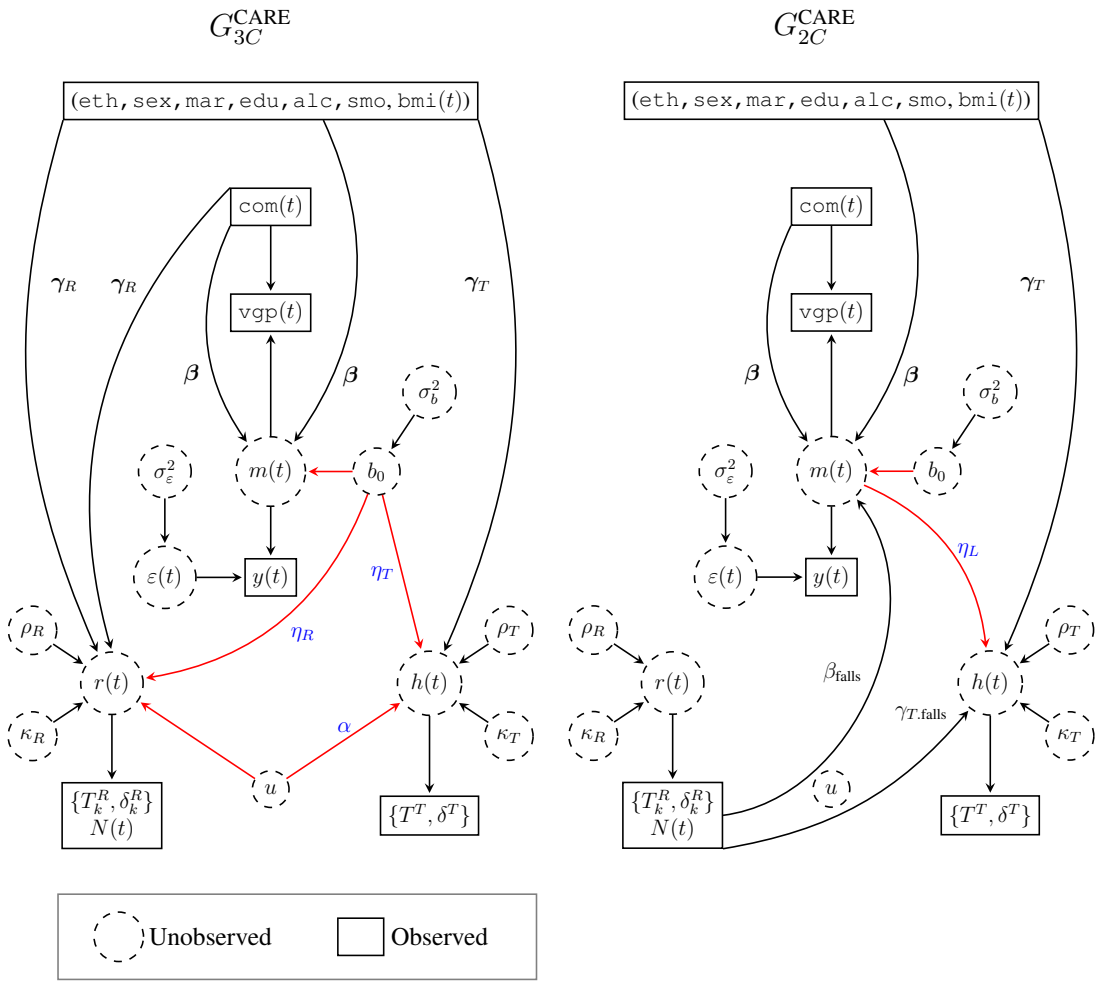
$$M_{2C}^{\text{CARE}}: \begin{cases} \text{L:} & y_i(t \mid b_{i0}) = \underbrace{(\beta_0 + b_{i0}) + \mathbf{x}_i^\top(t)\boldsymbol{\beta} + \beta_{\text{falls}} N_i(t)}_{m_i(t)} + \varepsilon_i(t) & \text{(6.4a)} \\ \text{T:} & h_i(t \mid b_{i0}) = h_0(t) \exp\{\mathbf{w}_i^\top(t)\boldsymbol{\gamma}_T + \gamma_{T.\text{falls}} N_i(t) + \eta_L m_i(t)\}, & \text{(6.4b)} \end{cases}$$

where

$$\begin{aligned} \mathbf{w}_i^\top(t) &= (\texttt{ethnicity}_i,\ \texttt{sex}_i,\ \texttt{marital}_i,\ \texttt{education}_i,\ \texttt{alcohol}_i, \\ &\quad\ \texttt{smoker}_i,\ \texttt{bmi}_i(t)) \\ \mathbf{x}_i^\top(t) &= \left(\mathbf{w}_i^\top(t),\ \texttt{comorbidities}_i(t)\right) \\ \boldsymbol{\beta}^\top &= \left(\beta_{\texttt{eth}},\ \beta_{\texttt{sex}},\ \beta_{\texttt{mar}},\ \beta_{\texttt{edu}},\ \beta_{\texttt{alc}},\ \beta_{\texttt{smo}},\ \beta_{\texttt{bmi}}\right). \\ \boldsymbol{\gamma}_R^\top &= \left(\gamma_{R.\texttt{eth}},\ \gamma_{R.\texttt{sex}},\ \gamma_{R.\texttt{mar}},\ \gamma_{R.\texttt{edu}},\ \gamma_{R.\texttt{alc}},\ \gamma_{R.\texttt{smo}},\ \gamma_{R.\texttt{bmi}},\ \gamma_{R.\texttt{com}}\right). \\ \boldsymbol{\gamma}_T^\top &= \left(\gamma_{T.\texttt{eth}},\ \gamma_{T.\texttt{sex}},\ \gamma_{T.\texttt{mar}},\ \gamma_{T.\texttt{edu}},\ \gamma_{T.\texttt{alc}},\ \gamma_{T.\texttt{smo}},\ \gamma_{T.\texttt{bmi}}\right). \\ \mathbf{v}_i^\top &= (b_{i0}, u_i). \end{aligned}$$

The vector $\mathbf{w}_i(t)$ contains the common causes of the three outcomes, and $\mathbf{x}_i(t)$ the common causes of frailty and falls, with regression coefficients vectors $\boldsymbol{\gamma}_T$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_R$ respectively. Note that the vectors $\mathbf{w}_i(t)$ and $\mathbf{x}_i(t)$ have variables in common. The links between the three outcomes are the same as those in models $\widehat{M}_3^{\text{CARE}}$ and $\widehat{M}_2^{\text{CARE}}$.

Table 6.1 shows the estimates of the associations between frailty, falls and mortality of models $\widehat{M}_{3C}^{\text{CARE}}$ and $\widehat{M}_{2C}^{\text{CARE}}$. For comparison, we copied on this same table the estimates of the corresponding associations of models $\widehat{M}_3^{\text{CARE}}$ and $\widehat{M}_2^{\text{CARE}}$ previously fitted in Chapter 3. Table 6.2 contains all the parameter estimates of models $\widehat{M}_{3C}^{\text{CARE}}$ and $\widehat{M}_{2C}^{\text{CARE}}$.



sex : Gender	vgp(t) : # visits to a GP by t	$\varepsilon(t)$: Error of frailty score
eth : Ethnicity	$N(t)$: # falls at t	b_0 : Random intercept
mar : Married or remarried	$m(t)$: Frailty at t free of measurement error	u : Random effect shared by the two hazards
edu : Highest education	$y(t)$: Frailty score at t with measurement error	σ_ε^2 : Variance of $\varepsilon(t)$
alc : Drinks alcohol	$r(t)$: Hazard risk of falls	σ_b^2 : Variance of b_0
smo : Smoker	$h(t)$: Hazard risk of death	$\rho.$: Shape of Weibull
bmi(t) : BMI at t		$\kappa.$: Rate of Weibull
com(t) : # comorbidities by t		

Figure 6.2: DAGs G_{3C}^{CARE} (left) and G_{2C}^{CARE} (right) for frailty, recurrent falls and mortality in the CARE75+ data set, corresponding to models $\widehat{M}_{2C}^{\text{CARE}}$ and $\widehat{M}_{2C}^{\text{CARE}}$, respectively.

Parameter		Estimate	Std.Err	p-value
$\widehat{M}_3^{\text{CARE}}$ vs. $\widehat{M}_{3C}^{\text{CARE}}$				
	$\widehat{\eta}_T$ (Frailty→Mortality)			
	$\widehat{M}_3^{\text{CARE}}$	0.430	0.206	0.037
	$\widehat{M}_{3C}^{\text{CARE}}$	0.408	0.198	0.040
	$\widehat{\alpha}$ (Falls→Mortality)			
	$\widehat{M}_3^{\text{CARE}}$	−0.962	0.703	0.171
	$\widehat{M}_{3C}^{\text{CARE}}$	−0.817	0.439	0.063
	$\widehat{\eta}_R$ (Frailty→Falls)			
	$\widehat{M}_3^{\text{CARE}}$	0.439	0.081	< 0.001
	$\widehat{M}_{3C}^{\text{CARE}}$	0.452	0.080	< 0.001
$\widehat{M}_2^{\text{CARE}}$ vs. $\widehat{M}_{2C}^{\text{CARE}}$				
	$\widehat{\eta}_L$ (Frailty→Mortality)			
	$\widehat{M}_2^{\text{CARE}}$	0.511	0.150	< 0.001
	$\widehat{M}_{2C}^{\text{CARE}}$	0.587	0.178	0.001
	$\widehat{\gamma}_{T.\text{falls}}$ (Falls→Mortality)			
	$\widehat{M}_2^{\text{CARE}}$	−0.259	0.235	0.269
	$\widehat{M}_{2C}^{\text{CARE}}$	−0.276	0.248	0.267
	$\widehat{\beta}_{\text{falls}}$ (Falls→Frailty)			
	$\widehat{M}_2^{\text{CARE}}$	0.329	0.051	< 0.001
	$\widehat{M}_{2C}^{\text{CARE}}$	0.330	0.051	< 0.001

Table 6.1: Comparison of the estimated effects of frailty and falls on mortality: $\widehat{M}_3^{\text{CARE}}$ & $\widehat{M}_2^{\text{CARE}}$ against $\widehat{M}_{3C}^{\text{CARE}}$ & $\widehat{M}_{2C}^{\text{CARE}}$.

Table 6.1 shows that with respect to the estimates of model $\widehat{M}_3^{\text{CARE}}$, the estimated effects of frailty on mortality $(\widehat{\eta}_T)$ and falls on mortality $(\widehat{\alpha})$ in model $\widehat{M}_{3C}^{\text{CARE}}$ decreased in absolute terms, and the effect of frailty on falls $(\widehat{\eta}_R)$ is larger. The standard errors of all three estimates in model $\widehat{M}_{3C}^{\text{CARE}}$ are smaller than those estimated in model $\widehat{M}_3^{\text{CARE}}$. The sign of all three estimates did not change and the p-values of $\widehat{\eta}_T$, $\widehat{\alpha}$ and $\widehat{\eta}_R$ changed as little such that the conclusions of the hypotheses tests remain the same. The smaller standard error of $\widehat{\alpha}$ in $\widehat{M}_{3C}^{\text{CARE}}$ suggests that adjusting for confunding has reduced noise and has made this estimate more stable, resulting in a smaller p-value (0.063) almost to the point of becoming significant at a level of 0.05. The interpretation of the associations between frailty, falls and mortality in the CARE75+ data estimated with model $\widehat{M}_{3C}^{\text{CARE}}$ are:

- Effect of frailty on mortality. The relative risk of death increases $\exp(0.408) = 1.504$ with a unit increase in frailty, *ceteris paribus*.
- Effect of falls on mortality. Higher risk of falls decreases the risk of death. As discussed in Chapter 4, this result might be due to relatively short follow-up period and small number of deaths and falls. Additionally, it is possible that the falls reported by the CARE75+ participants are not severe enough to pose important deterioration in their general health condition or mobility, and they implement precautionary measures, preventing further falls or making them less severe although more frequent.
- Effect of frailty on falls. The relative risk of falls increases $\exp(0.452) = 1.571$ with a unit increase in frailty, *ceteris paribus*.

The estimate of the effect of frailty on mortality $(\widehat{\eta}_L)$ in model $\widehat{M}_{2C}^{\text{CARE}}$ is 15% greater than its estimate in model $\widehat{M}_2^{\text{CARE}}$, and its standard error greater by 19%. According to DAG G_2^{CARE} (Figure 6.2) in model $\widehat{M}_{2C}^{\text{CARE}}$, frailty mediates the falls $\rightarrow$ mortality relationship. The direct effect of falls on the relative hazard of mortality is $\exp(\widehat{\gamma}_{T.\text{falls}}) = 0.759$ and the indirect effect is $\exp(\widehat{\beta}_{\text{falls}}\widehat{\eta}_L) = 1.214$ (VanderWeele, 2011) . So the total effect of falls on mortality is 0.921, calculated as the sum of direct and indirect effects.

6.2.2 Choosing the mean structure for joint modelling frailty, falls and mortality

Comparing the two casual models for frailty, falls and mortality that we proposed in this chapter, we notice that to some extent the qualitative conclusions we can make from $\widehat{M}_{3C}^{\text{CARE}}$ and $\widehat{M}_{2C}^{\text{CARE}}$ are similar: direct association between frailty and mortality and between frailty and falls, and weak to no association between falls and mortality. The parameter $\widehat{\eta}_T$ in model $\widehat{M}_{3C}^{\text{CARE}}$ accounts for the effect frailty $\rightarrow$ mortality and, just like the effect frailty $\rightarrow$ falls, although the interpretation of this parameter depends entirely on the interpretation of the random intercept b_{i0}, as discussed in Chapter 4. In contrast, in model $\widehat{M}_{2C}^{\text{CARE}}$ the parameter that accounts for the frailty $\rightarrow$ mortality effect is $\widehat{\eta}_L$ and, as shown in DAG G_2^{CARE} of Figure 6.2, this relationship is characterized by $m_i(t)$, i.e. the true and latent frailty that includes b_{i0}. It is not clear to us which model to choose between $\widehat{M}_{3C}^{\text{CARE}}$ and $\widehat{M}_{2C}^{\text{CARE}}$ and it is not straightforward to decide on the best way to characterize the frailty $\rightarrow$ mortality relationship. In Section 6.3 we conduct a simulation study to explore the consequences of misspecifying the mean structure in joint modelling longitudinal and time-to-event data.

There is still more to explore about the relationships between these three outcomes, in particular about the association between frailty and falls. On the one hand, according to model $\widehat{M}_{3C}^{\text{CARE}}$ we cannot rule out the hypothesis of an effect frailty $\rightarrow$ falls, although the interpretation of the parameter that accounts for this relationship, $\widehat{\eta}_R$, depends on the interpretation of the random intercept b_{i0}, just as the effect frailty $\rightarrow$ mortality. On the other hand, in $\widehat{M}_{2C}^{\text{CARE}}$ the estimate $\widehat{\beta}_{\text{falls}}$ accounts for an effect in the opposite direction i.e. from falls $\rightarrow$ frailty with a straightforward interpretation (every additional fall contributes to increase frailty as much as $\widehat{\beta}_{\text{falls}}$). In this thesis we did not address exhaustively this relationship, and it a topic that we regard as future work.

	$\widehat{M}_{3C}^{\text{CARE}}$			$\widehat{M}_{2C}^{\text{CARE}}$		
Parameter	Estimate	Std.Err	p-value	Estimate	Std.Err	p-value
Fixed effects						
Frailty						
$\widehat{\beta}_0$	6.571	0.622	< 0.001	6.043	0.598	< 0.001
$\widehat{\beta}_t$	-0.032	0.112	0.774	-0.094	0.111	0.396
$\widehat{\beta}_{\text{sex}}$	-0.028	0.243	< 0.909	-0.028	0.230	0.905
$\widehat{\beta}_{\text{eth}}$	-1.967	0.366	< 0.001	-2.151	0.350	< 0.001
$\widehat{\beta}_{\text{mar}}$	-0.849	0.235	< 0.001	-0.825	0.223	< 0.001
$\widehat{\beta}_{\text{edu}}$	-0.830	0.237	< 0.001	-0.814	0.224	< 0.001
$\widehat{\beta}_{\text{alc}}$	-1.353	0.239	< 0.001	-1.238	0.235	< 0.001
$\widehat{\beta}_{\text{smo}}$	-0.010	0.205	0.963	-0.014	0.197	0.944
$\widehat{\beta}_{\text{bmi}}$	-0.011	0.019	0.563	0.003	0.018	0.880
$\widehat{\beta}_{\text{com}}$	0.229	0.031	< 0.001	0.250	0.030	< 0.001
$\widehat{\beta}_{\text{falls}}$	$+$	$+$	$+$	0.330	0.051	< 0.001
Falls						
$\widehat{\gamma}_{R.\text{sex}}$	-0.322	0.279	0.249	$+$	$+$	$+$
$\widehat{\gamma}_{R.\text{eth}}$	1.028	0.439	0.019	$+$	$+$	$+$
$\widehat{\gamma}_{R.\text{mar}}$	-0.470	0.262	0.073	$+$	$+$	$+$
$\widehat{\gamma}_{R.\text{edu}}$	0.210	0.257	0.414	$+$	$+$	$+$
$\widehat{\gamma}_{R.\text{alc}}$	-0.512	0.250	0.041	$+$	$+$	$+$
$\widehat{\gamma}_{R.\text{smo}}$	-0.074	0.257	0.766	$+$	$+$	$+$
$\widehat{\gamma}_{R.\text{bmi}}$	-0.002	0.020	0.903	$+$	$+$	$+$
$\widehat{\gamma}_{R.\text{com}}$	-0.064	0.042	0.122	$+$	$+$	$+$
Mortality						
$\widehat{\gamma}_{T.\text{sex}}$	-0.327	0.585	0.576	-0.237	0.566	0.675
$\widehat{\gamma}_{T.\text{eth}}$	0.222	0.859	0.796	1.551	0.881	0.078
$\widehat{\gamma}_{T.\text{mar}}$	0.275	0.560	0.623	0.633	0.546	0.246
$\widehat{\gamma}_{T.\text{edu}}$	-0.455	0.584	0.436	0.131	0.593	0.825
$\widehat{\gamma}_{T.\text{alc}}$	-0.104	0.650	0.874	0.445	0.675	0.509
$\widehat{\gamma}_{T.\text{smo}}$	-0.115	0.568	0.840	-0.123	0.550	0.823
$\widehat{\gamma}_{T.\text{bmi}}$	-0.036	0.053	0.497	-0.071	0.056	0.204
$\widehat{\gamma}_{T.\text{falls}}$	$+$	$+$	$+$	-0.276	0.248	0.267

Table 6.2 continued on next page

Continuation of Table 6.2

	$\widehat{M}_{3C}^{\text{CARE}}$			$\widehat{M}_{2C}^{\text{CARE}}$		
Parameter	Estimate	Std.Err	p-value	Estimate	Std.Err	p-value
Association						
$\widehat{\eta}_R$	0.452	0.080	< 0.001	$+$	$+$	$+$
$\widehat{\eta}_T$	0.408	0.198	0.040	$+$	$+$	$+$
$\widehat{\alpha}$	-0.817	0.439	0.063	$+$	$+$	$+$
$\widehat{\eta}_L$	$+$	$+$	$+$	0.587	0.178	0.001
Variance component						
Frailty						
$\widehat{\sigma}_\varepsilon$	2.152	$-$	$-$	1.459	$-$	$-$
$\widehat{\sigma}_b^2$	2.577	$-$	$-$	2.188	$-$	$-$
$\widehat{\phi}$	1.074	0.578	0.032	$+$	$+$	$+$
Baseline hazard						
Falls						
$\widehat{\kappa}_R$	1.030	$-$	$-$	$+$	$+$	$+$
$\widehat{\rho}_R$	0.477	$-$	$-$	$+$	$+$	$+$
Mortality						
$\widehat{\kappa}_T$	1.032	$-$	$-$	$+$	$+$	$+$
$\widehat{\rho}_T$	0.286	$-$	$-$	$+$	$+$	$+$

Table 6.2: Parameter estimates of models $\widehat{M}_{3C}^{\text{CARE}}$ and $\widehat{M}_{2C}^{\text{CARE}}$. "$-$" not directly available from the software output. "$+$" not a parameter to be estimated by the model on the column.

6.3 Simulation study

From our analysis of the previous section, based on the CARE75+ data we are not able to distinguish between models $\widehat{M}_{3C}^{\text{CARE}}$ and $\widehat{M}_{2C}^{\text{CARE}}$. Thus we consider important to do a sensitivity analysis.

In this section we conduct a simulation study with the purpose of exploring how different the conclusions might be when we fit the "wrong" model to the data for two different joint models of longitudinal, recurrent and terminal events data. Even though the two models we explore with are structurally different and not comparable in every single aspect, we are interested in knowing how badly estimated are the fixed effects parameters when using the "wrong" model and the extent to which conclusions made from the interpretation of the association parameters are different. We found that most of the fixed effects regression coefficients are well estimated with confidence intervals having coverage above 0.90 even when the wrong model is fitted, and that with one of the two models (M_2 explained in next paragraph) the conclusions about the association parameters is in the right direction, even if it is the "wrong" model for the data.

The two joint models we explore with, M_3 and M_2, are described by Equations (6.5a)–(6.5c)) and (6.6a)–(6.6b), with their corresponding DAGs G_3 and G_2 depicted in Figure 6.3.

$$M_3 : \begin{cases} \text{L:} & y_i(t \mid b_{i0}) = (\beta_0 + b_{i0}) + \beta_t t + \beta_4 w_{i4} + \beta_5 w_{i5} + \varepsilon_i(t) & \text{(6.5a)} \\ \text{R:} & r_i(t \mid \mathbf{v}_i) = u_i r_0(t) \exp\{\gamma_{R3} w_{i3} + \gamma_{R6} w_{i6} + \eta_R b_{i0}\} & \text{(6.5b)} \\ \text{T:} & h_i(t \mid \mathbf{v}_i) = u_i^{\alpha} h_0(t) \exp\{\gamma_{T1} w_{i1} + \gamma_{T2} w_{i2} + \eta_T b_{i0}\}, & \text{(6.5c)} \end{cases}$$

$$M_2 : \begin{cases} \text{L:} & y_i(t \mid b_{i0}) = \underbrace{(\beta_0 + b_{i0}) + \beta_t t + \beta_4 w_{i4} + \beta_5 w_{i5}}_{m_i(t)} + \varepsilon_i(t) & \text{(6.6a)} \\ \text{T:} & h_i(t \mid b_{i0}) = h_0(t) \exp\{\gamma_{T1} w_{i1} + \gamma_{T2} w_{i2} + \eta_L m_i(t) + \eta_N N_i(t)\}, & \text{(6.6b)} \end{cases}$$

where

$$
\begin{aligned}
m_i(t) &= (\beta_0 + b_{i0}) + \beta_t t + \beta_4 w_{i4} + \beta_5 w_{i5}, \\
N_i(t) &= \text{Counting process with intensity rate } r_i(t) = r_0(t)\exp\{\gamma_{R3} w_{i3} + \gamma_{R6} w_{i6}\}, \\
\varepsilon_i(t) &\sim \mathcal{N}(0, \sigma^2_\varepsilon), \\
b_{i0} &\sim \mathcal{N}(0, \sigma^2_b), \\
u_i &\sim \text{Gamma}(\phi^{-1}, \phi^{-1}), \\
r_0(t) &= \kappa_R \rho_R (\rho_R t)^{\kappa_R - 1}, \\
h_0(t) &= \kappa_T \rho_T (\rho_T t)^{\kappa_T - 1}.
\end{aligned}
$$

Model M_3 is a three-outcome joint model, where a regression submodel is specified for each one. In this case, the longitudinal outcome and the recurrent event are considered to be endogenous time-varying covariates for mortality and assumed to be measured with error. This model is similar to model $\widehat{M}^{\text{CARE}}_{3C}$ in terms of the associations between the quantitative outcome and the recurrent and terminal events, although they differ with each other in their confounding structure. For the sake of avoiding sources of confusion, we decided to keep confounding as simple as possible in the simulation study with no time-varying covariates but time (t) in the linear-mixed submodel (Equation (6.5a)).

Model M_2 is a two-outcome joint model for the longitudinal quantitative outcome and a terminal event. Its association between the longitudinal outcome and the terminal event is the same as in $\widehat{M}^{\text{CARE}}_{2C}$ it also includes a counting process as an exogenous time-varying covariate of the terminal event submodel. Confounding in M_2 is the same as in M_3 and differs with respect to $\widehat{M}^{\text{CARE}}_{2C}$ in the same way M_3 differs from $\widehat{M}^{\text{CARE}}_{3C}$, as described above.

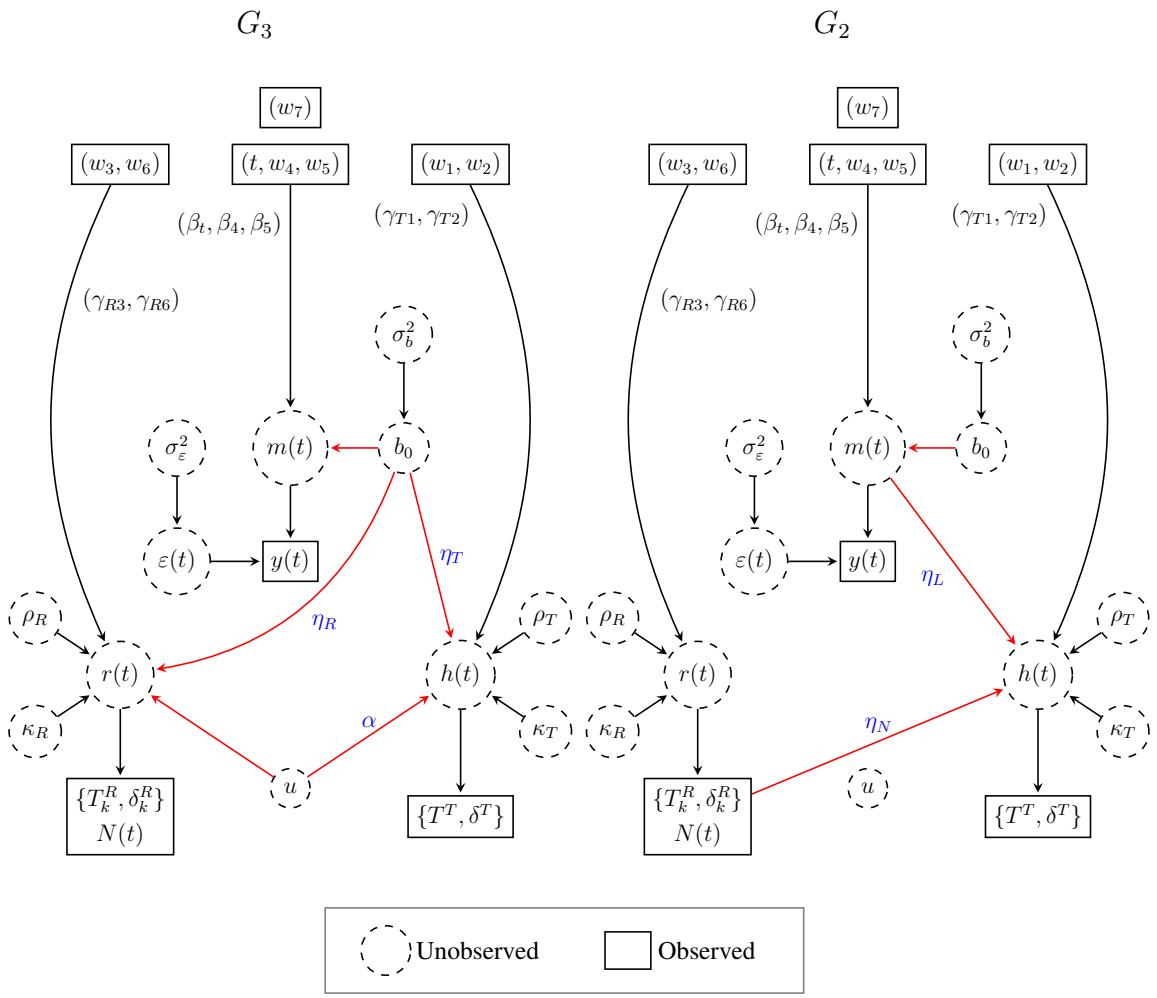


Figure 6.3: DAGs for simulations. Left: The DAG G_3 is the three-outcome joint model where they are linked through the random effects b_0 and u. Right: The DAG G_2 is the two-outcome joint model where the link between the longitudinal outcome and the terminal event is the current value of the "true" and unobserved $m(t)$.

In both models there are seven time-fixed covariates, $\mathbf{w}_i^\top = (w_{i1}, \cdots, w_{i7})$. All seven covariates are in each submodel, but only two have non-zero coefficient in each submodel: (w_{i1}, w_{i2}) of the terminal event, (w_{i4}, w_{i5}) of the longitudinal outcome, and (w_{i3}, w_{i6}) of the hazard rate of the recurrent event in M_3. In the terminology of DAGs, this means that (w_{i1}, w_{i2}) are parents of the terminal event $(h_i(t))$, (w_{i3}, w_{i6}) are parents of the recurrent event $(r_i(t))$, and (w_{i4}, w_{i5}) are parents of the longitudinal outcome $(m_i(t))$, as shown in both DAGs of Figure 6.3.

Model M_2 differs from M_3 mainly in three aspects. First, in M_2 the association between the quantitative outcome and the terminal event is through the current level of the latent variable $m_i(t)$, i.e. the link function is $g(\mathbf{b}_i, t) = m_i(t)$. Second, the observed cumulative count of recurrent event occurrences, $N_i(t)$, generated by the hazard rate $r_i(t)$, is assumed an exogenous time-varying covariate for the terminal event submodel, so M_2 does not have a regression equation for the recurrent event since it is not an outcome meant to be modelled while M_3 models the hazard rate of the recurrent event, $r_i(t)$. And third, in model M_2 there no arrows connecting the recurrent event and the quantitative outcome so they are not associated. In model M_2 there are only two association parameters: η_L the effect of the quantitative outcome on the terminal event, and η_N the effect of the cumulative count of the recurrent event on the terminal event. Model M_3 has three association parameters: η_R for the association between the quantitative outcome and the recurrent event, η_T between the quantitative outcome and the terminal event, and α between the recurrent and terminal events.

The fixed-effects regression coefficients are the vectors $(\beta_0, \beta_t, \boldsymbol{\beta}^\top)^\top = (\beta_0, \beta_t, \beta_1, \ldots, \beta_7)$ for the linear mixed model, $\gamma_R^\top = (\gamma_{R1}, \ldots, \gamma_{R7})$ for the recurrent event model, and $\gamma_T^\top = (\gamma_{T1}, \ldots, \gamma_{T7})$ for the terminal event model.

6.3.1 Simulation design

We simulated a total of $n_{\text{sim}} = 150$ data sets from each model, M_3 and M_2, each with a sample size $n = 500$ subjects. We denote by $D_k = (D_k^{(1)}, \ldots, D_k^{(150)})$ the collection of data sets generated from model $M_k,\ k = 3, 2$. For each subject in D_k, a vector of seven covariates $(\mathbf{w}_i,\ i = 1, \ldots, n)$ was sampled according to the following scheme: $(w_1, \ldots, w_4)$ are independent Bernoulli$(p = 0.5)$ trials, and (w_5, w_6, w_7) are uncorrelated instances of the multivariate normal distribution, $\mathcal{N}_3(\mathbf{0}, \Sigma)$, with $\Sigma = \text{diag}(\sigma_5^2 = 1, \sigma_6^2 = 4, \sigma_7^2 = 4)$. The n covariate vectors, $\mathbf{w}_i$, were sampled once for each i and kept the same values for all simulations.

The longitudinal outcome, $m_i(t)$, is assumed unobservable and instead the error-prone variable $y_i(t) = m_i(t) + \varepsilon_i(t)$ is observable, where $\varepsilon_i(t) \sim \mathcal{N}(0, \sigma_\varepsilon^2 = 1.25)$ is the measurement error. The variance of the normally distributed and zero-mean random intercept was set at $\text{var}(b_{i0}) = \sigma_b^2 = 1.25$. The random effect linking the two hazard

rates was chosen to be instances of the random variable $u_i \sim \text{Gamma}\,(\phi^{-1}, \phi^{-1})$, with $\phi = 0.64$. The parameters of the Gamma distribution are shape and rate, such that $\mathbb{E}(u_i) = \text{shape} \times \text{rate}^{-1} = 1$ and $\text{var}(u_i) = \text{shape} \times \text{rate}^{-2} = \phi$.

An administrative censoring time was set to 5.5 ($C_i = 5.5\ \forall i$), and repeated measures of the longitudinal outcome were generated at regular intervals every 0.2 time units, starting from $t = 0$, and with a maximum of 25 repeated measures, so this process stopped shortly before C_i. We opted for baseline hazards of both the recurrent and terminal event processes to be from the $\text{Weibull}(\kappa, \rho)$ distribution, where κ is the shape and ρ is the rate parameters (the reciprocal of the scale). There are several ways to parametrize the Weibull distribution, and the one we used corresponds to a cumulative hazard $H(t) = (\rho t)^{\kappa}$ and hazard rate $h(t) = \kappa\rho(\rho t)^{\kappa-1}$.

The association parameters were chosen to be all positive and not too large. All the values of the parameters for the simulation study are summarized in Table 6.3.

The estimates of $\widehat{M}_{3C}^{\text{CARE}}$ and $\widehat{M}_{2C}^{\text{CARE}}$ guided our choice of the values of the parameters in the simulation models M_3 and M_2. However, we were interested in producing data sets with more terminal events (less censoring) and more repeated measures of the longitudinal outcome, so some the parameters are substantially different, like $\phi = 0.64$, $\alpha = 2.6$ and $\rho_T = 1.5$. Consequently, a large proportion of terminal event times are short and hence not censored by C_i, and both M_3 and M_2 produced much more repeated measures of the longitudinal outcome. Recall from Chapter 4 that only 12.4% of the participants of the CARE75+ study have as many as 4 repeated measures of frailty, while in the simulations subjects can have up to 25 repeated measures, as shown in the right plots of Figure 6.5. Compare the left column plots of Figure 6.5 against the left plot in Figure 4.3 to see how the simulated longitudinal outcome differs from the frailty data of the CARE75+ data set, and compare the red lines of the right column plots of Figure 6.5 against Kaplan–Meier curve in Figure 4.6 to see the differences between simulated terminal event times and mortality.

The mean and median terminal event times for the baseline are $\mathbb{E}(T) = \int_0^{\infty} S_0(t)\text{d}t \approx 0.67$ and $T_{(0.5)} = S_0^{-1}(0.5) = (-\log(0.5))^{1/\kappa_T}\ /\rho_T \approx 0.46$ respectively. The observed mean and median survival times in the 150 simulated data sets differs between models M_3 and M_2, due to the mean structure.

	Related variable	M_3	M_2
Fixed effects			
Longitudinal			
β_0	1	3	3
β_t	time	0.5	0.5
β_4	w_4	0.5	0.5
β_5	w_5	0.5	0.5
Recurrent			
γ_{R3}	w_3	0.1	0.1
γ_{R6}	w_6	0.1	0.1
Terminal			
γ_{T1}	w_1	0.1	0.1
γ_{T2}	w_2	0.1	0.1
Association			
η_T	$m(t) \leftarrow b_{i0} \rightarrow h(t)$	0.5	–
α	$r(t) \leftarrow u_i \rightarrow h(t)$	2.6	–
η_R	$m(t) \leftarrow b_{i0} \rightarrow r(t)$	0.2	–
η_L	$m(t) \rightarrow h(t)$	–	0.2
η_N	$N(t) \rightarrow h(t)$	–	0.2
Variance comonent			
σ^2_ε		1.25	1.25
σ^2_b		2.25	2.25
ϕ		0.64	–
Baseline hazard (Weibull)			
Recurrent			
κ_R (shape)		2	2
ρ_R (rate)		0.5	0.5
Terminal			
κ_T (shape)		1	1
ρ_T (rate)		1.5	1.5
Link function			
$g_R(m_i(t))$ (Longitudinal-Recurrent)		b_{i0}	–
$g_T(m_i(t))$ (Longitudinal-Terminal)		b_{i0}	$m_i(t)$

Table 6.3: Simulation scheme. Table entries with "–" mean that the parameter on the row is not (or does not have the same interpretation, as described on the "Related variable" column) in the model of that column. Only non-zero regression coefficients are shown in the table, so $(\beta_1, \beta_2, \beta_3, \beta_6, \beta_7) = \mathbf{0}$, $(\gamma_{R1}, \gamma_{R2}, \gamma_{R4}, \gamma_{R5}, \gamma_{R7}) = \mathbf{0}$, and $(\gamma_{T3}, \ldots, \gamma_{T7}) = \mathbf{0}$

The values of the shape and rate parameters of the Weibull baseline hazards were chosen to keep both the expected count of the recurrent event and the proportion of censored observations low in order to ease the computational burden during the estimation process. The expected cumulative number of events by time τ_i for a subject with covariate vector $\mathbf{w}_i$ and possibly random effects vector $\mathbf{v}_i^\top = (b_{i0}, u_i)$, is given by (Cook & Lawless, 2007)

$$\mathbb{E}\{N_i(\tau_i \mid \mathbf{w}_i, \mathbf{v}_i)\} = \int_0^{\tau_i} r_i(t \mid \mathbf{w}_i, \mathbf{v}_i)\mathrm{d}t = \int_0^{\tau_i} u_i r_0(t) \exp\{\mathbf{w}_i^\top \boldsymbol{\gamma} + \eta b_{i0}\}\mathrm{d}t.$$

So, given the values of the Weibull parameters shown in Table 6.3, the expected number of recurrent events by the censoring time, C_i, for the baseline is $\mathbb{E}\left\{N_0\left(C_i|\mathbf{w}_i = \mathbf{0}, \mathbf{v}_i^\top = (1,0)\right)\right\} = \int_0^{\tau_i} r_0(t)\mathrm{d}t = 7.56$. In our simulation study, the number of recurrent events per subject is generally smaller than this because the recurrent event process is interrupted either by censoring or the terminal event (terminal event times of two instances of the simulated data sets are represented by • in the right column plots of Figure 6.5).

Algorithms 6.1 and 6.2 describe the simulation process of models M_3 and M_2 respectively, which are written assuming a more general linear mixed submodel having a random intercept and random slope, $\mathbf{b}_i^\top = (b_{i0}, b_{i1})$.

Algorithm 6.1 describes the simulation process from model M_3, where the link between the longitudinal outcome and the other two outcomes is only through the random effects, $\mathbf{b}_i$. We fixed the n covariate vectors $\mathbf{w}_i$ $(i = 1, \ldots, n)$ along the 150 simulated data sets to restrict the variability in the data to the random variables of the joint model; however, the $\mathbf{w}_i$ vectors used to simulate from model M_3 are different from those used to simulate from model M_2. Finding a closed expression for $H_i(t)$ and $H_i^{-1}(t)$, T_i^* is straightforward for common distributions of the baseline hazard, e.g. Weibull, Log-logistic, Gompertz and Makeham. However, when the linear predictor of the time-to-event submodel includes time-varying covariates or when the baseline hazard function involves complicated functions of time, then numerical methods are required to compute the integral $H_i(t) = \int_0^t h_i(s)\mathrm{d}s$ and its inverse, $H_i^{-1}(t)$, as discussed in Section 2.2.1.

Simulation of the inter-event times of the recurrent event process is done sequentially

Algorithm 6.1 Simulation from a joint of longitudinal, recurrent and terminal events outcomes linked with a random intercept (model M_3).

Set an administrative censoring time $C_i = C \; \forall i$, $C \in \mathbb{R}^+$.
For $i = 1, \ldots, n$:

1: Sample a baseline covariate vector, $\mathbf{w}_i$, according to the length of $\boldsymbol{\beta}, \boldsymbol{\gamma}_R, \boldsymbol{\gamma}_T$.
2: Sample an instance of the random effects, $\mathbf{b}_i \sim \mathcal{N}_2(\mathbf{0}, \Sigma)$ and $u_i \sim \text{Gamma}\,(\phi^{-1}, \phi^{-1})$.
3: Simulate the time to the terminal event, T_i:
   (i) Sample $\xi_i \sim \mathcal{U}(0,1)$
   (ii) $T_i^* = H_i^{-1}\left(-\log(\xi_i) \mid \mathbf{b}_i, u_i, \mathbf{w}_i, \boldsymbol{\gamma}_T, h_0(t)\right)$
   (iii) $T_i = \min(C_i, T_i^*)$
   (iv) $\delta_i = \mathbb{1}(T_i^* < C_i)$
4: Simulate the longitudinal outcome for time points $t_{ij},\ j = 1, \ldots, n_i$, where $t_{in_i} < T_i$:
   (i) Sample $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$
   (ii) $y_{ij} = (\beta_0 + b_{i0}) + (\beta_t + b_{i1})t_{ij} + \mathbf{w}_i^\top \boldsymbol{\beta} + \varepsilon_{ij}$.
5: Simulate the time to the k^{th} recurrent event, $k = 1, \ldots$:
   Set $T_{i0} = 0$.
   For $k = 1, \ldots$
   (i) Sample $\xi_{ik} \sim \mathcal{U}(0,1)$
   (ii) $T_{ik}^* = T_{ik-1} + \tilde{R}_{T_{ik-1}}^{-1}\left(-\log(\xi_{ik})\right)$
   (iii) $T_{ik} = \min(T_{ik}^*, T_i)$
   (iv) $\delta_{ik} = \mathbb{1}(T_i^* < T_i)$
   (v) Continue while $T_{ik} \leq T_i$.
   (vi) $N_i(T_{ik}) = \sum_{k=1}^{K_i} \mathbb{1}(T_{ik})$, where $K_i = \max\{k : T_{ik} \leq T_i\}$

Repeat n_{sim} times steps (2)–(5), i.e. fix the n covariate vectors $\mathbf{w}_i$ $(\text{1} = 1, \ldots, n)$ in all n_{sim} data sets.

by inverting the conditional cumulative hazard, $\tilde{R}_t(x) = R(x+t) - R(t)$, as described in Algorithm 2.3 in Section 2.2.2, which has a closed form if the baseline hazard comes from the Weibull distribution (for other common distributions this is also the case).

Model M_2 has both $m_i(t)$ (the longitudinal outcome free of measurement error) and $N_i(t)$ (the cumulative recurrent event counts) as a covariates of the terminal event submodel. Since $N_i(t)$ is a step function with jumps at the recurrent event times and $N_i(t)$ is a factor of the hazard function, $h_i(t)$, the latter is piecewise continuous. Figure 6.4 plots the data of an instance of simulated data of one individual, showing the longitudinal outcome and recurrent event process, the hazard rate, cumulative hazard and survival curves. The top plot depicts the quantitative outcome and the recurrent event counts, showing that the recurrent event process $(N_i(t))$ is a step function, making the hazard (second plot) a piecewise continuous function with jumps at the recurrent event times. Therefore the cumulative hazard and survival functions are continuous, but not differentiable at the recurrent event times (third and bottom plots respectively).

Therefore, data simulation from model M_2, as described in Algorithm 6.2, is done by sequentially generating the inter-event times and verifying whether at each time interval $H_i(t \mid \cdot)$ should be inverted. Only in the case where the baseline hazard of the terminal event is governed by the $\text{Exponential}(\rho)$ distribution, it is possible to obtain a closed expression for the (inverse) cumulative hazard. For other specifications of the baseline hazard, in order to find the inverse of the (conditional) cumulative hazard of the terminal and the recurrent event, numerical methods can be used. This becomes clear if we break down the cumulative hazard as follows:

$$
\begin{aligned}
H_i\left(t \mid b_{i0}, \mathcal{M}_i(t), \mathcal{F}_i^N(t)\right) &= \int_0^t h_0(s) \exp\left\{\mathbf{w}^\top \boldsymbol{\gamma} + \eta_L m_i(s) + \eta_N N_i(s)\right\} \mathrm{d}s \\
&= \int_0^t h_0(s) \exp\left\{\mathbf{w}^\top \boldsymbol{\gamma} + \eta_L(\beta_0 + \mathbf{w}_i^\top \boldsymbol{\beta}) + \eta_L b_{i0} + \eta_L \beta_t s + \eta_N N_i(s)\right\} \mathrm{d}s \\
&= \exp\left\{\mathbf{w}^\top \boldsymbol{\gamma} + \eta_L(\beta_0 + \mathbf{w}_i^\top \boldsymbol{\beta}) + \eta_L b_{i0}\right\} \int_0^t h_0(s) \exp\left\{\eta_L \beta_t s + \eta_N N_i(s)\right\} \mathrm{d}s \\
&= \exp\left\{\mathbf{w}^\top \boldsymbol{\gamma} + \eta_L(\beta_0 + \mathbf{w}_i^\top \boldsymbol{\beta}) + \eta_L b_{i0}\right\} \int_0^t h_0(s) \exp\left\{\eta_L \beta_t s + \eta_N N_i(s)\right\} \mathrm{d}s \\
&= \exp\left\{\mathbf{w}^\top \boldsymbol{\gamma} + \eta_L(\beta_0 + \mathbf{w}_i^\top \boldsymbol{\beta}) + \eta_L b_{i0}\right\} \sum_{q=1}^{Q_i} \int_{\Omega_{iq}} h_0(s) \exp\left\{\eta_L \beta_t s + \eta_N N_i(s)\right\} \mathrm{d}s,
\end{aligned}
$$

Algorithm 6.2 Joint model simulation for longitudinal and time-to-event outcomes with a counting process as an external time-varying covariate (model M_2).

Set an administrative censoring time $C_i = C \; \forall i$, $C \in \mathbb{R}^+$.
For $i = 1, \cdots, n$:

1: Sample a baseline covariate vector, $\mathbf{w}_i$, according to the length of $\boldsymbol{\beta}, \boldsymbol{\gamma}_R, \boldsymbol{\gamma}_T$.
2: Sample an instance of the random effects, $\mathbf{b}_i \sim \mathcal{N}_2(\mathbf{0}, \Sigma)$.
3: Sample $\xi_i \sim \mathcal{U}(0, 1)$, to be used to simulate the terminal event, T_i^*.
4: Simulate the inter-event times of the recurrent event:
   (i) Set $T_{i0}^* = 0$.
   (ii) For $k = 1, \cdots$
      (a) Sample $\xi_{ik} \sim \mathcal{U}(0, 1)$
      (b) k^{th} recurrent event time: $T_{ik}^* = T_{ik-1}^* + \tilde{R}_{T_{k-1}^*}^{-1} \left(-\log(\xi_{ik}) \mid \mathbf{w}_i, \boldsymbol{\gamma}_R \right)$
      (c) Counting process: $N_i(T_{ik}^*) = \sum_{k \geq 1} \mathbb{1}(T_{ik}^*)$
   (iii) Continue until $L_{ik} \leq -\log(\xi_i) < U_{ik}$ or $T_{ik}^* \leq C_i$, where
   C_i is the censoring time,
   $L_{ik} = H\left(T_{ik-1}^* \mid \mathcal{M}_i\left(T_{ik-1}^*\right), N_i(T_{ik-1}^*), \mathbf{b}_i, \mathbf{w}_i, \boldsymbol{\gamma}_T, \eta_R, \eta_L\right)$ and
   $U_{ik} = H\left(T_{ik}^* \mid \mathcal{M}_i\left(T_{ik}^*\right), N_i(T_{ik}^*), \mathbf{b}_i, \mathbf{w}_i, \boldsymbol{\gamma}_T, \eta_R, \eta_L\right)$,
   $\mathcal{M}_i(t)$ is the trajectory of $m_i(t)$ up to time t.
5: If the counting process was stopped by C_i, then the time to the terminal event is unobserved and thus censored. Otherwise, simulate the time to the terminal event, which lies within the interval $[L_{iK_i}, U_{iK_i}]$, where $K_i = \max\{k : T_{ik} \leq C_i\}$:
   $T_i^* = \left\{t \in [L_{ik}, U_{ik}] : H\left(t \mid \mathcal{M}_i\left(t\right), N_i(t), \mathbf{b}_i, \mathbf{w}_i, \boldsymbol{\gamma}_T, \eta_R, \eta_L\right) + \log(\xi_i) = 0\right\}$,
6: Create the event indicators for both the recurrent event and terminal event times:
   1. $T_i = \min(T_i^*, C_i)$
   2. $\delta_i = \mathbb{1}(T_i^* \leq C_i)$
   3. $T_{ik} = \min(T_{ik}^*, T_i, C_i), \; k = 0, \cdots, K_i$
   4. $\delta_{ik} = \mathbb{1}(T_{ik} = T_{ik}^*, T_{ik} > 0)$
7: Simulate the longitudinal outcome for time points $t_{ij}, \; j = 1, \cdots, n_i$, where $t_{in_i} \leq T_i$:
   1. Sample $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$
   2. $y_{ij} = (\beta_0 + b_{i0}) + (\beta_t + b_{i1}) t_{ij} + \mathbf{w}_i^\top \boldsymbol{\beta} + \varepsilon_{ij}$.

Repeat n_{sim} times steps (2)–(7), i.e. fix the n covariate vectors $\mathbf{w}_i$ $(\imath = 1, \ldots, n)$ in all n_{sim} data sets.

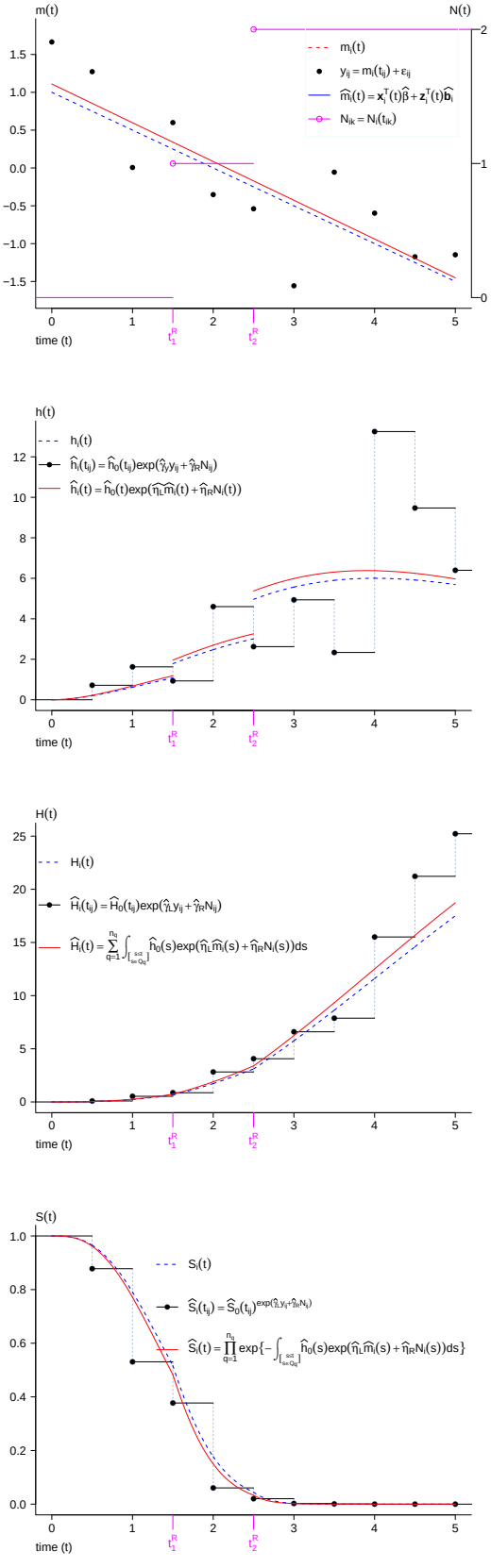


Figure 6.4: Top: repeated measures of a quantitative outcome with error (●), counts of the recurrent event (○—), true and unobserved quantitative measure without error (- - -), estimate of the quantitative outcome with joint modelling (——). Bottom three: hazard, cumulative hazard and survival curves with $N(t)$ being a time-varying covariate of the survival analysis submodel. the true & unobserved (- - -), estimates based on the marginal Cox model (●—), estimated by joint modelling (——).

where $\mathcal{M}_i(t)$ and $\mathcal{F}_i^N(t)$ are the history of the longitudinal outcome $(m_i(t))$ and the counting process up to time t, and the sets $\{\Omega_{iq}, q = 1, \ldots, Q_i\}$ denote the time intervals in which the time-varying covariates of subject i are assumed constant. The integrals involve the baseline hazard and the factors of the linear predictor that are, potentially complicated, functions of time. When the hazards are constant, as is the case of the Exponential distribution, these integrals are simplified.

Appendices A.6 and A.7 give sample R code to simulate data from these two joint models. The function of Appendix A.7 allows using baseline hazards from the Weibull, Log-logistic, Gompertz and Makeham distributions and can be easily tailored for other forms of the hazard function. It can also be adapted with little programming for non-linear functions of time in the linear mixed model part. The numerical integrals to compute the (conditional) cumulative baseline hazards use the Gauss–Kronrod method (Ziegel, 1987) implemented in the `integrate()` function of R. The inverse of the (conditional) cumulative hazard function (step 5 of Algorithm 6.2) is obtained by applying the crude bisection method (Monaco *et al.*, 2018), i.e. by finding the root of $H_i(t|\cdot) + \log(\xi_i)$, which is implemented in R as the univariate root finder function `uniroot()` within a specified interval, say $[L, U]$:

$$T^* = \{t \in [L, U] : H(t \mid \cdot) + \log(\xi) = 0\}.$$

6.3.2 Estimation

The parameters of model M_3 are

$$\Theta_3 = \{\beta_0, \beta_t, \boldsymbol{\beta}^\top, \boldsymbol{\gamma}_R^\top, \boldsymbol{\gamma}_T^\top, \eta_T, \alpha, \eta_R, \kappa_R, \kappa_T, \rho_R, \rho_T, \sigma_\varepsilon^2, \sigma_b^2, \phi\}.$$

In Section 3.3 we discussed parameter estimation of joint models like M_3. In this simulation study we followed the same approach as for fitting model $\widehat{M}_3^{\text{CARE}}$ described in Section 4.3.3.

The parameters of model M_2 are

$$\Theta_2 = \{\beta_0, \beta_t, \boldsymbol{\beta}^\top, \boldsymbol{\gamma}_R^\top, \boldsymbol{\gamma}_T^\top, \eta_L, \eta_N, \kappa_T, \rho_T, \sigma_\varepsilon^2, \sigma_b^2\}.$$

The estimation of Θ_2 is done by optimizing the log-likelihood function of Equation (3.2). Rizopoulos (2012) proposed a hybrid optimization procedure of the log-likelihood function starting with the EM algorithm or a fixed number of iterations (100) and, if convergence is not achieved, switching to a quasi-Newton algorithm until convergence. The R package `JM` (Rizopoulos, 2012) implements this procedure in the function `jointModel()`. The design matrices of models like M_2 are time-dependent due to the inclusion of $N_i(t)$ as a covariate. For joint models with exogenous time-varying covariates `JM` will not estimate a parametric baseline, instead it is approximated with regression splines. This is achieved by expanding $\log h_0(t)$ into B-spline basis functions for cubic splines:

$$\log h_0(t) = k_0 + \sum_{l=1}^{L} k_l B_l(t, q),$$

where the vector $\mathbf{k}^\top = (k_0, \cdots, k_L)$ denotes the spline coefficients, q is the degree of the B-spline basis functions (B), and $L = L' + q - 1$, with L' denoting the number of internal knots.

6.3.3 Results

Each simulated data set $(D_k^{(s)},\ s = 1, \ldots, 150)$ contains data of $n = 500$ subjects. The total number of records of each data set is different because the number of repeated measures and recurrent events per subject is not known *a priori*, and depends on the time elapsed before the terminal event, which is also simulated.

Figure 6.6 illustrates what the 150 simulated data sets look like, no modelling attempt is made. On the top row of Figure 6.6, the Kaplan–Meier curves show that the proportion of censored observations of the samples from model M_3 is approximately 20%. Due to the value of the association parameter for the $m(t) - h(t)$ relationship $(\eta_L = 0.2)$ in model M_2 and the link being the current value of the complete longitudinal outcome (without measurement error), the survival probability of these samples decreases faster, so the proportion of censored terminal events in model M_2 is practically zero.

The population survival probabilities were obtained by evaluating the survival function of each model at the true parameter values, the average covariate values $((\bar{w}_1, \bar{w}_2) = (0.5, 0.5))$ and the expected value of the random effects ($\mathbb{E}(u_i) = 1$ and $\mathbb{E}(b_{i0}) = 0$). The reason why the Kaplan-Meier curve differs from the population survival probabilities in data D_3 (top right of Figure 6.5 and top left of Figure 6.6) is because of the large value of α in the former and, in the latter, the survival function is evaluated at $\mathbb{E}(u_i) = 1$ and $\mathbb{E}(b_{i0}) = 0$ and does not take into account censoring.

The bottom row of Figure 6.6 shows the sample longitudinal outcome profiles, $m(t)$, of the data generated from models M_3 and M_2 of the 150 simulations. Each line represents the general trajectory of each sample. As expected, the lines of model M_3 are very close to the "true" sample profile $m(t) = 3 + 0.5t$, and the differences are due to randomness of the measurement error and covariates. The lines of model M_2 reflect the mutual dependence of the longitudinal and terminal event outcomes. The association parameter, $\eta_L = 0.2$, means that subjects with larger $m_i(t)$ values tend to have fewer repeated measures because they have a shorter survival times and *vice versa*, subjects with small $m_i(t)$ values tend to have larger survival times and thus more repeated measures. That is the reason why we observe positive and negative slopes in the overall profiles (see left of bottom half of Figure 6.5) In particular, the strong deviation of the slopes from the true population trend, $m(t) = 3 + 0.5t$, is what we lose by neglecting the dependence between the longitudinal outcome and terminal event. Figure 6.7 depicts the sample profile estimates, $\widehat{m}^{(k)}(t) = \widehat{\beta}_0^{(k)} + \widehat{\beta}_t^{(k)} t$ $(k = 1, \cdots, 150)$, illustrating how by joint modelling these two outcomes it is possible to correct for the bias caused by the dependence between the two outcomes.

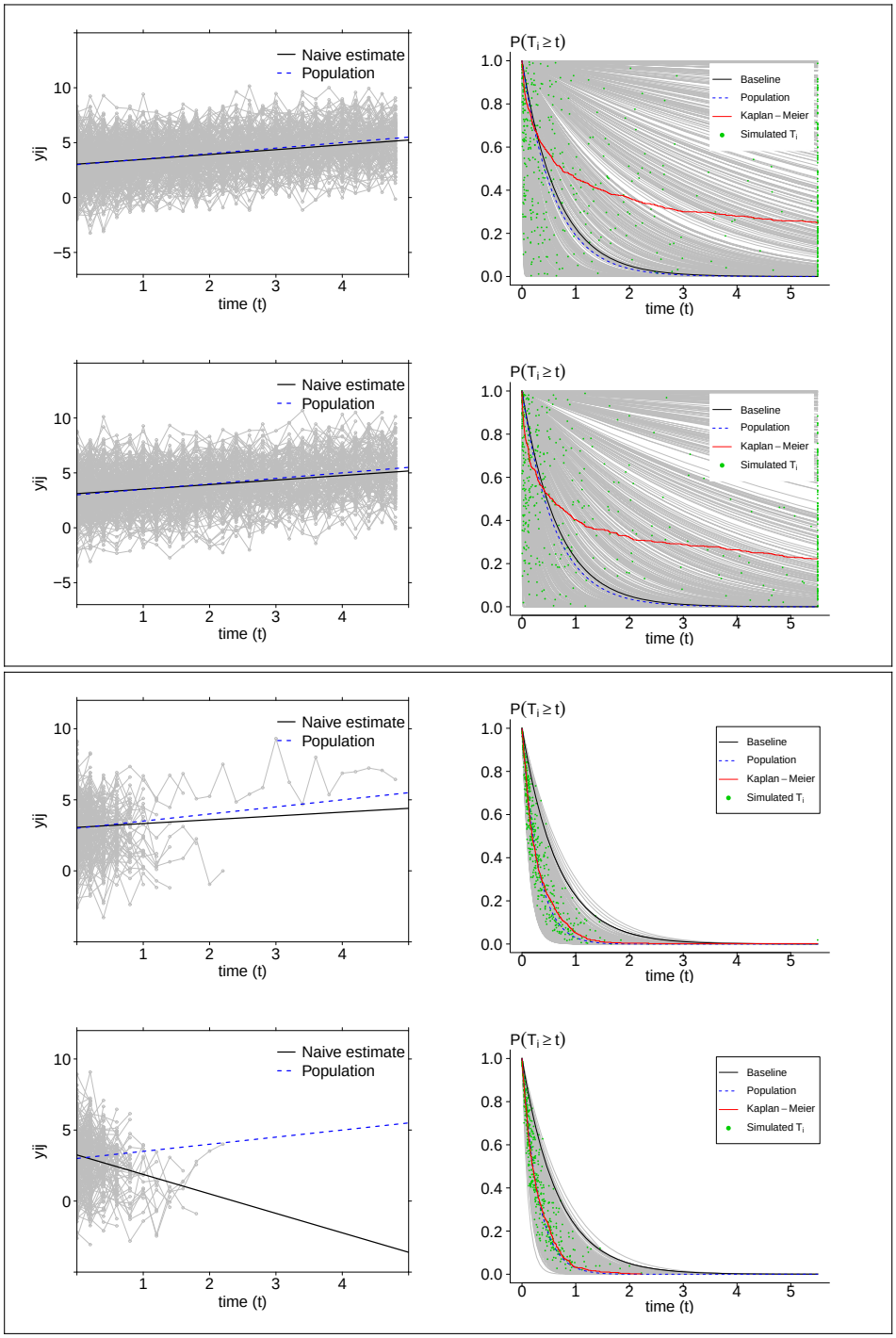


Figure 6.5: Top half: Two samples generated from model M_3; bottom half: Two samples generated from model m_2. Left plots: Individual profiles of the longitudinal outcome y_{ij} (–●–); population profile (- -), and naively estimated population profile ignoring the time-to-event outcome (—). Right plots: Survival probabilities (—); general population survival probability (- -); simulated time-to-event, T_i (●), and the Kaplan–Meier estimate (—).

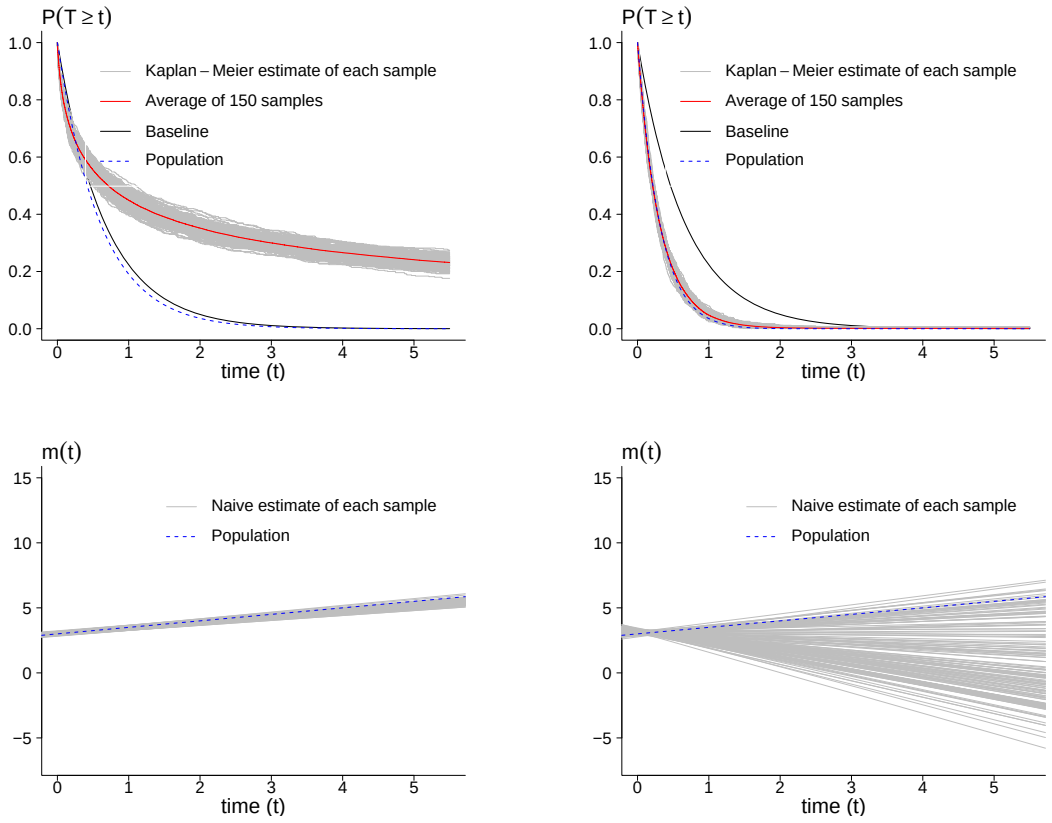


Figure 6.6: Left: D_3, Right: D_2. Top: Kaplan–Meier estimates of the 150 simulated data sets (—); average of the 150 Kaplan–Meier curves (—), and population survival curve (- -). Bottom: naive estimate of the longitudinal outcome population profile ignoring the time-to-event outcome (—), and true population profile (- -).

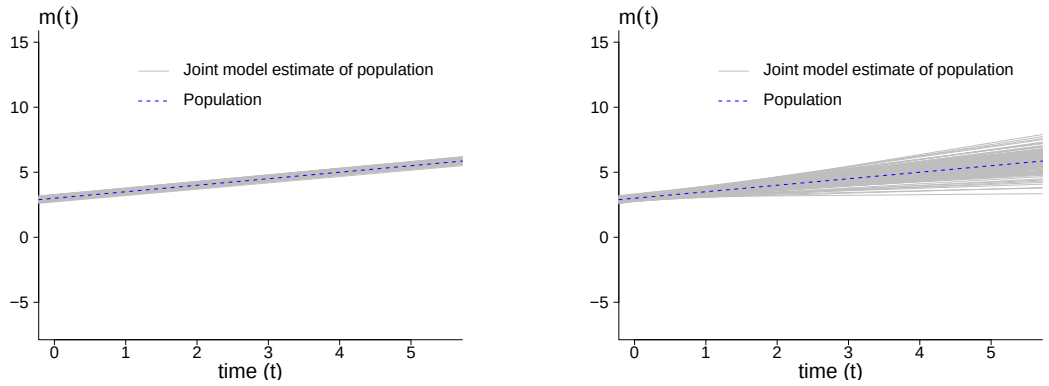


Figure 6.7: Left: D_3, Right: D_2. Joint model estimate of the longitudinal outcome population profile (——) of each of the 150 samples, and true population profile (- - -).

Figure 6.8 shows a histogram of the average cumulative count of recurrent events occurred by C_i in the 150 data sets. The larger number of recurrent events in the three-outcome data D_3 occurs because of the larger survival times mainly because of the stronger relationship between the recurrent event process on the terminal event process through u_i^{α} $(\alpha = 2.6)$, consequently for many subjects of D_3 the data generating process of $N_i(t)$ and $y_i(t)$ continued for longer.

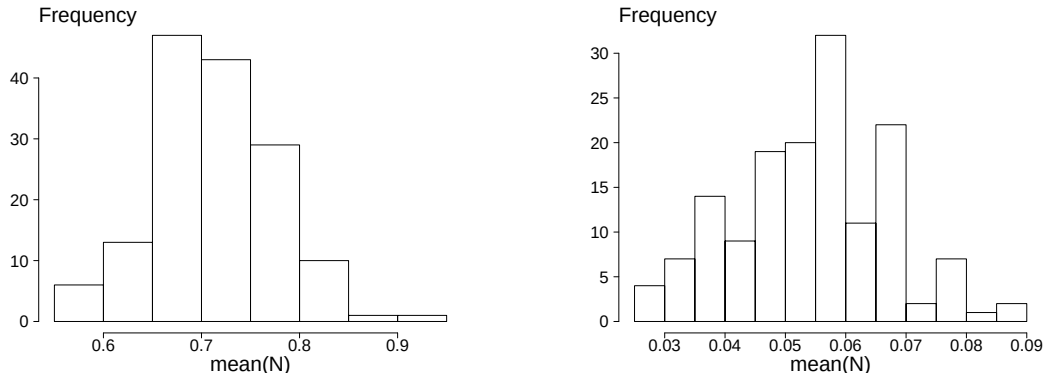


Figure 6.8: Average of the cumulative recurrent events that the n subjects observed by the censoring time in each data set: $n^{-1}\sum_{i=1}^{n} N_i(C_i)$. Left: D_3; right: D_2.

For each $D_k^{(s)},\ s = 1, \ldots, 150$ we fit first the "correct" model and then the "wrong"

one, which results in two sets of parameter estimates per data set: one for the correct model and one for the wrong one. Even though models M_3 and M_2 are different, mainly in their association structure, and not comparable with each other in every single parameter, they do have several common parameters: the fixed effects $(\beta_0, \beta_t, \boldsymbol{\beta}, \boldsymbol{\gamma}_R, \boldsymbol{\gamma}_T)$, the variance of the measurement error of the longitudinal outcome (σ^2_ε), the variance of the random intercept (σ^2_b), and the Weibull parameters of the baseline hazard of the terminal event (κ_T, ρ_T). Our interest is to investigate the possible consequences of misspecifying the association structure in a joint model for longitudinal, recurrent and terminal events data. In particular:

1. if by fitting the wrong model the common parameters can be still correctly estimated as if they were estimated with the correct model, and
2. if the conclusions we would make from interpreting the estimated association parameters of the wrong model would be different from the conclusions we would make from fitting the correct model. For instance, we already know that data D_2 is simulated with association parameters' values $\eta_L = 0.2$ and $\eta_N = 0.2$, and we want to know if fitting model M_3 to data D_2 results in $\widehat{\eta}_T > 0$ and $\widehat{\alpha} > 0$.

D_3 (data simulated with model M_3)

Table 6.4 show the relative bias (R.Bias) and mean-squared error (MSE) of the estimates and the coverage (C) of their 95% confidence intervals of models M_3 and M_2 fitted to D_3, and Table 6.5 for D_2. We can see in both tables that the fixed effects are correctly estimated even when the wrong model is fitted, especially fixed effects of the longitudinal and the recurrent event submodels. The interval estimates of the fixed effects of the terminal event submodel are not as good as in the other two submodels, especially when model M_3 is fitted to data D_2.

When model M_2 is fitted to data D_3 (Table 6.4), the fixed effects estimates of the linear mixed submodel $(\beta_0, \beta_t, \boldsymbol{\beta})$ are nearly unbiased, have small MSE and the corresponding 95% interval estimates have large coverage C $> 92\%$. The fixed effects of the terminal event submodel $(\boldsymbol{\gamma}_T)$ seem to be more difficult to estimate, whether with model M_3 or M_2. By fitting M_2 to data D_3, $(\widehat{\gamma}_{T1}, \widehat{\gamma}_{T2})$ have small MSE $(0.014, 0.013)$, but

		M_3			M_2		
Parameter	True (M_3)	R.Bias	MSE	C (%)	R.Bias	MSE	C (%)
Fixed effects							
β_0	3	0.001	0.011	89.4	-0.006	0.011	92.7
β_t	0.5	0.004	0.000	96.2	-0.004	0.000	96.0
β_4	0.5	0.000	0.025	89.4	-0.002	0.022	96.0
β_5	0.5	-0.002	0.006	88.6	-0.012	0.005	95.3
γ_{R3}	0.1	0.256	0.020	95.5	–	–	–
γ_{R6}	0.1	0.099	0.002	93.9	–	–	–
γ_{T1}	0.1	1.806	0.707	88.5	-0.589	0.015	92.7
γ_{T2}	0.1	1.218	0.918	71.3	-0.532	0.013	90.7
Association							
η_T	0.5	2.862	4.353	68.9	–	–	–
α	2.6	0.834	16.753	3.8	–	–	–
η_R	0.2	1.058	0.055	28.8	–	–	–
η_L	–	–	–	–	$-0.282^{(\dagger)}$	0.004	$97.3^{(\ddagger)}$
η_N	–	–	–	–	$-0.844^{(\dagger)}$	4.822	$100.0^{(\ddagger)}$
Variance component							
σ^2_ε	1.25	$4.397^{(\star)}$	$104.916^{(\star)}$	$0.8^{(\star)}$	0.000	0.001	95.3
σ^2_b	2.25	-0.007	0.030	94.7	-0.008	0.030	95.3
ϕ	0.64	5.712	13.640	32.6	–	–	–
Baseline hazard (Weibull)							
κ_R	2	-0.060	0.016	13.6	–	–	–
κ_T	0.5	0.890	1.228	71.2	–	–	–
ρ_R^{-1}	1	-0.247	0.246	0.0	–	–	–
ρ_T^{-1}	1.5	0.519	0.129	2.3	–	–	–

Table 6.4: Relative bias (R.Bias), mean-squared error (MSE) and coverage (C) of estimates of models M_3 and M_2 fitted to data D_3. "–" means that the parameter on the row does not exist in the model of that column. $^{(\star)}$ Refers to σ_ε. $^{(\dagger)}$ refers difference with respect to zero since parameter does not exist in true model hence there is no true value. $^{(\ddagger)}$ must be read as the relative frequency in which both end points of the of interval estimate have the same sign as the true value of the association parameter.

		M_3			M_2		
Parameter	True (M_2)	R.Bias	MSE	C (%)	R.Bias	MSE	C (%)
Fixed effects							
β_0	3	0.004	0.011	95.9	−0.003	0.012	94.0
β_t	0.5	−0.006	0.019	92.6	0.021	0.019	94.7
β_4	0.5	−0.070	0.023	96.6	0.004	0.026	96.0
β_5	0.5	−0.061	0.007	89.3	0.015	0.006	94.7
γ_{R3}	0.1	0.607	0.328	95.3	–	–	–
γ_{R6}	0.1	0.629	0.020	93.3	–	–	–
γ_{T1}	0.1	−0.148	0.010	89.9	0.063	0.008	96.0
γ_{T2}	0.1	−0.219	0.006	79.9	−0.016	0.008	96.0
Association							
η_T	–	$-0.569^{(\dagger)}$	0.083	$100.0^{(\ddagger)}$	–	–	–
α	–	$-0.896^{(\dagger)}$	5.428	$54.4^{(\ddagger)}$	–	–	–
η_R	–	$0.871^{(\dagger)}$	0.100	$41.6^{(\ddagger)}$	–	–	–
η_L	0.2	–	–	–	−0.018	0.001	96.7
η_N	0.2	–	–	–	0.315	0.052	88.7
Variance component							
σ_ε^2	$1.25^{(\star)}$	$-0.957^{(\star)}$	$1.145^{(\star)}$	0.0	0.002	0.005	94.7
σ_b^2	2.25	−0.013	0.046	93.9	−0.011	0.046	94.0
ϕ	–	$3.365^{(\dagger)}$	4.832	0.67	–	–	–
Baseline hazard (Weibull)							
κ_R	2	0.038	0.025	87.2	–	–	–
κ_T	0.5	0.058	0.004	41.6	–	–	–
ρ_R^{-1}	1.0	−0.458	0.843	0.0	–	–	–
ρ_T^{-1}	1.5	−0.461	0.282	81.9	–	–	–

Table 6.5: Relative bias (R.Bias), mean-squared error (MSE) and coverage (C) of estimates of models M_3 and M_2 fitted to data D_2. "–" means that the parameter on the row does not exist in the model of that column. $^{(\star)}$ Refers to σ_ε. $^{(\dagger)}$ refers difference with respect to zero since parameter does not exist in true model hence there is no true value. $^{(\ddagger)}$ must be read as the relative frequency in which both end points of the of interval estimate have the same sign as the true value of the association parameter.

they are underestimated with relative bias of $(-0.58, -0.56)$ respectively, and interval estimates with lower coverage compared to the other fixed effects $(92.7\%, 90.7\%)$.

Model M_2 does not include (w_{i3}, w_{i6}), the parents of $N_i(t)$, in the terminal event submodel since $N_i(t)$ is a covariate of the terminal event submodel, and by the *causal effect rule* in Equation (2.56) $(w_{i3}, w_{i6}) \perp \{t_i\delta_i\}$. It was verified in a separate simulation study that $\widehat{\gamma}_{T3} = \widehat{\gamma}_{T6} = 0$ (not shown in this document).

It is important to note that the fixed effects of the three submodels of M_3 are at least as well estimated by fitting model M_2 (the wrong model) as by M_3 (the correct model) to data D_3. When estimated with model M_3 the coverage of the interval estimates of $\beta_0, \beta_4, \beta_5, \gamma_{T1}$ and γ_{T2} is lower than 90.0%, and $\boldsymbol{\gamma}_{T1}, \boldsymbol{\gamma}_{T2}$ have large relative bias and MSE due to some large estimates as shown in Figures B.18 and B.21 in Appendix B.3.1. The reason for this might be that the complexity of model M_3 makes the estimation process more challenging: (a) the recurrent event process is also modelled; (b) the recurrent and terminal events submodels have two random effects, and (c) the parameters of the baseline hazards, which are nonlinear functions of time, are directly estimated.

The challenging optimization process of the log-likelihood function of M_3 is also reflected in the poorly estimated association, variance components and Weibull parameters, with σ_b^2 being the only one of these correctly estimated when M_3 is the fitted model (R.Bias $= 0.007$, MSE $= 0.030$, C$= 94.7\%$). The variance of the random intercept is correctly estimated with model M_3 even when it is the wrong model. The rest of the association, variance components and Weibull parameters estimates have large relative bias and MSE, and 95% interval estimates with poor coverage. In particular, as shown in Table 6.4 and Figure B.18 of Appendix B.3.1, the estimated variance of the random effect $\widehat{\text{var}}(u_i) = \widehat{\phi}$ is poorly estimated (R.bias $= 5.712$, MSE **=** $= 13.640$, C $= 32.6\%$). Barker & Henderson (2005) provide some evidence from simulation studies of the difficulty to estimate the variance parameter of the random effect in the Gamma-frailty model. They found ϕ to be underestimated in small or medium sized samples $(n = 200, 500, 1000)$. Underestimation of the random effect variance results in fixed effects being underestimated (Barker & Henderson, 2005; Henderson & Oman, 1999). In our simulation studies, biases of $\widehat{\phi}$ and the associations $\widehat{\eta}_T, \widehat{\alpha}$ and $\widehat{\eta}_R$ are pos-

itive. Extending our simulations to larger samples might be necessary to find out if biases get smaller.

It is worth noting that even when fitting M_3 to data D_3 the variance of the measurement measurement error (σ_ε^2) is poorly estimated (R.bias $= 4.397$, MSE $= 104.916$, C $= 0.8$), as shown in Table 6.4 and Figure B.29 of Appendix B.3.2. We noticed that when fitting M_3 some of the estimates are unusually highly correlated correlations higher than 0.7 (see Figures B.17 and B.28 in Appendices B.3.1 and B.3.2)

- Fit M_3 to D_3

$$\begin{aligned}
\text{cor}(\widehat{\eta}_T, \widehat{\alpha}) &= 0.92 \\
\text{cor}(\widehat{\eta}_T, \widehat{\sigma}_\varepsilon) &= 0.98 \\
\text{cor}(\widehat{\eta}_T, \widehat{\kappa}_T) &= 0.98 \\
\text{cor}(\widehat{\alpha}, \widehat{\sigma}_\varepsilon) &= 0.87 \\
\text{cor}(\widehat{\alpha}, \widehat{\kappa}_T) &= 0.98 \\
\text{cor}(\widehat{\sigma}_\varepsilon, \widehat{\kappa}_T) &= 0.87
\end{aligned}$$

- Fit M_3 to D_2

$$\text{cor}(\widehat{\eta}_T, \widehat{\sigma}_\varepsilon) = 0.99$$

Figures B.17 and B.28 in Appendices B.3.1 and B.3.2 show that regardless of which data we fit model M_3 to, $\widehat{\eta}_T$ and $\widehat{\sigma}_\varepsilon$ are almost perfectly correlated. We have been in communication with the author of the R package `frailtypack`, but we do not suspect programming problems. We need to investigate further the reason why these estimates are so highly correlated since they affect the association parameters estimates, which are among the parameters of primary interest.

Although the association parameters (η_T, α, η_R) of model M_3 are overestimated the range of the interval estimates of α and η_R are of the same sign as their true values across the 150 simulations, C $= (100.0\%, 99.2\%$ not shown in Table). The practical implication is that by fitting model M_3 we are able to correctly estimate the sign of the association between $r_i(t)$ & $h_i(t)$ and $m_i(t)$ & $r_i(t)$, but their magnitude might be

overestimated. We arrive to the same conclusion with respect to the direction of the association parameters when fitting model M_2 to D_3.

Table 6.4 and Figure B.21 show that fitting model M_3 to data D_3 understimates the parameters of the recurrent event baseline hazard, and overestimates those of the terminal event. The relative bias and MSE of these estimates is not as large as the other poorly estimated parameters, but the coverage of the interval estimates is very low, especially for the Weibull parameters of the terminal event hazard.

D_2 (data simulated from model M_2)

The simulation experiment carried out on data D_2 yields similar results in the sense that the fixed effects regression coefficients are correctly estimated even with the wrong model, M_3, as shown in Table 6.4. Not surprisingly, when fitting model M_2 (the correct model) the coverage of the interval estimates are larger compared to those obtained from fitting M_2 to D_3, especially $(\widehat{\gamma}_{T1}, \widehat{\gamma}_{T2})$ which have much smaller relative bias and MSE. This is also the case when fitting M_3 (the wrong model); nonetheless $(\widehat{\gamma}_{R3}, \widehat{\gamma}_{R6})$ have larger relative bias compared to fitting M_3 to D_3. The fact that we observe interval estimates of model M_3 with $\mathrm{C} < 90\%$ raises the question if $n = 500$ and $n_{\text{sim}} = 150$ are large enough sample size and simulations for a three-outcome joint model like M_3.

The association parameter of the $N_i(t) \rightarrow h_i(t)$ relationship (η_N) is the parameter of model M_2 most poorly estimated with R.Bias $= 0.315$, MSE $= 0.052$ and $\mathrm{C} = 88.7\%$. The rest of the parameters of model M_2 are correctly estimated in terms of relative bias, MSE and coverage.

When model M_3 is fitted to these data, the conclusions about the association parameters is correct for the association between the longitudinal and the terminal event, where 100% of the interval estimates will contain only positive values $(\eta_L = 0.2)$, but not for the $N_i(t) \rightarrow h_i(t)$. In such a case, the range of the interval estimate of α of the same sign as $\eta_N = 0.2$ is 54.6%.

6.4 Discussion and future work

The results of our simulation study suggest that for the two specific models we have considered the fixed effects are generally correctly estimated, even when a misspecified model is fitted. There are however limitations about this statement that can be easily overlooked, for instance by fitting M_2 we cannot estimate fixed effects for the recurrent event since it is not being modelled. Thus it is important to know exactly what features of the data each joint model is trying to estimate and, in the application context, what is the research question that we want to answer with the analysis of the data.

The large MSE and low coverage of some of the estimates in our simulation study might be due to the relatively small number of simulations and sample size given the complexity of joint modelling. The relatively large number of parameters to be estimated in joint models require larger samples than the marginal models for separate outcomes. Fitting the joint model of longitudinal, recurrent and terminal events (M_3) is particularly complicated due to the inclusion of multiple random effects in both hazard rate submodels.

When fitting the two-outcome joint model M_2, it is possible to estimate the sign of the association between the longitudinal outcome and the terminal event and between the recurrent and terminal events. A practical implication of these results is that by fitting model M_2 while analyzing real data we can still learn about the fixed effects and the associations longitudinal outcome $\rightarrow$ terminal event and recurrent $\rightarrow$ terminal events even if it is not the correct model that represents the data. The association parameters of Model M_2 have a straightforward interpretation analogous to the proportional hazards model, fitting M_2 requires accommodating less sources of within-subject variability represented by the random effects, and addresses in a flexible way the variability from the baseline hazard of the terminal event. More research would still be needed though to explore the extent to which this conclusion holds when the recurrent event is an endogenous time-varying covariate of the terminal event submodel and such endogeneity is ignored.

Additionally, we are interested in knowing if even when the some parameters of model M_3 are wrongly estimated they compensate in such a way that the log-likelihood eval-

uated at these estimates is close to the log-likelihood evaluated at the true values, if the estimated survival probabilities of these model differ from the survival probabilities evaluated at the true parameters ($S_i(t|\widehat{\Theta}_3) \approx S_i(t|\Theta_3)$), and if the fitted values of the longitudinal outcome are similar to the true values ($y_i(t|\widehat{\Theta}_3) \approx m_i(t|\Theta_3)$).

6.5 Future work and extensions

In our simulation studies of this chapter we noticed that some of the parameters of model M_3 are not correctly estimated and some are highly correlated. As we pointed out, this kind of joint model is challenging to fit and require of long processing times. We would like to use the parameter estimates $\widehat{\Theta}_3$ from our simulation study to calculate $\widehat{h}_i(t)$, $\widehat{S}_i(t)$ or $\widehat{H}_i(t)$ and compare against the values of these functions at the true parameters. What we are looking for is if the wrong estimates of model M_3 compensate each other giving unbiased estimates of the survival probabilities (or hazards). This requires predictions of the random effects.

We summarize the topics of future work and possible extensions in the following points:

- Extend the simulation study to explore with other possible relationships between frailty, falls and mortality:
  - Falls as a common cause of frailty and mortality.
  - Falls as a mediator of the frailty $\rightarrow$ mortality relationship.

  This requires adapting our data simulation algorithm and R programs by (a) making the recurrent event counts, $N_i(t)$, become part of the linear-mixed submodel accommodating the recurrent event times, and (b) making the longitudinal outcome, $m_i(t)$, part of the recurrent event process just as it is for the terminal event.
- Extend the simulation study to explore with other values of the association parameters. Will our conclusions change?
- Extend the simulation study to address other confounding structures:

  - Common causes of frailty, falls and mortality,
  - Time-varying confounding. This requires deciding on appropriate functions of time to simulate reasonable covariates in line with real data.

- Extend the simulation study to include a random slope. Although this does not seem necessary for analyzing the CARE75+ data, it seems reasonable in a broader context. As we mentioned along this thesis, the computer processing time increases dramatically when an additional random effect is added to a joint model. So exploring joint models with a random intercept and slope can be challenging, specially for joint models that include a recurrent event submodel in addition to the longitudinal and terminal event submodels, like M_3.

- In this thesis we worked with time-independent regression coefficients and incorporated the effect of time via a random slope the simulation studies of Chapter 5. An alternative is to use time-varying regression coefficients.

- We suspect that what makes the likelihood optimization of M_3 so challenging is that it involves integrating out 2 random effects and estimating the Weibull parameters of two hazard rates, which involve non-linear functions of time. As part of our possible future simulations we consider increasing the sample size $(n = 500, 1000)$ to examine whether the biases get smaller and explore with alternative optimization strategies. Additionally, it would be important to try with alternatives for joint modelling frailty, falls and mortality. For instance, van Houwelingen (2014) distinguishes pattern mixture models as an approach to joint modelling, starting with a marginal model for the time-to-event outcome $\{T_i, \delta_i\}$ and adding a model for the longitudinal response conditional on $\{T_i, \delta_i\}$ defined in terms of the time before the occurrence of the event. Henderson *et al.* (2000) propose a latent variable model where the association between longitudinal and the time-to-event outcomes is described by the cross-correlation of a latent bivariate Gaussian process.

Chapter 7

Conclusions

In this final chapter we summarize our main conclusions, indicating future work and possible extensions. We consider that our work makes contributions in two areas: statistical methodology and applications. Our contributions to the statistical methodology are (1) proposing a variable selection strategy for simultaneously optimizing prediction of the two outcomes in joint modelling of longitudinal and time-to-event data, (2) approaching the joint modelling framework from the causality perspective using DAGs, and (3) evaluating the consequences of misspecifying the mean and association structures of joint modelling a longitudinal outcome and recurrent and terminal events. In the applications context, geriatric frailty, recurrent falls and mortality have been analyzed before with marginal models, and we investigated their relationships with the joint modelling methodology.

Throughout Chapters 3–6 we discussed several methodological aspects of jointly modelling longitudinal outcomes and recurrent and terminal events, including variable selection, description, prediction, causal inference and model specification. The methods we discussed were applied to the CARE75+ data set using joint models for exploring plausible underlying mechanisms between frailty, falls and mortality. Frailty and falls, and their relationship with mortality have been analyzed before with marginal models. However,due to their nature, it is more appropriate to model them jointly as separate models do not fully account for their possible dependence. The joint modelling methodology seems well suited for analyzing the relationship among them.

In Chapter 4, we fitted a joint model to describe the relationship between frailty, falls and mortality in the CARE75+ data set ($\widehat{M}_3^{\text{CARE}}$) under the assumption that their association is completely characterized by a set of random effects. The fitted model suggests that being white, married or remarried, having higher education and drinking alcohol are associated with lower frailty, and that frailty increases with increasing number of comorbidities. The risk of falls is higher among white people with a relative risk of falls with respect to the other ethnicities of $\exp(1.429) = 4.175$. Once we condition on these covariates of frailty and falls, no covariate is significantly associated with mortality.

Frailty is strongly associated with both falls and mortality, an association that is characterized by a random intercept in the linear mixed submodel of frailty. The relative risk of falls is $\exp(0.439) = 1.551$ and the relative risk of mortality $\exp(0.430) = 1.537$ for subjects whose frailty is one unit above the population average, *ceteris paribus*.

The analyzed data set shows no signs of an association between falls and mortality after conditioning on covariates, since the corresponding parameter is not significant at the 5% level. It is worth considering at this point the complex epidemiology of falls, and that the falls $\rightarrow$ mortality relationship can be mediated by its multiple adverse outcomes, in particular hospitalization (Masud & Morris, 2001). Even though we consider an association between falls and mortality plausible, a longer follow-up period might be needed than is available in the CARE75+ study to observe this association.

The model diagnostics indicate some signs of skewness due to many large values of frailty. Further refinements and considerations might be required since in both falls and mortality submodels the fitted model $\widehat{M}_3^{\text{CARE}}$ seems to overestimate the risk of falls and mortality. Perhaps this is due to (1) the relatively small sample size for a joint model of longitudinal data and recurrent and terminal events, (2) the small number of terminal events, and (3) the lack of precise times at which falls occur.

The existing diagnostic tools to check the fit of joint models are lacking. A sensitivity analysis to explore with model uncertainty might allow for a closer inspection of the model fit and investigate new options.

We fitted an alternative joint model for frailty, falls and mortality ($\widehat{M}_2^{\text{CARE}}$). In this alternative model we considered falls as exogenous time-varying covariate of both frailty

and mortality, and specified a different link of the frailty $\rightarrow$ mortality relationship, in this case through the current frailty value free of measurement error $(m_i(t))$, i.e. frailty being an endogenous time-varying covariate of mortality. In qualitative terms, our conclusions about the association between frailty, falls and mortality with this alternative model are the same as those derived from model $\widehat{M}_3^{\text{CARE}}$.

In statistical modelling, variable selection is carried out in different ways depending on the intended use of the fitted model: *description*, *causal inference* or *prediction* (Shmueli *et al.*, 2010). The variable selection process to fit the joint models of Chapter 4 was carried out by a stepwise procedure which tries to optimize goodness of fit to the data (i.e. description), first in the marginal model of each outcome and then in the joint model. This approach becomes more difficult with correlated covariates, in addition to the long processing times required to fit a joint model. Moreover, we were interested in a joint model with a set of covariates that optimizes prediction of longitudinal and a time-to-event outcomes. By being able to get accurate predictions of frailty and mortality for the CARE75+ data it might be possible to adapt the management plans of care homes.

In Chapter 5 we proposed a variable selection strategy that aimed at optimizing prediction of a joint model for longitudinal and time-to-event outcomes. Our strategy consisted of penalizing the log-likelihood function with separate penalties for the fixed effects regression coefficients of each submodel and using a K-fold cross-validation design to estimate the parameters and score out-of-sample data to assess prediction by the mean-squared error (MSE) and the Integrated Brier score (IBS). We conducted simulation studies under different covariate scenarios. The results suggested that with highly correlated covariates, the region of the hyperpameters (the two penalties) of optimal MSE might not overlap with the region of optimal IBS. In such a case, our strategy chooses among values within a small region defined by two penalties that require a compromise between MSE and IBS depending on which outcome is the priority.

As a secondary criterion of performance, we assessed our variable selection strategy in the simulation studies by comparing the regression coefficient estimates against their true values under a binary classifier (zero and non-zero) in terms of accuracy, sensitivity and specificity. The values of these metrics are not as high as we have seen in the standard marginal regression models for normally distributed outcomes. Nonetheless,

this is of secondary interest for our proposed strategy, the principal criterion being to optimize prediction.

We constructed two additional binary classifiers that impose other restrictions to the binary classifier used in our simulation study. Even though we obtained similar results they seem too ambitious for the joint modelling context and we would rather explore how they perform in simpler contexts, like the marginal model of each outcome. These two binary classifiers are discussed briefly in the Section 5.5.1 and are considered for future work.

The goal of the two joint models fitted in Chapter 4 was to find a plausible model describing the relationship of frailty, falls and mortality and covariates in the CARE75+ data set. These models have limitations for causal inference because they do not explicitly take confounding into account. In Chapter 6, we used DAGs to state upfront our hypotheses about the relationships among all the variables in the CARE75+ data set, identifying all possible confounders to the frailty-falls-mortality relationship, and we fitted the joint model corresponding to the relationships stated on the DAGs: $\widehat{M}_{3C}^{\text{CARE}}$ and $\widehat{M}_{2C}^{\text{CARE}}$.

According to model $\widehat{M}_{3C}^{\text{CARE}}$, the relative risk of death increases by a factor of $\exp(0.408) = 1.504$ and relative risk of falls increases $\exp(0.452) = 1.571$ with a unit increase in frailty, *ceteris paribus*. The association between the risk of falls and the risk of mortality was not significant at the 5% level.

The results of model $\widehat{M}_{2C}^{\text{CARE}}$ suggest a direct effect of frailty on the relative risk of mortality by a factor $\exp(0.587) = 1.8$, from falls to frailty of 0.330, and a total effect of falls on mortality of 0.921.

There is still more to explore about the relationships between frailty, falls and mortality, in particular about the association between frailty and falls. On the one hand, according to model $\widehat{M}_{3C}^{\text{CARE}}$ we cannot rule out the hypothesis of an effect frailty $\rightarrow$ falls, although the interpretation of the parameter that accounts for this relationship, $\widehat{\eta}_R$, depends on the interpretation of the random intercept b_{i0}, just as the effect frailty $\rightarrow$ mortality. On the other hand, in $\widehat{M}_{2C}^{\text{CARE}}$ the estimate $\widehat{\beta}_{\text{falls}}$ accounts for an effect in the opposite direction i.e. from falls $\rightarrow$ frailty with a straightforward interpretation (every additional fall contributes to increase frailty as much as $\widehat{\beta}_{\text{falls}}$). It would be important to continue

exploring this relationship in order to determine the direction of the corresponding effect.

In Chapter 6 we conducted a simulation study to explore the consequences of model misspecification in joint modelling. More specifically our interest was to assess the extent in which the model parameters and regression coefficients are correctly/wrongly estimated when the data is analyzed with the wrong model. We simulated a series of datasets from two different joint models, M_3 and M_2 and analyze each data set with the two models.

The results of our simulation study suggest that for the two specific models we have considered the fixed effects are generally correctly estimated, even when a misspecified model is fitted. There are however limitations about this statement that can be easily overlooked, for instance by fitting M_2 we cannot estimate fixed effects for the recurrent event since it is not being modelled. Thus it is important to know exactly what features of the data each joint model is trying to estimate and, in the application context, what is the research question that we want to answer with the analysis of the data.

When fitting M_2, it is possible to estimate the sign of the association between the longitudinal outcome and the terminal event and between the recurrent and terminal events. A practical implication of these results is that by fitting model M_2 while analyzing real data we can still learn about the fixed effects and the associations longitudinal outcome $\rightarrow$ terminal event and recurrent $\rightarrow$ terminal events even if it is not the correct model that represents the data. The association parameters of Model M_2 have a straightforward interpretation analogous to the proportional hazards model. Fitting M_2 requires accommodating less sources of within-subject variability represented by the random effects, and addresses in a flexible way the variability from the baseline hazard of the terminal event.

More research would still be needed though to explore the extent to which the conclusion for model M_2 holds when the recurrent event is an endogenous time-varying covariate of the terminal event submodel and such endogeneity is ignored.

When fitting model M_3, estimates of the variance of the measurement error were poor (largely biased and inaccurate) and some parameter estimates were highly correlated regardless of which data (generated with M_3 or M_2) the model was fitted to. We would

like to use the parameter estimates from model M_3 to calculate $\widehat{h}_i(t)$, $\widehat{S}_i(t)$ or $\widehat{H}_i(t)$ and compare against the values of these functions at the true parameters. What we are looking is if the fit of model M_3 compensates for some biased parameter estimates by adjusting others, giving unbiased estimates of the survival probabilities (or hazards). This requires prediction of the random effects. This is part of our future work.

Due to the complex nature of frailty and falls, there might be several plausible underlying mechanisms for their relationship with mortality. It would be important to extend our simulation studies to investigate other modelling strategies, considering alternative association and confounding structures, as pointed out in Section 6.5.

Joint modelling longitudinal and time-to-event outcome has been an area of active research in latest years and, although all of the examples used in this document are from the medical context, its applications are not restricted to this area. Fitting a joint model is complex and computationally challenging hence the processing times are much longer than those for the marginal models. It might be important to explore with alternatives parameter estimation, even with methods that give an approximate solution. An option worth exploring is the possibility to use the ideas of *variational Bayes approximations*, see for example Ormerod & Wand (2010), or *message passing*, Wand (2017). The general idea is to approximate the posterior distribution of the parameters given the observed data, $f(\boldsymbol{\theta}|\mathcal{D})$ by another density, $q(\boldsymbol{\theta})$ that minimizes the Kullback-Leibler divergence.

Appendix A

R code

A.1 Simulation of survival analysis data (includes frailty)

Sample code to simulate data from Proportional Hazards models. Corresponds to Algorithm 2.1 and includes models of Sections 2.2.1, 2.2.1.

```
##================================================================
## Last update: 31/Dec/2017
## Version 1.0
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## R code Simulate data from a frailty model with 3
## time-fixed covariates and Weibull baseline hazard.
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Chapter: Preliminaries
## For first year 201617
## (c) Jacob Cancino-Romero (JHD,LB,SB) 2017
##================================================================

#================================================================
## Simulate Frailty PH survival data ~ Weibull and 3 covariates.
## k = 1,..., K_{i}  (here K_{i} = K)
## i = 1,..., n
##
##    The model from which the data is simulated from:
##
##    h_ik(t) = h_{0}(t)* u_{i} *exp(x_{ik}^{T}*beta),
##
##    h_{0}(t) ~ Weibull(shape=k, rate=r): k*r*(r*t)^(k-1)
##
##    u_{i}:
##       ~ LN(meanlog = 0, sdlog=sqrt(theta))
##       ~ gamma(shape=1/theta, scale=theta)
##
## Gamma distribution: dgamma(shape=a, scale=s):
## f(x)= 1/(s^a Gamma(a)) x^(a-1) e^-(x/s)
## E(X) = a*s = 1/theta * theta = 1
## V(X) = a*s^2 = 1/theta * theta^2 = theta
## ..............................................................
## The general inverse transfor method for the model with frailty:
##
```

```
##  T_{ik} = - 1/u_{i} * log(z) * exp(-x_{ik}^{T}*beta)
##
#=================================================================

#=================================================================
f.simFrail <- function(
n=nn, K=nK, k=kap, r=rho, P1=p1, P2=p2, m=mu3, s=sig3, seed=1234,
b=betas, fix.cens=TRUE, ct=CENST, kc=k.cens, rc=r.cens,
frailty, frailty.par=fpar){
        #~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
        writeLines('\nUsing␣Version␣1.0␣of␣function␣f.simFrail')
        writeLines('Last␣update:␣31/Dec/2017')
        #~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
        ## Acummulate according to G
        set.seed(seed)
        g  <- 1
        dt <- survt <- delta <- matrix(rep(NA, n*K), ncol=K)
        Xl <- vector("list", K)
        X  <- NULL
        #~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
        ## Fralties
        f.frailty <- function(frailty, n, p){
                fnofrailty <- function(n,p){rep(as.numeric(1), n)}
                flognormal <- function(n,p){ rlnorm(n, meanlog = 0, sdlog = sqrt(p)
                fgamma     <- function(n,p){ rgamma(n, shape = 1/p, scale = p) }
                #.....................................................................
                flist <- list("none"=fnofrailty, "lognormal"=flognormal, "gamma"=fg
                d <- match(frailty, names(flist))
                return(flist[[d]](n = n, p = p))
        }
        u <- f.frailty(frailty=frailty, n=n, p=frailty.par)
        #~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
        ## Censoring time & Survival times from Weibull(kappa,rho)
        if(fix.cens){
                cat("\nFixed␣censoring␣time␣=",ct)
        } else {
                cat("\nCensoring␣process␣follows␣a␣Weibull␣distribution␣with␣parame
                cat("shape␣=", kc, "rate␣=", rc)
        }
        while(g <= K){
                #.....................................................
```

```
                ## Produce the covariance matrix, X
                x1 <- (rbinom(n,1,P1))
                x2 <- (rbinom(n,1,P2))
                x3 <- round(rnorm(n, mean=m, sd=s),1)
                Xg  <- cbind(x1,x2,x3)
                #.....................................................
                v  <- runif(n, 0,1)  ## Censoring time
                z  <- runif(n, 0,1)  ## Survival time
                censt <- ifelse(fix.cens,
                ct,
                1/rc*(-log(v))^(1/kc) )
                survt[, g]   <- 1/(r*u)*( -log(z)*1/exp(Xg %*% b) )^(1/k)
                dt[, g]      <- pmin(survt[, g], censt)
                delta[, g]   <-  1*(survt[, g] <= censt)
                colnames(Xg) <- paste0(colnames(Xg),"_",g)
                X <- cbind(X, Xg)
                #.....................................................
                g <- g+1
                #.....................................................
        }
        survt <- data.frame(survt)
        dt    <- data.frame(dt)
        delta <- data.frame(delta)
        X     <- data.frame(X)
        if(K==1){
                names(survt) <- "survt"; names(dt) <- "dt"; names(delta)<-"delta"
                names(X) <- paste0("x", 1:3)
        } else {
                names(survt) <- paste0("survt", 1:K)
                names(dt)    <- paste0("dt", 1:K)
                names(delta) <- paste0("delta", 1:K)
        }
        #~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
        ## Survival outcome is min(survt, censt)
        simSurv <- data.frame(cbind(
        id = seq(1:n), survt, dt, delta, X ) )
}
#============================================================

# nn      <- 100  ## number of clusters
# nK      <- 2    ## number of repetitions for each i
```

```
# kap     <- 2    ## hazard ~ Weibull(shape=kap, rate=rho)
# rho     <- 0.5  ## hazard ~ Weibull(shape=kap, rate=rho)
# p1      <- 0.5  ## X1 ~ Binomial(n, p1)
# p2      <- 0.5  ## X2 ~ Binomial(n, p2)
# mu3     <- 0    ## X3 ~ N(mean=mu3, sd=sig3)
# sig3    <- 4    ## X3 ~ N(mean=mu3, sd=sig3)
# betas   <- c(0.5, 0.5, 0.5) ## Regression coefficients.
# CENST   <- 5    ## Fixed censoring time for all id
# k.cens  <- 1    ## shape of the censoring process if "fix.cens" = FALSE
# r.cens  <- 0.1  ## rate of censoring process if "fix.cens" = FALSE
# frailty <- "gamma"  ## "lognormal" or "gamma"
# fpar    <- 0.1  ## Parameter of the frailty distribution.

d <- f.simFrail(
  n=nn, K=nK, k=kap, r=rho, P1=p1, P2=p2, m=mu3, s=sig3, seed=1234,
  b=betas, fix.cens=TRUE, ct=CENST, kc=k.cens, rc=r.cens,
  frailty = "none", frailty.par=fpar)
names(d)
par(mfrow=c(1,nK), las=1)
for (g in 1:nK){ hist(d[, 1+g], main=bquote(.(names(d)[1+g])), xlab="time") }
```

A.2 Simulation of Cox model data with endogenous time-varying covariate and parametric baseline hazard

Sample code to simulate data from Proportional Hazards models. Corresponds to Algorithm 2.2 for the Cox model with endongenous time-varying covariate of Section 2.2.1.

```
##=================================================================
## Last update: 23/Aug/2019
## Version 1.0
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## R code to simulate recurrent event times from the
## Andersen-Gill model (includes random effect)
##
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Chapter: 2
## For third year 201819
## (c) Jacob Cancino-Romero (JHD,LB,SB) 2019
##=================================================================
##=================================================================
## Contents
## --------
## Function to simulate recurrent event times from the AG model
##
##=================================================================


##=================================================================
Version     <- 2.0
Last_update <- format(Sys.Date(), "%d␣%B␣%Y")
# Last_update <- "7/September/2019"
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
library(MASS) ## mvrnorm(n, mu, Sigma)
library(survival)

if(!require(actuar)) install.packages("actuar")      ## Shut down in office
if(!require(flexsurv)) install.packages("flexsurv") ## Shut down in office
if(!require(VGAM)) install.packages("VGAM")          ## Shut down in office
```

```
library(actuar)          ## log-logistic            ## Shut down in office
library(flexsurv)        ## gompertz               ## Shut down in office
library(VGAM)            ## makeham                ## Shut down in office

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Time functions
ft <- function(x, fnt, p){
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  fn.names <- c("linear", "banana", "cos-sin", "logistic", "normal")
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  f1 <- function(x,p){
    p[1] + p[2]*x }
  f2 <- function(x,p){
    p[1] + p[2]*x + p[3]*x^3 + p[4]*exp(p[5]*x^2) }
  f3 <- function(x,p){
    p[1] + p[2]*x + p[3]*sin(p[4]*x) + p[5]*cos(p[6]*x) +
      p[7]*(x+p[8])^2 + p[9]*exp(p[10]*x)}  ##
  f4 <- function(x,p){
    p[2] / (1 + exp(-p[3]*(x-p[1])))}
  f5 <- function(x, p){
    p[1]*x + p[2]*dnorm(x, mean=p[3], sd=p[4])}
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  flist <- list("linear"=f1, "banana"=f2, "cos-sin"=f3, "logistic"=f4,
                "normal"=f5)
  d <- match(fnt, fn.names)
  return(flist[[d]](x=x,p=p))
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Baseline hazard
h0 <- function(x, sdist, g){
  ## Simulate survival times.
  ## Baseline hazard. Using pweibull()/deweibull() might cause
  ## computational difficulties as time gets larger (0/0 = NaN)
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  d.names <- c("weibull", "log-logistic", "gompertz", "makeham", "bathtub")
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  h1 <- function(x,g){ g[1]*g[2]*(g[2]*x)^(g[1]-1) }
  h2 <- function(x,g){ g[1]*g[2]*(g[2]*x)^(g[1]-1) / (1+g[2]*x)^g[1]}
  h3 <- function(x,g){ g[1]*exp(g[2]*x) }
  h4 <- function(x,g){ g[1] + g[2]*exp(-g[3]*x)}
```

```
  h5 <- function(x,g){ g[1]*x + g[2]/(1 + g[3]*x) }
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  hlist <- list(h1, h2, h3, h4, h5)
  names(hlist) <- d.names
  distrib <- match(sdist, d.names)
  return(hlist[[distrib]](x=x,g=g))
}
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Subject specific hazard rate
hx <- function(x, fnt, p, betas, b, w, etaL, gammas, etaR=0, N=0,
               sdist, g, v){
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  ## h(t) = h_{0}(t) * exp{ v + eta*m(t) + etaR*N + w^{T}gammas }
  ## v = log(frailty), where frailty is an instance of the
  ## gamma or log-normal distribution.
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  gammas <- as.vector(gammas)
  w      <- as.vector(w)
  exp(etaL*mx(x, fnt=fnt, p=p, w=w, betas=betas, b=b) +
        v + as.numeric(crossprod(gammas,w))) *
    h0(x, sdist=sdist, g=g)
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Random sample from the frailties distribution.
f.frailty <- function(n, fpar, fdist){
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  frailty.dist <- c("gamma", "log-normal")
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  f1    <- function(n,fpar){rgamma(n, shape=1/fpar, scale=fpar)}
  f2    <- function(n,fpar){rlnorm(n, meanlog=0, sdlog=sqrt(fpar))}
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  flist <- list("gamma"=f1, "log-normal"=f2)
  d     <- match(fdist, frailty.dist)
  return(flist[[d]](n=n,fpar=fpar))
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Cumulative hazard. Notice different vectorization method via sapply():
cumhaz <- function(
  x, fnt, p, betas, b, w, etaL, gammas, etaR=0, N=0, sdist, g, v, LOWER){
```

```
  sapply(x, function(x)
    integrate(hx, lower=LOWER, upper=x,
              fnt=fnt, p=p, betas=betas, b=b, w=w, etaL=etaL, etaR=etaR,
              gammas=gammas, N=N, sdist=sdist, g=g, v=v)$value)
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Conditional cumulative hazard with frailty
cHaz <- function(x, t, fnt, p, betas, b, w, etaL, gammas, etaR=0, N=0,
                 sdist, g, v, LOWER){
  cumhaz(x=x+t, fnt, p, betas, b, w, etaL, gammas, etaR=0, N=0,
         sdist, g, v, LOWER) -
    cumhaz(x=t,  fnt, p, betas, b, w, etaL, gammas, etaR=0, N=0,
           sdist, g, v, LOWER)
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Inverse of the cumulative hazard
icumhaz <- function(u, fnt, p, betas, b, w, etaL, gammas, etaR=0, N=0,
                    sdist, g, v, LOWER, FROM=0, TO, shift, tolexp=0.5){
  ## u stands for S(t). Make sure 0<=u<= 1
  phi <- (-log(u))
  h   <- function(x){
    cumhaz(x, fnt=fnt, p=p, betas=betas, b=b, w=w, etaL=etaL, gammas=gammas,
           etaR=etaR, N=N, sdist=sdist, g=g, v=v, LOWER=LOWER) + shift - phi
  } ## Must subtract phi, not u!
  x <- uniroot(h, interval=c(FROM, TO), tol=.Machine$double.eps^tolexp)$root
  return(x)
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Inverse of the conditional cumulative hazard with frailty
icHaz <- function(u, t, fnt, p, betas, b, w, etaL, gammas, etaR=0, N=0,
                  sdist, g, v, LOWER, FROM=0, TO, tolexp=0.5){
  ## u stands for S(t). Make sure 0<=u<= 1
  phi <- (-log(u))
  h   <- function(x){
    cHaz(x, t=t, fnt=fnt, p=p, betas=betas, b=b, w=w, etaL=etaL, gammas=gammas,
         etaR=etaR, N=N, sdist=sdist, g=g, v=v, LOWER=LOWER)-phi
  } ## Must subtract phi, not u!
  x <- uniroot(h, interval=c(FROM, TO), tol=.Machine$double.eps^tolexp)$root
```

```
  return(x)
}

#-----------------------------------------------------------------
f.simAG <- function(seed, n, nsim,
                    p1=0.5, p2=0.5, p3=0.5, p4=0.5, S_567.rand = FALSE,
                    mu_567     = c(0,0,0),
                    S_567      = diag(c(1,1,1)),
                    nRecEv     = 20,  ## How many recurrent events per subject?
                    ran.censt  = FALSE,
                    censt      = 1,  ## censoring time
                    sdist      = "weibull", ## "weibull","log-logistic","gamma","go
                    g          = c(4, 1.5),  ## c(shape, rate)
                    gammas     = rep(-0.1, 7),
                    frailty    = FALSE,
                    fpar       = 0.4^2 ,       ## Accounts for the variance of the
                    fdist      = "log-normal", ## Distribution of frailty
                    tolexp     = 0.5,
                    savefile   = FALSE,
                    outdir     = getwd(),
                    outfile    = "simAG.txt"
){
  cat('================================================================\n')
  cat('Function␣simAG(),␣version', Version, '\n')
  cat('Last␣update:', Last_update,'\n')
  cat('file␣name:␣f.simAGmodel.R␣\n')
  cat('For␣tests␣and␣details␣check␣file␣f.simAGmodel(develop).R\n')
  # Create empty datasets to acummulate the data of a number of simulations
  simRec <- NULL
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  set.seed(seed) ## sample(1:100, 1)
  # Generate covariates once and keep the values fixed for all simulations
  x1 <- (rbinom(n,1,p1))  # binary. Alternative, LaplacesDemon::rbern
  x2 <- (rbinom(n,1,p2))  # binary
  x3 <- (rbinom(n,1,p3))  # binary
  x4 <- (rbinom(n,1,p4))  # binary
  X1 <- mvrnorm(n, mu=mu_567, Sigma=S_567) # continuous, potentially correlated

  mvrnorm(n, mu=mu_567, Sigma=S_567)
  colnames(X1) <- c("x5","x6","x7")
  X    <- cbind(x1,x2,x3,x4,X1)
```

```
ID   <- seq(1,n)
temp <- data.frame(id=ID, X)

cat('--------------------------------------------------------------------\n')
# Three levels of "for" loops: s=simulations, ID=individuals, k=recurrent events
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
for (s in 1:nsim){ ## s<-1
  set.seed(seed) ## sample(1:100, 1)
  dRec <- NULL
  cat('sim = ', s, 'out of', nsim, '--> ')
  # ..........................................................
  ## Define censoring and random effects
  ## Censoring
  if (ran.censt){
    censi <- runif(n, min=0, max=censt) } else { censi <- rep(censt, n)
    }
  # Random effects
  if (frailty){
    z <- f.frailty(n=n, fpar=fpar, fdist=fdist)
  }
 # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  for (i in seq_along(ID)){  ## i<-1
    ## Recurrent event process
    R        <- NULL
    CONTINUE <- TRUE
    k      <- 1
    K      <- 0             ## Control the maximum number of rec. events.
    N      <- c(0)          ## Counting process starts at 0
    deltaR <- c(0)          ## Event process starts at 0
    gap    <- gap.c <- c(0) ## time-gaps vector: inter-arrival times
    L      <- c(0)          ## L=Lower end of interval
    U      <- vector(mode='numeric', length(0)) ## U=Upper end of interval
    r_stop <- vector(mode='numeric', length(0))
    u      <- vector(mode='numeric', length(0)) ## runif(1)
    while(CONTINUE){
      u[k]   <- runif(1)
      gap[k] <- icHaz(u=u[k], sdist=sdist, g=g, w=X[i,], gammas=gammas,
                      z=z[i], t=L[k], FROM=0, TO=1e3)
      # gap[k] <- icHazWeib(u=u, k=g[1], r=g[2], w=X[i,], gammas=gammas, t=L[k],
      #                 FROM=0, TO=1e3)
      U[k]      <- L[k] + gap[k]
```

```
      r_stop[k] <- min(U[k], censi[i])
      deltaR[k] <- 1*( N[k] > 0 )
      censored  <- ( U[k] > censi[i] )
      gap.c[k]  <- ifelse( !censored, gap[k], r_stop[k]-U[k-1])
       #.....................................................
      ## Collect the data here!
      R <- rbind(R, cbind(
        sim=s, seed=seed, id=ID[i], u=u[k], frailty=z[i], deltaR=deltaR[k], N=N[k
        censt=censi[i], gap=gap[k], gap.c=gap.c[k], r.time=L[k], L=L[k],
        U=U[k], r_start=L[k], r_stop=r_stop[k]
      ) )
      #.....................................................
      ## Update values. If CONTINUE = FALSE, these will be reset.
      N <- c(N, N[k]+1)
      L <- c(L, U[k])
      k <- length(N)
      K <- K <- K+1
      CONTINUE <- (K <= nRecEv) && (!censored)
    } ##----------while
    ## Merge with covariates and acumulate records of all id of simulation s
    dRec <- rbind(dRec, data.frame(R))
  }  ##----------id
  dRec   <- merge(x=dRec, y=temp, by='id')
  simRec <- rbind(simRec, dRec)
  seed   <- seed + 1
  cat('Finished\n')
} ##-------------sim
cat('Preparing', length(outfile), 'files.\n')
simRec <- cbind(subset(simRec, select=sim), subset(simRec, select=-sim))
simList <- list(simRec=simRec)
if(savefile){
  setwd(outdir)
  for (f in seq(simList)){ ##f<-1
    write.table(simList[[f]], outfile[f], row.names=FALSE, col.names=TRUE, sep='
  }
  cat('-----------------------------------------------------------------------\n')
  cat('Output files:', paste0(outfile, collapse=", "), '\n')
  cat('Location:', outdir, '\n')
  cat('=======================================================================\n')
}
return(simList) }
```

A.3 Simulation of recurrent events with the Andersen–Gill model

Sample code to simulate a data set with recurrent events and covariates from the Andersen–Gill model of Section 2.2.2. Uses Algorithm 2.3 to simulate the recurrent events.

```
##=================================================================
## Last update: 23/Aug/2019
## Version 1.0
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## R code to simulate recurrent event times from the
## Andersen-Gill model (includes random effect)
##
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Chapter: 2
## For third year 201819
## (c) Jacob Cancino-Romero (JHD,LB,SB) 2019
##=================================================================
##=================================================================
## Contents
## --------
## Function to simulate recurrent event times from the AG model
##
##=================================================================


##=================================================================
Version     <- 2.0
Last_update <- format(Sys.Date(), "%d␣%B␣%Y")
# Last_update <- "7/September/2019"
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
library(MASS) ## mvrnorm(n, mu, Sigma)
library(survival)

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Baseline hazard. Note exp(etaT*x), so can include time-varying covs:
h0 <- function(x, sdist, g){
```

```
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  d.names <- c("weibull", "log-logistic", "gompertz", "makeham", "bathtub")
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  h1 <- function(x,g){ g[1]*g[2]*(g[2]*x)^(g[1]-1) }
  h2 <- function(x,g){ g[1]*g[2]*(g[2]*x)^(g[1]-1) / (1+g[2]*x)^g[1]}
  h3 <- function(x,g){ g[1]*exp(g[2]*x) }
  h4 <- function(x,g){ g[1] + g[2]*exp(-g[3]*x)}
  h5 <- function(x,g){ g[1]*x + g[2]/(1 + g[3]*x) }
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  hlist <- list(h1, h2, h3, h4, h5)
  names(hlist) <- d.names
  distrib <- match(sdist, d.names)
  return(hlist[[distrib]](x=x,g=g))
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Baseline cumulative baseline hazard: H_{0}(t) = -log(S_{0}(t))
H0 <- function(x, g, sdist){
  sapply(x, function(x)
    integrate(h0, lower=0, upper=x,
              sdist=sdist, g=g)$value)
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Random sample from the frailties distribution.
f.frailty <- function(n, fpar, fdist){
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  frailty.dist <- c("gamma", "log-normal")
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  f1    <- function(n,fpar){rgamma(n, shape=1/fpar, scale=fpar)}
  f2    <- function(n,fpar){rlnorm(n, meanlog=0, sdlog=sqrt(fpar))}
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  flist <- list("gamma"=f1, "log-normal"=f2)
  d     <- match(fdist, frailty.dist)
  return(flist[[d]](n=n,fpar=fpar))
}
v.frailty <- Vectorize(f.frailty) ## Vectorized version

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Cumulative hazard with frailty
cumhaz <- function(x, w, gammas, sdist, g=g, z){
```

```
  ## z is an instance of the frailty distribution.
  sapply(x, function(x)
    H0(x, sdist=sdist, g=g)*exp(as.numeric(crossprod(w,gammas)) + log(z)))
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Conditional cumulative hazard with frailty
cHaz <- function(x, t, w, gammas, sdist, g=g, z){
  cumhaz(x=x+t, w, gammas, sdist, g=g, z=z) -
    cumhaz(x=t, w, gammas, sdist, g=g, z=z)
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Inverse of the cumulative hazard
icumhaz <- function(u, sdist, g, w, gammas, z, FROM=0, TO, tolexp=0.5){
  ## u stands for S(t). Make sure 0<=u<= 1
  phi <- (-log(u))
  h   <- function(x){
    cumhaz(x, sdist=sdist, g=g, w=w, gammas=gammas, z=z)-phi
  } ## Must subtract phi, not u!
  x <- uniroot(h, interval=c(FROM, TO), tol=.Machine$double.eps^tolexp)$root
  return(x)
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Inverse of the conditional cumulative hazard with frailty
icHaz <- function(u, t, sdist, g, w, gammas, z, FROM=0, TO, tolexp=0.5){
  ## u stands for S(t). Make sure 0<=u<= 1
  phi <- (-log(u))
  h   <- function(x){
    cHaz(x, sdist=sdist, g=g, w=w, gammas=gammas, z=z, t=t)-phi
  } ## Must subtract phi, not u!
  x <- uniroot(h, interval=c(FROM, TO), tol=.Machine$double.eps^tolexp)$root
  return(x)
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Weibull cumulative hazard
Haz_Weib <- function(x, k, r, w, gammas, z){
  lp  <- as.numeric(crossprod(w,gammas))+log(z)
  Haz <- (r*x)^k * exp(lp)
```

```
  return(Haz)
}
## Weibull inverse of the conditional cumulative hazard
icHaz_Weib <- function(u, t, k, r, w, gammas, z){
  phi <- -log(u)
  lp  <- as.numeric(crossprod(w,gammas))+log(z)
  gap <- 1/r * ( phi*exp(-lp) + (r*t)^k )^(1/k) - t
  return(gap)
}
##================================================================
## Simulator
# Version     <- 2.0
# Last_update <- format(Sys.Date(), "%d %B %Y")
#
# seed=0; n=5; nsim=2; p1=0.5; p2=0.5; p3=0.5; p4=0.5; S_567.rand = FALSE;
# mu_567      = c(0,0,0)
# S_567       = diag(c(1,1,1))
#
# nRecEv      = 20  ## How many recurrent events per subject?
# ran.censt   = FALSE
# censt       = 1   ## censoring time
# sdist       = "weibull"  ## "weibull","log-logistic","gamma","gompertz","makeham"
# g           = c(4, 1.5)   ## c(shape, rate)
# gammas      = rep(0, 7)
# frailty     = TRUE
# fpar        = 0.4^2       ## Accounts for the variance of the frailty.
# fdist       = "log-normal" ## Distribution of frailty
# tolexp      = 0.5
# savefile    = FALSE
# outdir      = getwd()
# outfile     = "simAG.txt"
#----------------------------------------------------------------
f.simAG <- function(first.event.in.row.1 = FALSE,
                    seed, n, nsim,
                    p1=0.5, p2=0.5, p3=0.5, p4=0.5, S_567.rand = FALSE,
                    mu_567      = c(0,0,0),
                    S_567       = diag(c(1,1,1)),
                    nRecEv      = 20,  ## How many recurrent events per subject?
                    ran.censt   = FALSE,
                    censt       = 1,   ## censoring time
                    sdist       = "weibull", ## "weibull","log-logistic","gamma","go
```

```
                    g          = c(4, 1.5),  ## c(shape, rate)
                    gammas     = rep(-0.1, 7),
                    frailty    = FALSE,
                    fpar       = 0.4^2 ,       ## Accounts for the variance of the
                    fdist      = "log-normal", ## Distribution of frailty
                    tolexp     = 0.5,
                    savefile   = FALSE,
                    outdir     = getwd(),
                    outfile    = "simAG.txt"
){
  cat('================================================================\n')
  cat('Function␣simAG(),␣version', Version, '\n')
  cat('Last␣update:', Last_update,'\n')
  cat('file␣name:␣f.simAGmodel.R␣\n')
  cat('For␣tests␣and␣details␣check␣file␣f.simAGmodel(develop).R\n')
  # Create empty datasets to acummulate the data of a number of simulations
  simRec <- NULL
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  set.seed(seed) ## sample(1:100, 1)
  # Generate covariates once and keep the values fixed for all simulations
  x1 <- (rbinom(n,1,p1))  # binary. Alternative, LaplacesDemon::rbern
  x2 <- (rbinom(n,1,p2))  # binary
  x3 <- (rbinom(n,1,p3))  # binary
  x4 <- (rbinom(n,1,p4))  # binary
  X1 <- mvrnorm(n, mu=mu_567, Sigma=S_567) # continuous, potentially correlated

  mvrnorm(n, mu=mu_567, Sigma=S_567)
  colnames(X1) <- c("x5","x6","x7")
  X    <- cbind(x1,x2,x3,x4,X1)
  ID   <- seq(1,n)
  temp <- data.frame(id=ID, X)

  cat('----------------------------------------------------------------\n')
  # Three levels of "for" loops: s=simulations, ID=individuals, k=recurrent events
  # ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
  for (s in 1:nsim){ ## s<-1
    set.seed(seed) ## sample(1:100, 1)
    dRec <- NULL
    cat('sim␣=', s, 'out␣of', nsim, '-->␣')
    # ..........................................................
    ## Define censoring and random effects
```

```
 ## Censoring
 if (ran.censt){
   censi <- runif(n, min=0, max=censt) } else { censi <- rep(censt, n)
   }
 # Random effects
 z <- 1 + double(n)
 if (frailty){
   z <- f.frailty(n=n, fpar=fpar, fdist=fdist)
 }
#  ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 for (i in seq_along(ID)){  ## i<-1
   #~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
   ## Recurrent event process
   R        <- NULL
   CONTINUE <- TRUE
   k      <- 1
   K      <- 0             ## Control the maximum number of rec. events.
   N      <- c(0)          ## Counting process starts at 0
   deltaR <- c(0)          ## Event process starts at 0
   gap    <- gap.c <- c(0) ## time gaps with closed form of Weibull conditional
   gap_Weib <- gap_Weib.c <- c(0)## time-gaps (inter-arrival times) with root f
   L      <- c(0)          ## L=Lower end of interval
   U      <- vector(mode='numeric', length(0)) ## U=Upper end of interval
   r_stop <- vector(mode='numeric', length(0))
   u      <- vector(mode='numeric', length(0)) ## runif(1)
   phi    <- vector(mode='numeric', length(0)) ## runif(1)

   while(CONTINUE){
     u[k]   <- runif(1)
     phi[k] <- -log(u[k]);
     gap[k] <- icHaz(u=u[k], sdist=sdist, g=g, w=X[i,], gammas=gammas,
                     z=z[i], t=L[k], FROM=0, TO=1e3)
     gap_Weib[k] <- icHaz_Weib(u=u[k], t=L[k], k=g[1], r=g[2], w=X[i,],
                               gammas=gammas, z=z[i])
     U[k]      <- L[k] + gap[k]
     r_stop[k] <- min(U[k], censi[i])
     #.....................................................
     #' If the recurrent event process starts with the 1st event in
     #' the 1st row (deltaR=1) then tha last one will be censored.
     #' In such a case "first.event.in.row.1 = TRUE".
     #' This simulator was origininally written considering
```

```
    #' "first.event.in.row.1 = FALSE". This is, the 1st
    #' event is recorded "after" it had occurred, so in the time interval
    #' [0,t1), deltaR_1 = 0, and in [t1,t2) delta_R_2 = 1, unless
    #' tj > censt, in which case deltaR_j = 0.
    #' Also N below should be replaced
    if (first.event.in.row.1){
      deltaR[k] <- 1*( r_stop[k] < censi[i] )
      N[k]      <- ifelse(deltaR[k] == 1, N[k]+1, N[k])
    } else {
      deltaR[k] <- 1*( N[k] > 0 )
      # deltaR[k] <- 1*( r_start[k] < censi[i] )
    }
    #.....................................................
    censored  <- ( U[k] > censi[i] )
    gap.c[k]  <- ifelse( !censored, gap[k], r_stop[k]-L[k])
    gap_Weib.c[k] <- ifelse( !censored, gap_Weib[k], r_stop[k]-L[k])
    #.....................................................
    ## Collect the data here!
    R <- rbind(R, cbind(
      sim=s, seed=seed, id=ID[i], u=u[k], phi=phi[k], frailty=z[i],
      deltaR=deltaR[k], N=N[k], censt=censi[i], gap=gap[k], gap.c=gap.c[k],
      gap_Weib=gap_Weib[k], gap_Weib.c=gap_Weib.c[k],
      r.time=L[k], L=L[k], U=U[k], r_start=L[k], r_stop=r_stop[k]
    ) )
    #.....................................................
    ## Update values. If CONTINUE = FALSE, these will be reset.
    if (first.event.in.row.1){
      N <- c(N, N[k])
    } else {
      N <- c(N, N[k]+1) ## <------- Replace for N <- c(N, N[k])
    }
    L <- c(L, U[k])
    k <- length(N)
    K <- K <- K+1
    ## Stop the data generating process if K has reached the RecEv limit
    ## or if the last recurrent event occurred after the censoring time.
    ## Note that K started at 0 and is updated after the kth iteration,
    ## so (K <= nRecEv) means we will generate K complete event times.
    CONTINUE <- (K <= nRecEv) & (!censored)
  } ##----------while
  #~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

```
      ## Merge with covariates and acumulate records of all id of simulation s
      dRec <- rbind(dRec, data.frame(R))
      i<-i+1
    }  ##-----------id
    dRec   <- merge(x=dRec, y=temp, by='id')
    simRec <- rbind(simRec, dRec)
    seed   <- seed + 1
    cat('Finished\n')
  } ##--------------sim
  cat('Preparing', length(outfile), 'files.\n')
  simRec <- cbind(subset(simRec, select=sim), subset(simRec, select=-sim))
  simList <- list(simRec=simRec)
  if(savefile){
    setwd(outdir)
    for (f in seq(simList)){ ##f<-1
      write.table(simList[[f]], outfile[f], row.names=FALSE, col.names=TRUE, sep='
    }
    cat('------------------------------------------------------------------------\n')
    cat('Output files:', paste0(outfile, collapse=", "), '\n')
    cat('Location:', outdir, '\n')
    cat('========================================================================\n')
  }
  return(simList)
}
##=============================================================
```

A.4 Brier Score and Integrated Brier Score

Sample code to compute the Brier Score and Integrated Brier Score explained in Section 2.3.1

```
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#' Last update: 9/July/2020
#' Version 1.0
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#' Brier Score and Integrated Brier Score
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#' Chapter: 2
#' For second year 201718
#' (c) Jacob Cancino-Romero (JHD,LB,SB) 2020
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#' Contents
#' ~~~~~~~~
#'
#' BS, IBS and W(t) function for weighting the contribuitons when data
#' are censored.
#' The file explain with a series of functions how to compute the BS & IBS
#' for the Cox model
#'
#'   (a) only baseline covariates
#'   (b) with an endogenous time-varying covariate
#'
#' It has a weight function to accommodate censoring.
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Version     <- 2.0
Last_update <- format(Sys.Date(), "%d␣%B␣%Y")
# Last_update <- "7/September/2019"
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
library(MASS) ## mvrnorm(n, mu, Sigma)
library(survival)
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# IBS (Cox model) ---------------------------------------------
```

```
dT <- 1.5
p  <- c(4,1)
w  <- c(1,1,1,1)
g  <- c(-0.5, -0.5, -0.5, -0.5)
A  <- 0
B  <- 3

#' Baseline survival function

S0 <- function(x,p){exp(-(p[2]*x)^p[1])}
curve(S0(x,p), A,B)
#' Survival function

S <- function(x,p,w,g){
lp <- as.numeric(crossprod(w,g))
S <- S0(x,p)^exp(lp)
return(S)
}
curve(S(x,p,w,g), A,B, col=4, lwd=2, add=T)

#' 1(T > x) function

f1 <- function(x,dT){
f <- function(x){ifelse((dT > x), 1, 0)}
sapply(x,f)
}
f1(seq(0,10), 5)

#' Function to plot f1()

plot.f1 <- function(x,dT,from,to){
xseq  <- seq(from=from, to=to, len=11)
steps <- stepfun(xseq, c(1, f1(x=xseq,dT)))
plot(steps, vertical=TRUE, pch=NA, lty=2, col=8, col.hor=NA, add=T)
plot(steps, pch=NA, col=3, lwd=2, vertical=FALSE, add=T)
}
plot.f1(x,dT,A,B)

#' Integral

int <- function(dT, p,w,g, A,B){
```

```
fl <- function(x,dT, p,w,g){(f1(x,dT)-S(x,p,w,g))}
up <- integrate(fl, lower=A, upper=dT, dT,p,w,g)$value
lo <- integrate(S, lower=dT, upper=B, p,w,g)$value
int <- up + lo
return(list(upper=up, lower=lo, integral=int))
}
int(dT=1,p,w,g,A,B)
integrate(S, lower=A, upper=B, p,w,g)

#' Brier score (BS)

bscore <- function(x, dT, p,w,g, A,B){
bscore <-
(f1(x,dT)==1) * (f1(x,dT) - S(x,p,w,g))^2 +
(f1(x,dT)==0) * (S(x,p,w,g) - f1(x,dT))^2
return(bscore)
}
x <- 1
bscore(x,dT,p,w,g,A,B)
(f1(x,dT) - S(x,p,w,g))^2

#' Integrated Brier score (IBS)

ibs <- function(dT, p,w,g, A,B){
ibs <- integrate(bscore, A,B, dT,p,w,g)$value
return(ibs)
}
xseq <- seq(A,B,len=1e4)
rbind(ibs(dT, p,w,g, A,B),
sum(bscore(xseq,dT,p,w,g,A,B)*diff(xseq)[1]))

ibs(dT,p,w,g,A=1,B=2)
ibs(dT=1, p,w,g, A,B)
int(dT=1, p,w,g, A,B)
S2 <- function(x,p,w,g){S(x,p,w,g)^2}
ibs(dT=0,p,w,g,A,B)
ibs(dT=B,p,w,g,A,B=B)
integrate(S2, lower=A, upper=B, p,w,g)

# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# IBS (Cox with time-varying covariate m(t) ---------------------
```

```
b    <- c(1,0.5,3)
eta <- 1

#' Time-varying covariate

mt <- function(x,b){
mt <- b[1] + b[2]*x + cos(b[3]*x)
return(mt)
}
curve(mt(x,b), A,B)

#' Baseline hazard

h0 <- function(x,p){p[1]*p[2]*(p[2]*x)^(p[1]-1)}
curve(h0(x,p), A,B*2)

#' Hazard rate

ht <- function(x,p,w,g,b,eta){
lp <- as.numeric(crossprod(w,g)) + eta*mt(x,b)
ht <- h0(x,p)*exp(lp)
}
curve(ht(x,p,w,g,b,eta), col=2, lwd=2, add=T)

#' Cumulative hazard

Ht <- function(x,p,w,g,b,eta,from=0){
Ht <- function(x){
integrate(ht, lower=from, upper=x, p,w,g,b,eta)$value
}
sapply(x,Ht)
}
curve(Ht(x,p,w,g,b,eta), col=2, lwd=2)

St <- function(x,p,w,g,b,eta,from=0){
exp(-Ht(x,p,w,g,b,eta))
}
curve(St(x,p,w,g,b,eta), col=2, lwd=2, A,B)
curve(St(x,p,w,g,b,eta=0), col=2, lwd=2, A,B)
curve(S(x,p,w,g), col=3, lwd=2, add=T)
```

```
int_mt <- function(dT, p,w,g,b,eta, A,B){
fl <- function(x,dT, p,w,g,b,eta){
(f1(x,dT)-St(x,p,w,g,b,eta))
}
up <- integrate(fl, lower=A, upper=dT, dT,p,w,g,b,eta)$value
lo <- integrate(St, lower=dT, upper=B, p,w,g,b,eta)$value
int <- up + lo
return(list(upper=up, lower=lo, integral=int))
}
int_mt(dT=1,p,w,g,b,eta,A,B)
int_mt(dT=0,p,w,g,b,eta,A,B)
integrate(St, lower=A, upper=B, p,w,g,b,eta)

#' Brier score (BS)

bs_mt <- function(x, dT, p,w,g,b,eta, A,B){
bscore <-
(f1(x,dT)==1) * (f1(x,dT) - St(x,p,w,g,b,eta))^2 +
(f1(x,dT)==0) * (St(x,p,w,g,b,eta) - f1(x,dT))^2
return(bscore)
}
#' Check for a single x-value
x <- 1
bs_mt(x,dT,p,w,g,b,eta,A,B)
(f1(x,dT) - St(x,p,w,g,b,eta))^2

#' Integrated Brier score (IBS)

ibs_mt <- function(dT, p,w,g,b,eta, A,B){
ibs_mt <- integrate(bs_mt, A,B, dT,p,w,g,b,eta)$value
return(ibs_mt)
}
ibs_mt(dT, p,w,g,b,eta, A,B)
ibs_mt(dT, p,w,g,b,eta=0, A,B)
ibs(dT, p,w,g, A,B)
xseq <- seq(A,B,len=1e4)
rbind(ibs_mt(dT, p,w,g,b,eta, A,B),
sum(bs_mt(xseq,dT,p,w,g,b,eta,A,B)*diff(xseq)[1]))

# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

```
# Censoring weights ------------------------------------------

n <- 1e4
l <- 4
W <- matrix(rnorm(l*n), ncol=l); colnames(W) <- paste0("w", 1:l)
g <- rep(0,l)
u <- runif(n)
survt <- 1/p[2] * (-log(u)*exp(-as.numeric(crossprod(t(W),g))))^(1/p[1])
hist(survt, prob=T)
curve(dweibull(x,shape=p[1], scale=1/p[2]),col=4, lwd=2, add=T)

set.seed(2)
n <- 1e3
W <- matrix(rnorm(l*n), ncol=l); colnames(W) <- paste0("w", 1:l)
g <- rep(0.5,l)
u <- runif(n)
survt <- 1/p[2] * (-log(u)*exp(-as.numeric(crossprod(t(W),g))))^(1/p[1])
hist(survt, prob=T)
curve(dweibull(x,shape=p[1], scale=1/p[2]),col=4, lwd=2, add=T)

censt <- runif(n,A,B)
dT <- ifelse(survt <= censt, survt, censt)
hist(dT, prob=T)
curve(dweibull(x,shape=p[1], scale=1/p[2]),col=4, lwd=2, add=T)
wi <- split(W, row(W))
wi[[1]]

delta <- as.numeric(dT == survt)
sum(delta)

library(survival)
kmd <- survfit(Surv(dT,delta)~1, conf.type="none")
kmc <- survfit(Surv(dT,1-delta)~1, data=d, conf.type="none")
par(mfrow=c(1,2))
plot(kmd, ylab="Surival␣probability", xlab="time-to-event", mark.time=TRUE)
plot(kmc, ylab="Censoring␣probability", xlab="time-to-censor", mark.time=TRUE)
par(mfrow=c(1,1))
str(kmc)
kmddata <- data.frame(time=kmd$time, surv=kmd$surv, n.risk=kmd$n.risk,
n.event=kmd$n.event, n.censor=kmd$n.censor)
kmcdata <- data.frame(time=kmc$time, surv=kmc$surv, n.risk=kmc$n.risk,
```

```
n.censor=kmc$n.event, n.event=kmc$n.censor)
head(kmddata,10)
head(kmcdata,10)
tail(kmcdata,10)
plot(kmc)
with(kmcdata, lines(time,surv, col=2))

KMcens <- kmcdata

#' G(t) function: is the Kapla-Meier of the censoring process.
#' Is the probability of being censored at time t, for
#' t = t.star and t=dT

f.Gt <- function(x,KMcens){
D      <- KMcens
f <- function(x){
w_time <- max(which(D[,'time'] <= x))   ## max(i) s.t. dT[i] <= x
Gt     <- D[w_time,'surv']
return(Gt)
}
sapply(x,f)
}

GT <- f.Gt(dT, KMcens)
sum(is.na(GT))
sum(GT==0)
Gt <- f.Gt(B, KMcens)
sum(is.na(Gt))
sum(Gt==0)
tail(KMcens)

cbind(head(GT[order(dT)]), tail(GT[order(dT)]))

#' If at least one entry of GT or Gt is 0, the weight will be NaN.
#' This occurs in GT when the largest dT is a censored, and
#' in Gt when x >= max(dT) and max(dT) is censored.
#' The Weight function will take care of it as long as this zero
#' is replaced by any number.  I replaced it with 1.

t.star <- min(max(dT), 0.75*(B-A)); t.star
```

```
GT <- f.Gt(dT, KMcens)
sum(is.na(GT))
sum(GT==0)
Gt <- f.Gt(t.star, KMcens); Gt
sum(is.na(Gt))
sum(Gt==0)
dT[which(GT==0)]
dT[which(Gt==0)]

f.Gt(dT[which(Gt==0)], KMcens)

#' Weight function: W(t.star, G, dT, delta)

f.Wt <- function(KMcens, t.star, dT, delta){
G.T <- f.Gt(x=dT,KMcens)
G.t <- f.Gt(x=t.star, KMcens)
#.............................................................
#' If larges  event time is censored its G.T and G.t is zero. Make
#' it take the value of 1.  It will not be added because delta will
#' take care of its weight
#.............................................................
G.T <- ifelse(G.T == 0, 1, G.T)
G.t <- ifelse(G.t == 0, 1, G.t)
Wt <- 1*(dT <= t.star)*delta/G.T + 1*(dT > t.star)/G.t
return(list(GT=G.T, Gt=G.t, Wt=Wt))
}
Wt <- f.Wt(KMcens,t.star,dT,delta)
sum(Wt$Wt)

#' Weighted IBS

IBSi <- sapply(1:n, function(i) ibs(dT=dT[i], p, w=wi[[i]], g, A,B) )
IBSi
IBSi.W <- IBSi * Wt$Wt
sum(IBSi.W)

mat <- cbind(IBSi, Wt=Wt$Wt, IBSi.W)
rbind(head(mat), tail(mat))

IBS <- crossprod(IBSi, Wt$Wt)/n; IBS
rbind(IBS, mean(IBSi.W))
```

```
as.matrix(c(IBS=mean(IBSi), W.IBS=IBS))

# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# Censoring weights ------------------------------------------

#' If the Cox model has time-varying covariates, compute the IBS with
#' the ibs_mt() function.
#' The weight function will be exactly the same f.Gt() with the
#' corresponding K-M data for the censoring process.
#'
#'  For joint models, the ibs_mt() function needs small changes to add
#'  the linear mixed model to the mt() function.  Do this by creating
#'  a lists of the relevant parts of the lmm:
#'  yi, Xi, Zi, bi and epsi.  Each element of the lists corresponds to
#'  one subject.
```

A.5 Simulation from a 2 Outcome Joint Model

Sample code to simulate data from model $M^{(S)}$ of Section 5.2.1.

File: `C:\...\00-Myphd\MyRfuns\f.simJoint2.R`

```
#------------------------------------------------------------------
# 2. Simulate data 2-outcome joint model data |
#---------------------------------------------|
#
# "simJoint2" function to simulate data.

library(MASS) # Use for multivariate normal: mvrnorm(n, mu=c(0,0), Sigma=S)
library(mvtnorm) # Use for multivariate normal: rmvnorm(n,mean=c(0,0), sigma=S)
library(plyr) # Use to transofm data set with function "ddply"
source("f.CovMatSim.R")

# beta has one more entry than gamma because of beta8*l.time
simJoint2 = function(
seed=0, nsamp=400, nsim=1, t0=0, dt=0.2, repmax=20,
p1=0.5, p2=0.5, p3=0.5, p4=0.5, mu5=0 ,mu67=c(0,0), S67.rand=TRUE, S67,
gamma=c(0.1, 0.1, 0,   0,0,0,0),
beta= c(0,   0.5, 0.5, 0,0,0,0,0), beta0=3,
kappa=1, rho=1.5, eta=0.5, censt=5.5,
var.eps=4, var.b=4
){
set.seed(seed)

# Create empty datasets to acummulate the data of a number of simulations
simLong <- simSurv <- NULL

# Produce a covariance matrix with high correlations for covariates
if(S67.rand){
S <- pdRmt(vars=c(2,2), min.rho=0.9, mseed=sample(1:1e2, 1))$S; S # covariance mati
} else {S <- S67}

# Time-fixed covariates. Produce the values once and keep them constant
# for all simulations
x1 <- (rbinom(nsamp,1,p1))  # binary. Alternative, LaplacesDemon::rbern
```

```
x2 <- (rbinom(nsamp,1,p2))  # binary
x3 <- (rbinom(nsamp,1,p3))  # binary
x4 <- (rbinom(nsamp,1,p4))  # binary
x5 <- rnorm(nsamp, mean=mu5, sd=1) # continuous
X1 <- rmvnorm(nsamp,mean=mu67, sigma=S) # continuous and highly correlated

colnames(X1) <- c("x6","x7")

iind <- seq(1, nsamp)
# Two levels of "for" loops: s=simulations and iind=individuals
for(s in 1:nsim){
set.seed(seed+s)
b <- rnorm(nsamp, mean=0, sd=sqrt(var.b)) # Random intercept
u <- runif(nsamp, 0,1) # Use to simulate the survival times

temp <- data.frame(sim = s,
id  = iind,
cbind(x1,x2,x3,x4,x5,X1),
b,
u,
censt)

beta <- as.vector(beta)
gamma <- as.vector(gamma)

# Simulate survival times from Weibull(kappa,rho)
W <- as.matrix(temp[,c(3:9)])
temp$survt <- 1/rho*( -log(u)*1/exp(W %*% gamma + eta*b) )^(1/kappa)

# Survival outcome is min(survt, censt)
temp <- transform(temp,
delta = 1*(survt <= censt),  ## If also recurrent events, 'deltaD'
t = pmin(survt, censt)       ## 'time' is a better name
)

# Repeated measures & terminal event
jind <- with(temp, pmin(repmax-1, floor((t - t0) / dt)))

# Expand the data frame repeating each id as many times as "jind" says
datlong <- temp[rep(row.names(temp), times=jind+1),]
```

```
datlong = ddply(datlong, "id", here(transform),
obstime = seq(1,length(id)),
l.time = dt * (seq(1,length(id))-1),
epsilon = rnorm(length(id), mean=0, sd=sqrt(var.eps)))

# Time-varying covariates
#    datlong = ddply(datlong, "id", here(transform),
#                    x8 = 0.5 * l.time + rnorm(length(id), 0, 1),
#                    x9 = 0.5 * l.time^2 + rnorm(length(id), 0, 1))

# Compute the longitudinal outcome

X <- as.matrix(datlong[,c(colnames(W),"l.time")])

datlong$y <- with(datlong, (beta0+b) + X %*% beta + epsilon)

# Acummulate the repeated measures and the survival times
simLong <- rbind(simLong,datlong)
simSurv <- rbind(simSurv,temp)

} # End of "for s" loop
write.table(simLong, "simLong.txt", row.names=F, col.names=T, sep="\t")
write.table(simSurv, "simSurv.txt", row.names=F, col.names=T, sep="\t")
returnlist = list(simLong=simLong, simSurv=simSurv)
}
#----------------------END OF FUNCTION simJoint2()--------------------------|
```

A.6 Simulation from a 3 Outcome Joint Model

Sample code to simulate data from model M_3 of Section 6.3. Corresponds to Algorithm 6.1.

```
##=================================================================
## Last update: 30/Sep/2019
## Version 4.1
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## R code to simulate from a 3 outcome joint model with
## baseline covariates and link function = random effects:
##
##  - Longitudinal outcome
##  - Recurrent events: user defined distribution
##  - Terminal Event
##
## The recurrent events are generated according to the
## Andersen-Gill model with inter-event from any distribution
## provided we specify H(t).
##
## The file "f.simAGmodel_td(develop).R" is a general formulation
## to simulate recurrent events times with an endogenous time-varying
## covariate and link = current value.
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Chapter: 2 and 5
## For third year 201819
## (c) Jacob Cancino-Romero (JHD,LB,SB) 2019
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Contents
## ~~~~~~~~
##
##  0) Visualize Weibull density and hazard rate for different scale and rate
##  1) See how the survival outcome will look like
##      2) "simJoint" is the function to simulate the data
##      3) Visualize simulation and check
##
##      See also the sweave file:
##
##                      04_-_Joint_modeling(Sim-Estim).RNW
```

```
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Baseline cumulative baseline hazard no time-varying covariates:
## H_{0}(t) = -log(S_{0}(t))
## See f.simAGmodel.R for more baseline hazard options.
H0 <- function(x, g, rdist){  ## rdist: distribution of baseline hazard
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
dist.names <- c("weibull", "log-logistic")
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
f1 <- function(x,g){
-pweibull(x, shape=g[1], scale=1/g[2], lower=FALSE, log=TRUE)}
f2 <- function(x,g){ log( 1 + (g[2]*x)^g[1] ) }
# f3 <- function(x,g){ ## "gamma" is not proportional hazads
#   -pgamma(q=x, shape=g[1], rate=g[2],lower.tail=FALSE, log.p=TRUE)}
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
flist <- list("weibull" = f1, "log-logistic" = f2)
d <- match(rdist, dist.names)
return(flist[[d]](x=x,g=g))
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Random sample from the frailties distribution.
f.frailty <- function(n, fpar, fdist){
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
frailty.dist <- c("gamma", "log-normal")
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
f1    <- function(n,fpar){rgamma(n, shape=1/fpar, scale=fpar)}
f2    <- function(n,fpar){rlnorm(n, meanlog=0, sdlog=fpar)}
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
flist <- list("gamma"=f1, "log-normal"=f2)
d     <- match(fdist, frailty.dist)
return(flist[[d]](n=n,fpar=fpar))
}
v.frailty <- Vectorize(f.frailty) ## Vectorized version

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Cumulative hazard with frailty
cumhaz <- function(x, w, gammas, b, eta, rdist, g, z){
## z is an instance of the frailty distribution.
sapply(x, function(x)
```

```
H0(x, rdist=rdist, g=g) * exp(
as.numeric(crossprod(w,gammas)) + as.numeric(crossprod(b,eta)) +
log(z)))
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Conditional cumulative hazard with frailty
cHaz <- function(x, t, w, gammas, b, eta, rdist, g=g, z){
cumhaz(x=x+t, w, gammas, b=b, eta=eta, rdist, g=g, z=z) -
cumhaz(x=t, w, gammas, b=b, eta=eta, rdist, g=g, z=z)
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Inverse of the cumulative hazard
icumhaz <- function(u, rdist, g, w, gammas, b, eta, z, FROM=0, TO, tolexp=0.5){
## u stands for S(t). Make sure 0<=u<= 1
phi <- (-log(u))
h   <- function(x){
cumhaz(x, rdist=rdist, g=g, w=w, gammas=gammas, b=b, eta=eta, z=z)-phi
} ## Must subtract phi, not u!
x <- uniroot(h, interval=c(FROM, TO), tol=.Machine$double.eps^tolexp)$root
return(x)
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Inverse of the conditional cumulative hazard with frailty
icHaz <- function(u, t, rdist, g, w, gammas, b, eta, z, FROM=0, TO, tolexp=0.5){
## u stands for S(t). Make sure 0<=u<= 1
phi <- (-log(u))
h   <- function(x){
cHaz(x, rdist=rdist, g=g, w=w, gammas=gammas, b=b, eta=eta, z=z, t=t)-phi
} ## Must subtract phi, not u!
x <- uniroot(h, interval=c(FROM, TO), tol=.Machine$double.eps^tolexp)$root
return(x)
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Weibull cumulative hazard
Haz_Weib <- function(x, g, w, gammas, b, eta, z){
k   <- g[1]
r   <- g[2]
```

```
lp  <- as.numeric(crossprod(w,gammas)) + log(z) +
as.numeric(crossprod(b,eta))
Haz <- (r*x)^k * exp(lp)
return(Haz)
}

## Weibull inverse of the conditional cumulative hazard
icHaz_Weib <- function(u, t, g, w, gammas, b, eta, z){
k   <- g[1]
r   <- g[2]
phi <- -log(u)
lp  <- as.numeric(crossprod(w,gammas)) + log(z) +
as.numeric(crossprod(b,eta))
gap <- 1/r * ( phi*exp(-lp) + (r*t)^k )^(1/k) - t
return(gap)
}

##=================================================================
library(MASS) ## mvrnorm(n, mu, Sigma)
library(survival)

##=================================================================

## Recurrent events generator with Weibull hazards.
f.simRec_Weibull <- function(rdist, g, w, gammas, b, eta, z, dT_cT, nRecEv,
FROM=0, TO=1e3, tolexp=0.5){
#.................................................................
## g      : parameters (a vector) of the baseline hazard
## w      : covariates' vector
## gammas : regresson coefficients of w
## b      : random effects (b0,b1). In this case b1 not multplied by time.
## eta    : regression coefficients of b
## z      : an instance of the frailty  distribution
## dT_ct  : the terminal event or censoring time. Indicates the moment to stop.
## nRecEv : maximum number of recurrent events
#.................................................................
## Recurrent event process
R        <- NULL
CONTINUE <- TRUE
k       <- 1
K       <- 0             ## Control the maximum number of rec. events.
```

```
N       <- c(0)           ## Counting process starts at 0
deltaR  <- c(0)           ## Event process starts at 0
gap     <- gap.c <- c(0) ## time-gaps vector: inter-arrival times
gap_num <- gap_num.c <- c(0) ## time-gaps vector: inter-arrival times
L       <- c(0)           ## L=Lower end of interval
U       <- vector(mode='numeric', length(0)) ## U=Upper end of interval
r_stop  <- vector(mode='numeric', length(0))
u       <- vector(mode='numeric', length(0)) ## runif(1)
phi     <- vector(mode='numeric', length(0)) ## -log(u)
error_RecEv <- FALSE
ERROR  <- "error_RecEv"

while(CONTINUE){
u[k]   <- runif(1)
phi[k] <- -log(u[k])
gap[k] <- icHaz_Weib(
u=u[k], t=L[k], g=g, w=w, gammas=gammas, b=b, eta=eta, z=z)
error_RecEv <- any(tryCatch(  ## This is for the numeric inverse
{gap_num[k] <- icHaz(u=u[k], t=L[k], rdist=rdist, g=g, w=w,
gammas=gammas, b=b, eta=eta, z=z, FROM=FROM, TO=TO)
}, error=function(e){ERROR}) == ERROR )
if(error_RecEv){ ## Increase the resolution of the root finder.
tolexp.new <- 2*tolexp.new
cat('\nError␣generating␣the', k, 'recurrent␣event\n')
cat('Increasing␣resolution␣of␣root␣finder␣to', .Machine$double.eps^(tolexp.new),'\n
gap[k] <- icHaz_Weib(u=u[k], t=L[k], rdist=rdist, g=g, w=w,
gammas=gammas, b=b, eta=eta, z=z)
error_RecEv <- any(tryCatch(
{gap_num[k] <- icHaz(u=u[k], t=L[k], rdist=rdist, g=g, w=w, gammas=gammas,
b=b, eta=eta, z=z, FROM=FROM, TO=TO,tolexp=tolexp.new)
}, error=function(e){ERROR}) == ERROR )
} else {
U[k]      <- L[k] + gap[k]
r_stop[k] <- min(U[k], dT_cT)
deltaR[k] <- 1*( N[k] > 0 )
censored  <- ( U[k] > dT_cT )
gap.c[k]  <- ifelse( !censored, gap[k], r_stop[k]-L[k])
#.....................................................................
## Collect the data here!
R <- rbind(R, cbind(
u=u[k], phi=phi[k], frailty=z, deltaR=deltaR[k], N=N[k],
```

```
dT=dT_cT, gap=gap[k], gap_num=gap_num[k], gap.c=gap.c[k],
r.time=L[k], L=L[k], U=U[k], r_start=L[k], r_stop=r_stop[k]
) )
#.............................................................
## Update values. If CONTINUE = FALSE, these will be reset.
N <- c(N, N[k]+1)
L <- c(L, U[k])
k <- length(N)
K <- K <- K+1
CONTINUE <- (K <= nRecEv) && (!censored)
}
} ##----------while
return(R)
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Recurrent events generator (alternative distributions)

f.simRec <- function(rdist="weibull", g, w, gammas, b, eta, z, dT_cT, nRecEv,
FROM=0, TO=1e3, tolexp=0.5){
#.............................................................
## rdist  : distribution of the baseline hazard
## g      : parameters (a vector) of the baseline hazard
## w      : covariates' vector
## gammas : regresson coefficients of w
## b      : random effects (b0,b1). In this case b1 not multplied by time.
## eta    : regression coefficients of b
## z      : an instance of the frailty  distribution
## dT_ct  : the terminal event or censoring time. Indicates the moment to stop.
## nRecEv : maximum number of recurrent events
## FROM   : left end of the follow-up time (usually 0)
## TO     : right end of the follow-up time (choose a large value)
#.............................................................
## Recurrent event process
tolexp.new <- tolexp
R        <- NULL
CONTINUE <- TRUE
k      <- 1
K      <- 0             ## Control the maximum number of rec. events.
N      <- c(0)          ## Counting process starts at 0
deltaR <- c(0)          ## Event process starts at 0
```

```
gap    <- gap.c <- c(0) ## time-gaps vector: inter-arrival times
L      <- c(0)          ## L=Lower end of interval
U      <- vector(mode='numeric', length(0)) ## U=Upper end of interval
r_stop <- vector(mode='numeric', length(0))
u      <- vector(mode='numeric', length(0)) ## runif(1)
phi     <- vector(mode='numeric', length(0)) ## -log(u)
error_RecEv <- FALSE
ERROR  <- "error_RecEv"
while(CONTINUE){
u[k]   <- runif(1)
phi[k] <- -log(u[k])
error_RecEv <- any(tryCatch(
{
gap[k] <- icHaz(
u=u[k], rdist=rdist, g=g, w=w, gammas=gammas,
b=b, eta=eta, z=z, t=L[k], FROM=FROM, TO=TO)
}, error=function(e){ERROR}) == ERROR )
if(error_RecEv){ ## Increase the resolution of the root finder.
tolexp.new <- 2*tolexp.new
cat('\nError␣generating␣the', k, 'recurrent␣event\n')
cat('Increasing␣resolution␣of␣root␣finder␣to', .Machine$double.eps^(tolexp.new),'\n
error_RecEv <- any(tryCatch(
{
gap[k] <- icHaz(u=u[k], rdist=rdist, g=g, w=w, gammas=gammas,
b=b, eta=eta, z=z, t=L[k], FROM=FROM, TO=TO,
tolexp=tolexp.new)
}, error=function(e){ERROR}) == ERROR )
} else {
U[k]      <- L[k] + gap[k]
r_stop[k] <- min(U[k], dT_cT)
deltaR[k] <- 1*( N[k] > 0 )
censored  <- ( U[k] > dT_cT )
gap.c[k]  <- ifelse( !censored, gap[k], r_stop[k]-L[k])
#.............................................................
## Collect the data here!
R <- rbind(R, cbind(
u=u[k], phi=phi[k], frailty=z, deltaR=deltaR[k], N=N[k],
dT=dT_cT, gap=gap[k], gap.c=gap.c[k], r.time=L[k], L=L[k],
U=U[k], r_start=L[k], r_stop=r_stop[k]
) )
#.............................................................
```

```
## Update values. If CONTINUE = FALSE, these will be reset.
N <- c(N, N[k]+1)
L <- c(L, U[k])
k <- length(N)
K <- K <- K+1
CONTINUE <- (K <= nRecEv) && (!censored)
}
} ##----------while
return(R)
}
# mmm <- f.simRec(
#   rdist="weibull", g=c(1,1.5), w=c(0.5,2), gammas=rep(0.2,2),
#   b=rnorm(2), eta=rep(0.5,2), z=10,
#   dT_cT=50000, nRecEv=1e4, FROM=0, TO=1e3)
#-------------------------------------------------------------
# 2. Simulate data 3-outcome joint model data |
#---------------------------------------------|
#
# "simJoint3" function to simulate data.

Version     <- 4.1
Last_update <- format(Sys.Date(), "%d␣%B␣%Y")
# Last_update <- "7/September/2019"

library(MASS)           ## Sample from multivariate normal mvrnorm(mu,Sigma)
library(plyr)           ## ddply()

simJoint3_Weibull <- function(
seed=20170920, n=500, nsim=2, t0=0, dt=0.2, repmax=20, reventsmax=6,
p1=0.5, p2=0.5, p3=0.5, p4=0.5, S_567.rand = FALSE,
mu_567     = c(0,0,0),
S_567      = diag(c(1,1,1)),
beta0      = 3,
betaT      = 0.5,
beta       = c(0.5, rep(0,6)),
gammaR     = c(0, 0.5, rep(0,5)), ## c(0.5, rep(0,6))
gammaT     = c(0, 0, 0.1, rep(0,4)), ## c(0, 0.5, rep(0,5))
rand.slope = FALSE,
var.eps    = 1.25,
var.b0     = 1.5^2,
var.b1     = 0.8^2,
```

```
cov.b      = -0.5,
var.v      = 0.8^2,
fdist      = "gamma",  ## or "log-normal"
etaR       = c(0.2, 0.2),
etaT       = c(0.5, 0.5),
alpha      = 2.6,
censt      = 5.5,
shapeT     = 1,   ## shape parameter of h_{0}(t) of Terminal
rateT      = 1.5, ## rate parameter of h_{0}(t) of Terminal
rdist      = "weibull", ## "log-logistic" distribution of the baseline hazard
shapeR     = 1,   ## shape parameter of r_{0}(t) of Recurrent
rateR      = 2,   ## rate parameter of r_{0}(t) of Recurrent
savefile   = TRUE,
outfiles   = c("simLong.txt", "simRec.txt", "simSurv.txt"),
outdir     = "/apps/amsta/mmjcr/JM3_td"
)
{
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
cat('Function simJoint3_Weibull(), version', Version, '\n')
cat('Last update:', Last_update,'\n')
# Create empty datasets to acummulate the data of a number of simulations
simLong <- NULL; simSurv <- NULL; simRec <- NULL
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
ID <- seq(1,n)

# Generate covariates once and keep the values fixed for all simulations
x1 <- (rbinom(n,1,p1))  # binary. Alternative, LaplacesDemon::rbern
x2 <- (rbinom(n,1,p2))  # binary
x3 <- (rbinom(n,1,p3))  # binary
x4 <- (rbinom(n,1,p4))  # binary
X1 <- mvrnorm(n, mu=mu_567, Sigma=S_567) # continuous, potentially correlated

colnames(X1) <- c("x5","x6","x7")
X  <- cbind(x1,x2,x3,x4,X1)
Xi <- split(X, ID)
for (i in seq_along(Xi)){names(Xi[[i]]) <- colnames(X)}

## Covariance matrix of random effects. Depends on distribution of frailties.
f.Sr <- function(fdist){
fdist_list   <- list("log-normal", "gamma")
Sr_list <- list(
```

```
"log-normal" = matrix(c(var.b0, cov.b, 0, cov.b, var.b1, 0, 0, 0, var.v),
byrow=TRUE, ncol=3),
"gamma" = matrix(c(var.b0, cov.b, cov.b, var.b1), byrow=TRUE, ncol=2)
)
d     <- match(fdist, fdist_list)
return(Sr_list[[d]])
}
Sr <- f.Sr(fdist=fdist)


# Three levels of "for" loops: s=simulations, ID=individuals, k=recurrent events.

# ------------------------------------------------------------
cat('------------------------------------------------------------------------\n')
for(s in 1:nsim){  ## s<-1
cat('sim = ', s, 'out of', nsim, '--> ')
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
if(fdist=="log-normal"){  ## The alternative is Gamma
randef <- mvrnorm(n, mu = c(0,0,0), Sigma = Sr)
} else {
randef <- cbind(mvrnorm(n, mu = c(0,0), Sigma = Sr),
log(rgamma(n, shape = 1/var.v, scale = var.v)))
}
b      <- randef[, c(1,2)]
v      <- randef[,3]
u1     <- runif(n) # To generate the survival times
censi  <- unlist(split(rep(censt,n), f=ID))

colnames(b) <- c("b0", "b1")
if (rand.slope){b[,2] <- b[,2]} else {b[,2] <- 0}
bi     <- split(b, ID)
for (i in seq_along(bi)){names(bi[[i]]) <- colnames(b)}

temp   <- data.frame(sim = s,
id     = ID,
X,
b0     = b[,1],
b1     = b[,2],
v      = v,
u1     = u1,
censt = censi)
```

```
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# Terminal event times
survt  <- 1/rateT * (
-log(u1) * exp(- alpha*v - X %*% gammaT - b %*% etaT)
)^(1/shapeT)

dT     = pmin(survt, censi)
event  = 1 * (survt <= censi)

temp <- transform(temp,
survt  = survt,
event  = event,
deltaD = NA, ## its value will come later
dT     = dT
)
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# Repeated measures & terminal event
jind <- with(temp, pmin(repmax-1, floor((dT - t0) / dt)))

# Expand the data frame repeating each id as many times as "jind" says
datlong <- temp[rep(row.names(temp), times=jind+1),]

datlong = ddply(datlong, "id", here(transform),
obstime = seq(1,length(id)),
l.time = dt * (seq(1,length(id))-1),
epsilon = rnorm(length(id),mean=0,sd=sqrt(var.eps)))

# Compute the longitudinal outcome
datlong$y <- unlist(lapply(seq(n), function(i) ## i<-2
with(subset(datlong, id==i),
(beta0+b0) + (betaT+b1)*l.time + as.numeric(X[i,]%*%beta) + epsilon)))

# Make deltaD = 1 only at the times when events occur.
# The ave() function is a shortcut for sapply(split(d,f,FUN)) or
# unlist(by(data,INDICES,FUN))
datlong$deltaD <- with(datlong,
ave(x=event, id,
FUN=function(x) c(rep(0,length(x)-1), x[1]) ) )

# Acummulate the repeated measures
simLong <- rbind(simLong,datlong)
```

```
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Recurrent events
## Create datrec to acummulate the recurrent events of all subjects.

datrec <- NULL

for(i in seq_along(ID)){ ## i<-2
R <- f.simRec_Weibull(rdist=rdist, g=c(shapeR,rateR), w=Xi[[i]], gammas=gammaR,
b=bi[[i]], eta=etaR, z=exp(v[i]), dT_cT=dT[i],
nRecEv=reventsmax)

R <- data.frame(
cbind(sim=s, seed=seed, id=ID[i], censt=censi[i],
survt=survt[i], event=event[i], deltaD=NA,
R,
matrix(rep(Xi[[i]], times=nrow(R)), byrow=T, nrow=nrow(R)),
matrix(rep(bi[[i]], times=nrow(R)), byrow=T, nrow=nrow(R))))
names(R) <- c(names(R[1:(ncol(R)-length(Xi[[i]])-length(bi[[i]]))]),
names(Xi[[i]]), names(bi[[i]]))

# Accumulate for all subjects, i = 1,...,n
datrec <- rbind(datrec,R)
} ## end of for(i) loop

# Make deltaD = 1 only at the times when events occur.
datrec$deltaD <- unlist(
with(datrec, by(data=event, INDICES=id, FUN=function(x)
c(rep(0,length(x)-1), x[1]))))

# Acummulate the recurrent events of all simulations
simRec <- rbind(simRec,datrec)
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Survival analysis in counting process format. This is needed to fit the
## extended Cox model with y(t) and N(t) as time-varying covariates.
cat('Expand␣simSurv␣for␣y(t)␣&␣N(t)␣-->␣')

lvars <- c('l.time','y','epsilon')
rvars <- c('r.time','deltaR','N')

yi <- lapply(split(datlong, f=datlong$id), function(d) d[,lvars])
Ni <- lapply(split(datrec, f=datrec$id), function(d) d[,rvars])
```

```
ID <- 1:n
surv <- NULL
LONG <- NULL
REC  <- data.frame(r.time=0, deltaR=0, N=0)
nn <- 1  ## Start counter for the number of rows in dataset
for (i in seq_along(ID)){ ## i<-213 i<-53
## Subset files by id and keep only relevant variables
l <- subset(datlong, id==ID[i], select=lvars)
r <- subset(datrec,  id==ID[i], select=rvars)
v <- subset(temp, id==ID[i])
## Times of all time-varying variables
tl <- subset(l, select = l.time)
tr <- subset(r, select = r.time)
td <- subset(v, select = dT)
time <- sort(unique(unlist(c(tl, tr, td))))
## Start and stop times
K <- length(time)
L <- time[-K]
U <- time[-1]
## Initialize recurrent events part at t=0. Not necessary for longitudinal.
for (j in 1:(K-1)){ ## j<-1
I    <- cbind(t_start=L[j], t_stop=U[j])
long <- subset(l, l.time >= L[j] & l.time < U[j])
rec  <- subset(r,  r.time >= L[j] & r.time < U[j])
if (nrow(long)==0){ long <- LONG[nn-1, ] }  ##
if (nrow(rec)==0) {
rec        <- REC[nn, ]
rec$deltaR <- 0 }
LONG <- rbind(LONG, long) ## LONG and REC are auxiliary sets
REC  <- rbind(REC, rec)    ## LONG and REC are auxiliary sets
surv <- rbind(surv, cbind(v, I, long, rec) )
nn   <- nn + 1
# surv[,-which(names(surv) %in% fcovs)]; j <- j+1
}
}
surv        <- data.frame(surv)
surv$deltaD <- with(surv,
ave(event, id,
FUN=function(x) c(rep(0, length(x)-1), x[1]) ) )
```

```
simSurv <- rbind(simSurv,surv)
cat('Finished\n')
} # End of "for s" loop
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
cat('Preparing', length(outfiles), 'files.\n')
simList <- list(simLong=simLong, simRec=simRec, simSurv=simSurv)
if(savefile){
setwd(outdir)
for (f in seq(simList)){ ##f<-1
write.table(simList[[f]], outfiles[f], row.names=FALSE, col.names=TRUE, sep='\t')
}
cat('=========================================================================\n')
cat('Output files:', paste0(outfiles, collapse=", "), '\n')
cat('Location:', outdir, '\n')
}
return(simList)
}
#----------------------END OF FUNCTION simJoint3_Weibull()-------
```

A.7 Simulation from a joint model of longitudinal and time to event data with a counting process as time-varying covariate

Sample code to simulate data from model M_2 of Section 6.3. Corresponds to Algorithm 6.2.

```
##=================================================================
# Last_update <- "08 September 2019"
Last_update <- format(Sys.Date(), "%d␣%B␣%Y")
Version <- 4.1
##-----------------------------------------------------------
## R code for simulating a joint model of longitudinal and
## time-to-event outcomes with an exogenous time-varying
## covariate (a counting process),
## This version, incontrast to v3, stops the recurrent event
## process at min(censoring time, time-to-event, tT).
##-----------------------------------------------------------
## Chapter: Causal Infernce with Joint Models
## For third year 201819
## (c) Jacob Cancino-Romero (JHD,LB,SB) 2019
##=================================================================
#=================================================================
#  CONTENTS
# ----------
#
#       1) Simulate data from a joint model for longitudinal and time-to-event
#    outcomes with Weibull baseline and time-varying covariates.
#    The model is:
#
#    y(t) = (beta_{0} + b_{i0}) + x^{T}beta + beta_{t}*time + epsilon(t)
#         = m(t) + epsilon(t)
#    h(t) = h_{0}(t)*exp(w^{T}*gamma + eta_{Y}*m(t) + eta_{R}*N(t))
#
#    where N(t) is a Poisson(lam) counting process.
#
#=================================================================
```

```
library(MASS)      ## mvrnorm(n, mu, Sigma)
library(survival) ##
library(plyr)      ## ddply()

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Time functions
ft <- function(x, fnt, p){
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
fn.names <- c("linear", "banana", "cos-sin", "logistic", "normal")
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
f1 <- function(x,p){
p[1] + p[2]*x }
f2 <- function(x,p){
p[1] + p[2]*x + p[3]*x^3 + p[4]*exp(p[5]*x^2) }
f3 <- function(x,p){
p[1] + p[2]*x + p[3]*sin(p[4]*x) + p[5]*cos(p[6]*x) +
p[7]*(x+p[8])^2 + p[9]*exp(p[10]*x)}  ##
f4 <- function(x,p){
p[2] / (1 + exp(-p[3]*(x-p[1])))}
f5 <- function(x, p){
p[1]*x + p[2]*dnorm(x, mean=p[3], sd=p[4])}
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
flist <- list("linear"=f1, "banana"=f2, "cos-sin"=f3, "logistic"=f4,
"normal"=f5)
d <- match(fnt, fn.names)
return(flist[[d]](x=x,p=p))
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
mx <- function(x, fnt, p, betas, b, w){
# ..........................................................
## x is time
## betas are the betas (fixed effects regression coefficients)
## m(t) = (beta1+b1) + (beta2+b2)f(t) + (beta3+b3)w1 +...+ (betak+bk)w_{k-2}
##      = Xbeta + Zb
## X[i,] = c(1, f(t), w)
## Z[i,] = c(1, f(t))
# ..........................................................
p      <- as.vector(p) ## parameters that determine f(t)
betas  <- as.vector(betas) ## fixed effects covariates
```

```
b      <- as.vector(b) ## random effects
# ..........................................................
fx     <- ft(x, fnt=fnt, p=p)
# ..........................................................
W      <- matrix(rep(w, length(fx)), byrow=T, ncol=length(w))
X      <- as.matrix(cbind(1, fx, W))
Z      <- cbind(1, fx)
mt     <- as.numeric(crossprod(t(X),betas)) + as.numeric(crossprod(t(Z),b))
return(mt)
}
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Functions to simulate time to event
## Baseline hazard.
h0 <- function(x, sdist, g){
## Simulate survival times.
## Baseline hazard. Using pweibull()/deweibull() might cause
## computational difficulties as time gets larger (0/0 = NaN)
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
d.names <- c("weibull", "log-logistic", "gompertz", "makeham", "bathtub")
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
h1 <- function(x,g){ g[1]*g[2]*(g[2]*x)^(g[1]-1) }
h2 <- function(x,g){ g[1]*g[2]*(g[2]*x)^(g[1]-1) / (1+g[2]*x)^g[1]}
h3 <- function(x,g){ g[1]*exp(g[2]*x) }
h4 <- function(x,g){ g[1] + g[2]*exp(-g[3]*x)}
h5 <- function(x,g){ g[1]*x + g[2]/(1 + g[3]*x) }
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
hlist <- list(h1, h2, h3, h4, h5)
names(hlist) <- d.names
distrib <- match(sdist, d.names)
return(hlist[[distrib]](x=x,g=g))
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Subject specific hazard rate
hx <- function(x, fnt, p, betas, b, w, etaL, etaR, gammas, N,
sdist, g, v){
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## h(t) = h_{0}(t) * exp{etaL*m(t) + etaR*N(t) + v + w^{T}gammas}
## v = log(frailty), where frailty is an instance of the
## gamma or log-normal distribution.
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

```
gammas <- as.vector(gammas)
w      <- as.vector(w)
exp(etaL*mx(x, fnt=fnt, p=p, w=w, betas=betas, b=b) +
etaR*N + v + as.numeric(crossprod(gammas,w))) *
h0(x, sdist=sdist, g=g)
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Cumulative hazard. Notice different vectorization method via sapply():
cumhaz <- function(
x, fnt=fnt, p, betas, b, w, etaL, etaR, gammas, N, sdist, g, v, LOWER){
sapply(x, function(x)
integrate(hx, lower=LOWER, upper=x,
fnt=fnt, p=p, betas=betas, b=b, w=w, etaL=etaL, etaR=etaR,
gammas=gammas, N=N, sdist=sdist, g=g, v=v)$value)
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Inverse cumhaz (!!Not fully vectorized!!):
## Here u cannot be a vector. Use sapply in if u is a vector.
icumhaz <- function(
u, fnt, p, betas, b, w, etaL, etaR, gammas, N, sdist, g, v, LOWER,
FROM=0, TO, shift, tolexp=0.5){ ## u stands for S(t). Make sure 0<=u<= 1
phi <- (-log(u))
h   <- function(x){
cumhaz(x, LOWER=FROM, fnt=fnt, p=p, betas=betas, b=b, w=w, etaL=etaL,
etaR=etaR, gammas=gammas, N=N, sdist=sdist, g=g, v=v) + shift - phi
} ## Must subtract phi, not u!
x <- uniroot(h, interval=c(FROM, TO), tol=.Machine$double.eps^tolexp)$root
return(x)
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Recurrent events require the conditional cumulative hazard
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Subject specific hazard rate (recurrent events)
rx <- function(x, w, gammas, sdist, g, v){
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## r(t) = r_{0}(t) * exp{w^{T}gammas}
## v = log(frailty), where frailty is an instance of the
## gamma or log-normal distribution.
```

```
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
gammas <- as.vector(gammas)
w      <- as.vector(w)
exp(as.numeric(crossprod(gammas,w)) + v) * h0(x, sdist=sdist, g=g)
}
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Cumulative hazard with frailty (recurrent events)
Rx <- function(x, w, gammas, sdist, g, v, LOWER){
## v = log(frailty), where frailty is an instance of the
## gamma or log-normal distribution.
sapply(x, function(x)
integrate(rx, lower=LOWER, upper=x,
w=w, gammas=gammas, sdist=sdist, g=g, v=v)$value)
}
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Conditional cumulative hazard with frailty (recurrent event)
condRx <- function(x, t, w, gammas, sdist, g, v, LOWER){
Rx(x=x+t, w=w, gammas=gammas, sdist=sdist, g=g, v=v, LOWER=LOWER) -
Rx(x=t, w=w, gammas=gammas, sdist=sdist, g=g, v=v, LOWER=LOWER)
}
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Inverse of the conditional cumulative hazard with frailty (recurrent event)
icondRx <- function(u, t, w, gammas, sdist, g, v, LOWER, TO, tolexp=0.5){
## u stands for S(t). Make sure 0<=u<= 1
phi <- (-log(u))
h   <- function(x){
condRx(x, t=t, sdist=sdist, g=g, w=w, gammas=gammas, v=v, LOWER=LOWER)-phi
} ## Must subtract phi, not u!
x <- uniroot(h, interval=c(LOWER, TO), tol=.Machine$double.eps^tolexp)$root
return(x)
}

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Random sample from the frailties distribution.
f.frailty <- function(n, fpar, fdist){
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
frailty.dist <- c("gamma", "log-normal")
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
f1    <- function(n,fpar){rgamma(n, shape=1/fpar, scale=fpar)}
f2    <- function(n,fpar){rlnorm(n, meanlog=0, sdlog=fpar)}
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

```
flist <- list("gamma"=f1, "log-normal"=f2)
d     <- match(fdist, frailty.dist)
return(flist[[d]](n=n,fpar=fpar))
}
v.frailty <- Vectorize(f.frailty) ## Vectorized version

#============================================================
library(MASS) # Use for multivariate normal: mvrnorm(n, mu=c(0,0), Sigma=S)
library(mvtnorm) # Use for multivariate normal: rmvnorm(n,mean=c(0,0), sigma=S)
library(plyr) # Use to transofm data set with function "ddply"

## Note: This model is sensitive to the values of etaL and etaR
## When one of etaL or etaR < -0.02
f.simJM3_tdW_b0b1 <- function(
savefiles  = TRUE,
outdir     = actout_dir,
outfiles   = c("simLong.txt", "simRec.txt", "simSurv.txt"),
seed = 0, nsamp = 10, nsim = 3, dt = 0.5, repmax = 10,
p1=0.5, p2=0.5, p3=0.5, p4=0.5,
t0         = 0,
tT         = 20,
mu_567     = rep(0,3),
S_567      = matrix(c(1, 0, 0, 0, 4, 0.4, 0, 0.4, 4), ncol=3),
rand.slope = TRUE, ## No need to modify Sb if rand.slope=FALSE
var.eps    = 4,
var.b0     = 2,
var.b1     = 2,
cov.b      = 0,
fnt        = "cos-sin",
p          = c(0, 1, 1, 2 ,1, 1, rep(0,4)),
sdist      = "weibull",
g          = c(1, 1.5), ## c(shape=kappa,rate=rho) of h_{0}(t) of Terminal
beta0      = 3,
beta.time  = 0.5,                       ## time
beta.fixed = c(0, 0, 0, 0.5 ,0.5, 0, 0), ## x4, x5
gammaT     = c(0.1, 0.1, 0, 0, 0, 0, 0), ## x1, x2
gammaR     = c(0, 0, 0.1, 0 ,0, 0.1, 0), ## x3, x6 for recurrent events
etaL       = -0.02,#-0.1,  ## Y: Longitudinal outcome in the survival submodel
etaR       = -0.02,#-0.1,  ## R: Recurrent events in the survival submodel
rdist      = "weibull",  ## Recurrent events: Baseline hazard
rateR      = 3,    ## rate parameter of r_{0}(t) of Recurrent
```

```
shapeR     = 1,    ## shape parameter of r_{0}(t) of Recurrent
rand.censt = FALSE,
censt      = 5,
frailty    = FALSE, ## if TRUE, sometimes survival times = NaNs
fdist      = "log-normal",
fpar       = 0.4^2, ## Needs to be chosen carefully.
tol_exp    = 0.5 ## Tolerance of unit root finder .Machine$double.eps^tolexp
){
cat('=================================================================\n')
cat('Function␣f.simJM3_tdW_b0b1(),␣version', Version, '\n')
cat('Last␣update:', Last_update,'\n')
cat('Check␣also␣f.simJoint3.td.dist(plots)_v3␣for␣details␣and␣plots.\n')
cat('-----------------------------------------------------------------\n')
Sys.sleep(1)

set.seed(seed) ## sample(1:100, 1)

# Empty datasets to acummulate nsim simulations
simLong <- simSurv <- simRec <- check_ctrl_all <- NULL

ns <- nsim
n  <- nsamp
ID <- seq(1, n)

betas   <- as.vector(c(beta0, beta.time, beta.fixed))
gammaT  <- as.vector(gammaT)
gammaR  <- as.vector(gammaR)
Sb      <- matrix(c(var.b0, cov.b, cov.b, var.b1), ncol=2)

# Time-fixed covariates. Produce the values once and keep them constant
# for all simulations
x1 <- (rbinom(n,1,p1))  # binary
x2 <- (rbinom(n,1,p2))  # binary
x3 <- (rbinom(n,1,p3))  # binary
x4 <- (rbinom(n,1,p4))  # binary
X1 <- mvrnorm(n, mu_567, S_567) # continuous -> see correlations

colnames(X1) <- c("x5", "x6", "x7")
X  <- cbind(x1,x2,x3,x4, X1)
Xi <- split(X, ID)
for (i in seq_along(Xi)){names(Xi[[i]]) <- colnames(X)}
```

```
# Two levels of "for" loops: s=simulations and i=individuals
#-------------------------------------------------------------------
for(s in 1:ns){  ## s<-34; s<-1

set.seed(seed+s)

## Random effects: b and v=log(frailty)
if (rand.slope) {
b  <- mvrnorm(n, c(0,0), Sb) ## Random intercepts and slopes
} else {
b <- cbind(rnorm(n, 0, sd = sqrt(Sb[1,1])), 0)
}
colnames(b) <- c("b0", "b1")
bi          <- split(b, ID)
for (i in seq_along(bi)){names(bi[[i]]) <- colnames(b)}
if (frailty){  ## Note: v = log(frailty), frailty ~ gamma(1/fpar, fpar)
v <- log(f.frailty(n=n, fdist=fdist, fpar=fpar))
} else {
v <- rep(0,n)
}
vi <- split(v, ID)

## Censoring
if(rand.censt){
ci <- split(runif(n, min=0, max=censt), ID)
} else {
ci <- split(censt+double(n), ID)
}

temp <- data.frame(sim = s,
id  = ID,
X,
b,
log_frailty = v,
censt = unlist(ci))

## Linear predictor of the model used to simulate recurrent events.
psi     <- as.numeric(exp(X %*% gammaR))
#-------------------------------------------------------------------
## Simulate survival times from Weibull(sh=kap, r=rho)
```

```
## Baseline functions. Do not use pweibull()/deweibull() since it causes
## computational difficulties as time gets larger (0/0 = NaN)
##.............................................................
u      <- runif(n) ## Recall U = S(t), where U ~ Unif[0,1]
phi    <- -log(u)
sim_dT <- all_dT <- NULL
#.............................................................
for (i in seq_along(ID)){ ## i<-44 ## i<-1
## Initialize objects
censt      <- ci[[i]]
CONTINUE   <- TRUE
compute_T  <- FALSE
censored_T <- FALSE  # Terminal event is censored
censored_R <- FALSE  ## Last recurrent event is censored
censored_F <- FALSE  ## The time limit tT has been reached
censored_any <- FALSE
R          <- NULL       ## data frame to acumulate the data
K          <- HtL <- 0
StU        <- 1
L          <- N <- Ck <- DK <- c(0)              ## L=Lower end of interval
U          <- vector(mode='numeric', length(0))  ## U=Upper end of interval
#.............................................................
while(CONTINUE){  ## Mind k and K: K={0,1,...} is the count of events up to t
k    <- length(N)
## Check. Suppose rate = r*exp(x^{T}beta)
## U ~ Uniform(0,1) <=> -log(U)/rate = X ~ Exp(rate)
## X ~ Exp(rate) <=>  exp(-rate*X) = U ~ Uniform(0,1)
# gap1   <- rexp(1, rate=lam*psi[i]) ## Sample a new time gap each iteration.
new_u  <- runif(1)
gap    <- icondRx(u=new_u, t=L[k], w=Xi[[i]],
gammas=gammaR, sdist=rdist, g=c(shapeR,rateR),
v=vi[[i]], LOWER=0, TO=1e3, tolexp=tol_exp)
U[k]   <- L[k]+gap

## Ht is H(t_[k]) - H(t_{k-1})
Ht  <- cumhaz(x=U[k], LOWER=L[k], N=N[k], b=bi[[i]], w=Xi[[i]], v=vi[[i]],
fnt=fnt, p=p, betas=betas, etaL=etaL, etaR=etaR,
gammas=gammaT, sdist=sdist, g=g)
HtU <- HtL + Ht; HtU
StL <- exp(-HtU); StL  ## St is S(t_{k}) - S(t_{k-1})
```

```
compute_T    <- (phi[i] >= HtL & -log(u[i]) < HtU); compute_T
censored_R   <- (L[k] <= censt & censt <= U[k])  ## censt in [L[k],U[k]]
censored_F   <- (U[k] >= tT)  ## Last T_{k} >= censt
censored_any <- any(censored_F, censored_R, censored_F); censored_any

if(censored_R){    ## compute H(t) at t=censt
Hcenst <- HtL + cumhaz(
x=censt, LOWER=L[k], N=N[k], b=bi[[i]], w=Xi[[i]], v=vi[[i]],
fnt=fnt, p=p, betas=betas, etaL=etaL, etaR=etaR,
gammas=gammaT, sdist=sdist, g=g)
}

## In this version, the recurrent events generating process stops if:
##   a) phi[[i]] lies in the interval (HtL, HtU), so survt is computed, or
##   b) U[k] >= censt, or
##   c) U[k] has reached the time limit.

if(compute_T){
#.............................................................
survt <- icumhaz(u=u[i], FROM=L[k], TO=U[k], N=N[k], shift=HtL,
b=bi[[i]], w=Xi[[i]], v=vi[[i]],
fnt=fnt, p=p, betas=betas, etaL=etaL, etaR=etaR,
gammas=gammaT, sdist=sdist, g=g, tolexp=tol_exp)
## Sometimes dT==0 exactly. This happens when u[i] is small, e.g.
## u[i]=3.2e-05 when s=24, nsamp=500, i=475, so St==1.
## If dT==0 decrease the rolerance of the root finder from
## .Machine$double.eps^0.5 to .Machine$double.eps^1.
if(survt == 0){
cat('Reduced␣root␣finder␣tolerance␣to', .Machine$double.eps^(2*tol_exp), '--->','su
cat('sim␣=', s, "id␣=", i, '\n')
survt <- icumhaz(u=u[i], FROM=L[k], TO=U[k], N=N[k], shift=HtL,
b=bi[[i]], w=Xi[[i]], v=vi[[i]],
fnt=fnt, p=p, betas=betas, etaL=etaL, etaR=etaR,
gammas=gammaT, sdist=sdist, g=g, tolexp=2*tol_exp)
Htstar <- HtL + cumhaz(
x=survt, LOWER=L[k], N=N[k], b=bi[[i]], w=Xi[[i]], v=vi[[i]],
fnt=fnt, p=p, betas=betas, etaL=etaL, etaR=etaR,
gammas=gammaT, sdist=sdist, g=g)
}
# cat("Problem point 1, survt =", survt, "\n")
censored_T <- (survt >= censt) ## ; censored_T Remove to avoid problems
```

```
dT         <- min(survt, censt)
#.............................................................
} else {
survt      <- NA ## icumhaz(p=StU-1e+10); survt
censored_T <- censored_T  ## Necessary to calculate deltaD below
dT         <- survt
}

CONTINUE     <- (!compute_T & !censored_any); CONTINUE

##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
## Break the while loop if at least one process is censored
## or the time limit has been reached.
if(censored_any){
survt <- censt
dT    <- survt
}
##~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

## Terminal and recurrent events indicators:
##  - deltaD = 1 if dT < censt & death in Ht interval
##  - deltaR = 1 at every time dN(t) > 0

# writeLines("Problem point 2")
deltaD <- 1*(!censored_any & compute_T)
deltaR <- 1*(K>0 & !censored_any)
r_stop <- min(U[k], censt, survt, na.rm=TRUE)
# writeLines("Beyond problem point 2")

## Collect the data HERE!.
R <- rbind(R, cbind(
sim=s, id=ID[i], gap=gap, r.time=L[k], N=N[k], deltaR=deltaR,
L=L[k], U=U[k], r_start=L[k], r_stop,
phi=phi[[i]], St=u[i], HtL=HtL, HtU=HtU, StU=StU, StL=StL,
survt=survt, dT=dT, deltaD=deltaD, event=deltaD,
t(Xi[[i]]), t(bi[[i]]), log_frailty=v[i]))
#.............................................................
## Update values. If CONTINUE = FALSE, these will be reset.
N   <- c(N, K+1)
K   <- K+1
L   <- c(L,U[k]) ## Next iteration L (start) takes current value of U (stop)
```

```
U   <- c(U,0)  ## Add an element to vector U, which will be replaced by L+gap
HtL <- HtU
StU <- StL
round(R,3)
} ## end of while ..........................................

## Repeat "dT" in all rows for each subject and make the "event" column
## keep constant the terminal event status (1 or 0) in all rows.
## "deltaD" = "event" only in the last row.
R[, 'dT']    <- dT
R[, 'event'] <- R[nrow(R), 'deltaD']
all_dT <- rbind(all_dT, R)
# print(round(all_dT, 2))
# i<-i+1

} ## end of for individuals' loop and continue with s loop ......
#-------------------------------------------------------------------
## "sim_dT" removes all rows where recurrent events occurred after censt.
sim_survt <- all_dT[!is.na(all_dT[, 'survt']) , c('id','survt')]
sim_dT    <- all_dT[all_dT[, 'r_start'] <= all_dT[, 'dT'],
c('id','r.time','deltaR','N','dT','deltaD','event','phi','St')]
sim_dT    <- merge(x=sim_dT, y=sim_survt, by='id')
## Keep only the last record
last_sim_dT <- sim_dT[!duplicated(sim_dT[,'id'], fromLast=TRUE), ]

## "all_dT" keeps all recurrent events records up to the observed terminal event.
all_dT <- data.frame(all_dT)
sim_dT <- data.frame(sim_dT)
round(all_dT, 4)
dT_check <- all(!is.na(sim_dT))
cat('sim␣=', s, 'survival␣times␣simulation␣is␣complete?', dT_check,'\n')
################################################################
## Check if Ht and St are consistent with formulas.
## I could not vectorize cumhaz() on LOWER. I think it is because
## of how I defined the function mx(). So I use lapply twice.
R   <- all_dT
idR <- R[,'id']
Ri  <- split(R, idR)
Nti <- lapply(1:n, function(x) Ri[[x]][,'N'])
tLi <- lapply(1:n, function(x) Ri[[x]][,'r_start'])
tUi <- lapply(1:n, function(x) Ri[[x]][,'r_stop'])
```

```
Li  <- lapply(1:n, function(x) Ri[[x]][,'L'])
Ui  <- lapply(1:n, function(x) Ri[[x]][,'U'])

Ht  <- unlist(lapply(1:n, function(i) ## i<-1
cumsum(unlist(lapply(seq_along(Ui[[i]]), function(j)
cumhaz(x=Ui[[i]][j], LOWER=Li[[i]][j], N=Nti[[i]][j],
b=bi[[i]], w=Xi[[i]], v=vi[[i]], fnt=fnt, p=p, betas=betas,
etaL=etaL, etaR=etaR, gammas=gammaT, sdist=sdist, g=g) ) ) ) ) )
Ht_check <- all(abs(Ht - R$HtU) <= 1e-10) ## cbind(Ht, R$HtU)
St       <- exp(-Ht); St
St_check <- all(abs(R[,'StL']-St) <= 1e-15);  cbind(R[,'StL'], St)
cat('sim = ', s, 'Checked Ht and St?', Ht_check, St_check,'\n')

#################################################################
## Merge temp dataset with last_sim_dT
temp <- merge(temp, last_sim_dT)
#----------------------------------------------------------------
# Repeated measures & terminal event
jind    <- with(temp, pmin(repmax-1, floor((dT - t0) / dt)))
temp$ni <- sapply(jind, function(d) d + 1)

# Expand the data frame repeating each id as many times as "jind" says
datlong <- temp[rep(row.names(temp), times=jind+1),]

datlong <- ddply(datlong, "id", here(transform),
obstime = seq(1,length(id)),
l.time  = dt * (seq(1,length(id))-1),
epsilon = rnorm(length(id), mean=0, sd=sqrt(var.eps)))
names(datlong)
#----------------------------------------------------------------
## Simulate the longitudinal outcome
ti         <- with(datlong, split(l.time, f=id))
datlong$mi <- unlist(lapply(1:n, function(i) ## i<-1
mx(x=ti[[i]], fnt=fnt, p=p, betas=betas, b=bi[[i]], w=Xi[[i]]) ) )
datlong$y <- with(datlong, mi + epsilon)

cat('sim = ', s, 'Adding terminal and recurrent event times to simLong.\n')
## Modify datlong to include the number of recurrent events
## between repeated measures times, and to have the event
## indicator at the last period
```

```
## Time-to-event
datlong$l_start <- datlong$l.time
splitID         <- split(datlong[c('l_start','dT')], datlong$id)
datlong$l_stop  <- unlist(lapply(
splitID, function(d) c(d$l_start[-1], d$dT[1]) ) )
datlong$deltaD   <- with(datlong, ave(event, id, FUN=function(x)
c(rep(0, length(x)-1), x[1]) ) )

## Recurrent events (Only those that occured before censoring)
splitN <- split(sim_dT[c('r.time', 'N', 'deltaR')], sim_dT$id)
Nac    <- unlist(lapply(splitN, function(d) cumsum(d$deltaR)))
all(sim_dT$N == Nac)
splitL <- split(datlong[c('l_start','l_stop')], datlong$id)

## Complete datlong by including the recurrent and terminal events.
datlong$r.time <- datlong$Nt <- datlong$dN <- NA
for (i in 1:n){ # i<-28
RTIME <- NT <- NULL
tr <- splitN[[i]]   ## start times of recurrent events
mr <- nrow(tr)      ## number of recurrent event times
tl <- splitL[[i]]   ## time intervals of longitudinal outcome
ml <- nrow(tl)      ## number of longitudinal outcome times
for (k in 1:ml){ # k<-5
whichN <- which(tl$l_start[k] <= tr$r.time & tr$r.time < tl$l_stop[k])
whichd <- which(tl$l_start[k] <= tr$r.time & tr$r.time < tl$l_stop[k] &
tr$r.time > 0)
dN     <- length(whichd)
if(length(whichN) > 1){ ## If more than 2 events, record the latest
whichN <- max(whichN)
}
rtime  <- tr$r.time[whichN]
nt     <- tr$N[whichN]
if(length(rtime) == 0){ ## means no Rec.event occurs after l.time
rtime  <- RTIME[k-1]
nt     <- NT[k-1]
}
RTIME  <- c(RTIME, rtime)
NT     <- c(NT, nt)
## dN has to be 0 if r.time = 0
dN     <- dN * (rtime > 0)
## Fill the column one time interval & one subject at a time.
```

```
datlong$r.time[datlong$id == ID[i] & datlong$l_start == tl$l_start[k]] <-
rtime
datlong$Nt[datlong$id == ID[i] & datlong$l_start == tl$l_start[k]] <-
nt
datlong$dN[datlong$id == ID[i] & datlong$l_start == tl$l_start[k]] <-
dN
# cat("i =", i, "k =", k, "\n")
}
}
## deltaR = 1 if at least 1 recurrent event occured between 2 repeated measures.
datlong$deltaR <- 1*(datlong$r.time > 0 &
datlong$r.time >= datlong$l_start &
datlong$r.time < datlong$l_stop)
cat('sim = ', s, 'Creating survival analysis dataset counting process format.\n')
#-------------------------------------------------------------
## Create survival analysis data in counting process format where
## recurrent events and repeated measures are time-varying covariates.
## Select the time-dependent variables and covariates required.
ID  ## Was defined at the begining
fcovs <- c('sim', 'id', paste0('x', seq(1,7)), 'b0', 'b1', 'log_frailty')
lvars <- c('l.time', 'obstime', 'y', 'epsilon')
rvars <- c('r.time','deltaR','N','dT','survt','deltaD','event','phi','St')

covs_data <- subset(temp, select=fcovs)
sim_dT <- data.frame(sim_dT)

surv <- LONG <- REC <- NULL
nn <- 1  ## Start counter for the number of rows in dataset
for (i in seq_along(ID)){ ## i<-213 i<-53
## Subset files by id and keep only relevant variables
l <- subset(datlong,   id==ID[i], select=lvars)
r <- subset(sim_dT,    id==ID[i], select=rvars)
v <- subset(covs_data, id==ID[i], select=fcovs)
## Times of all time-varying variables
tl <- subset(l, select = l.time)
tr <- subset(r, select = r.time)
td <- subset(r, select = dT)
time <- sort(unique(unlist(c(tl, tr, td))))
## Start and stop times
K <- length(time)
L <- time[-K]
```

```
U <- time[-1]
for (j in 1:(K-1)){ ## j<-1
I <- cbind(t_start=L[j], t_stop=U[j])
long <- subset(l, l.time >= L[j] & l.time < U[j])
rec <- subset(r,  r.time >= L[j] & r.time < U[j])
if (nrow(long)==0){ long <- LONG[nn-1, ]}  ##
if (nrow(rec)==0){
rec <- REC[nn-1, ]
rec$deltaR <- 0 }
LONG <- rbind(LONG, long) ## LONG and REC are auxiliary sets
REC <- rbind(REC, rec)    ## LONG and REC are auxiliary sets
surv <- rbind(surv, cbind(v, I, long, rec) )
nn <- nn + 1
# surv[,-which(names(surv) %in% fcovs)]; j <- j+1
}
}
surv <- data.frame(surv)
surv$deltaD   <- with(surv,
ave(event, id,
FUN=function(x) c(rep(0, length(x)-1), x[1]) ) )
#-----------------------------------------------------------------
## Acummulate: repeated measures, recurrent events and survival analysis files
simRec <-  rbind(simRec, all_dT)
simSurv <- rbind(simSurv, surv)
simLong <- rbind(simLong, datlong)
data_check <- all(!is.na(sim_dT))
cat('sim␣=', s, 'data␣integration␣is␣complete?', data_check,'\n')
# setwd(outdir)
# tables <- list(simLong, simRec, simSurv)
# # files <- c('simLong.txt', 'simRec.txt', 'simSurv.txt')
# files <- LRS_outfiles
# for(z in seq_along(tables)){
#   write.table(tables[[z]], files[z], row.names=F, col.names=T, sep='\t')
# }
# cat('Backup (do not open yet):',paste0(files,collapse=", "),'in:\n',outdir,'\n')
cat('=================================================================\n')
} ## End of "for s" loop
#-----------------------------------------------------------------
tables <- list(simLong=simLong, simRec=simRec, simSurv=simSurv)
if(savefiles){
setwd(outdir)
```

```
files <- outfiles
for(z in seq_along(tables)){
write.table(tables[[z]], files[z], row.names=F, col.names=T, sep='\t')
}
cat('Output␣files:',paste0(files,collapse=",␣"),'\n')
cat('Folder:', outdir, '\n')
}
return(tables)
}
#~~~~~~~~~~~~~~~~~~~~~~~~END OF FUNCTION f.simJM3_tdW_b0b1~~~~~~~~~
```

Appendix B

Plots

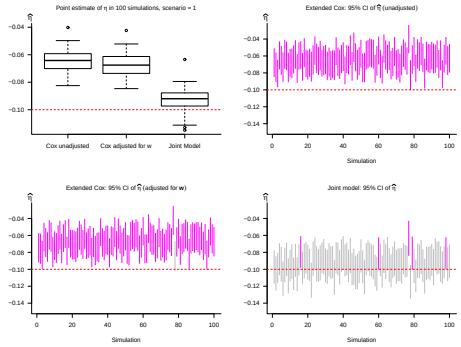
B.1 Compare extended Cox model and joint modelling

Simulation scheme to illustrate that estimates of the Cox model with endongenous time-varying covariates are biased. The plot of Section 3.1 corresponds to simulation scheme $\text{sim} = 3$.

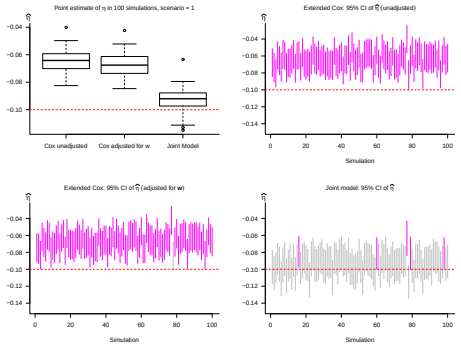
$$\begin{cases} \text{(Longitudinal)} & y_i(t|\mathbf{b}_i) = \overbrace{(\beta_0 + b_{i0}) + (\beta_t + b_{i1})f(t) + \mathbf{w}_i^\top \boldsymbol{\beta}}^{m_i(t)} + \varepsilon_i(t) \\ \text{(Terminal)} & h_i(t|\mathbf{b}_i) = h_0(t)\exp\{\mathbf{w}_i^\top \boldsymbol{\gamma} + \eta m_i(t)\} \\ & f(t) = \text{a function of time} \end{cases}$$

	sim = 1	sim = 2	sim = 3	sim = 4	sim = 5	sim = 6
$f(t)$	$x + \cos(0.5x) + \sin(0.5x)$		x			
η	-0.1	-0.2	-0.1	-0.2	0.1	0.2

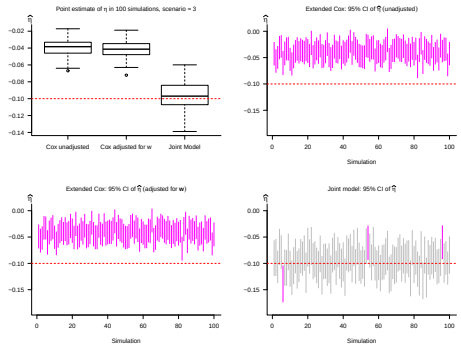
$$\begin{aligned} h_0(t) &\sim \text{Weibull}(\kappa = 4, \rho = 0.2) \\ \beta_t &= 0.1 \\ \beta_0 &= 0.5 \\ \boldsymbol{\beta}^\top &= (0.1, 0.2, -0.2, 0.3, -0.3) \\ \boldsymbol{\gamma}^\top &= (0.1, 0.1, 0.1, 0.1, 0.1) \\ n &= 500 \\ C_i &= 8 \; \forall i \\ \text{var}(\mathbf{b}_i) &= \begin{bmatrix} 3 & \\ 0 & 1 \end{bmatrix} \\ \text{var}(\varepsilon_i) &= 10 \end{aligned}$$



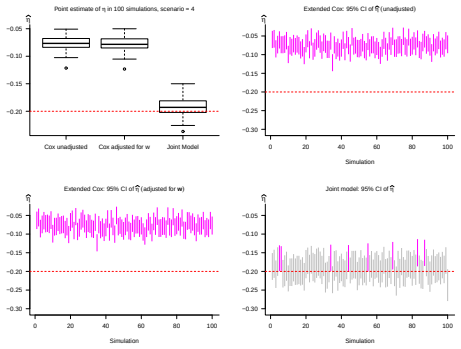
Simulation 1



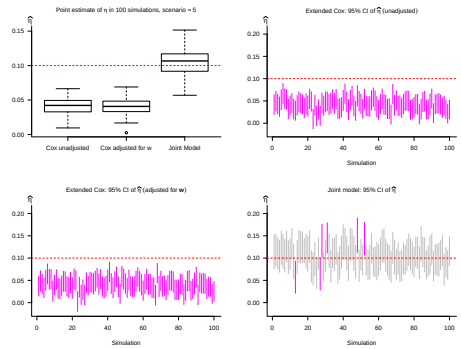
Simulation 2



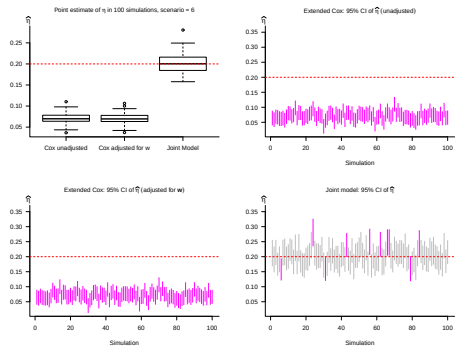
Simulation 3



Simulation 4



Simulation 5



Simulation 6

Figure B.1: Compare parameter estimates, $\widehat{\eta}$ for simulated data under different scenarios for the function of time in the longitudinal outcome and strength of the effect of the longitudinal outcome on the time-to-event outcome.

B.2 Simulation results of Chapter 4

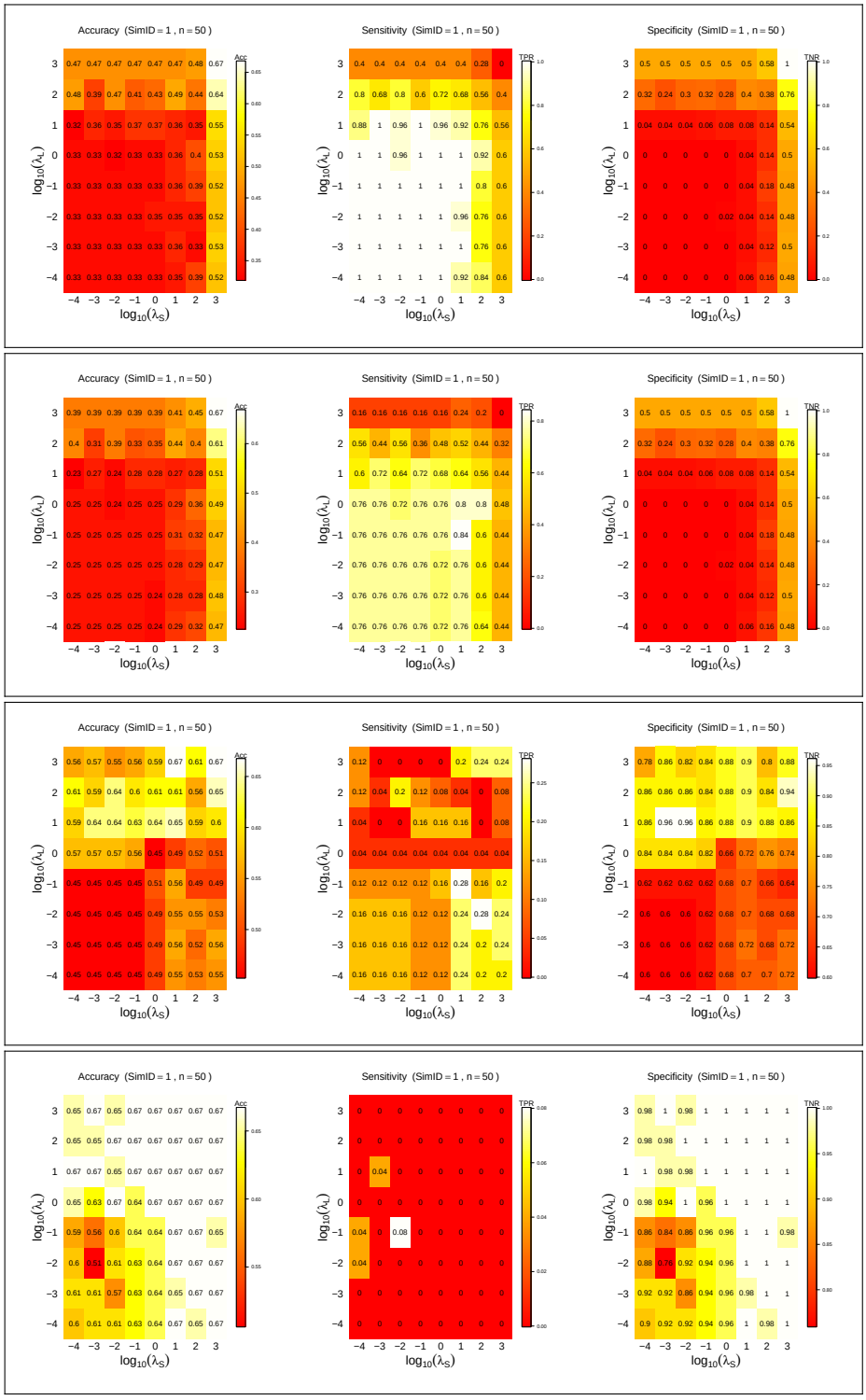


Figure B.2: Simulation 1, $M_{12}^{(S)}$. Performance metrics of the variable selection process. Top to bottom: Classifiers 1 to 4; left to right: Accuracy, Sensitivity and Specificity.

B.2.1 Performance metrics based on confusion matrix

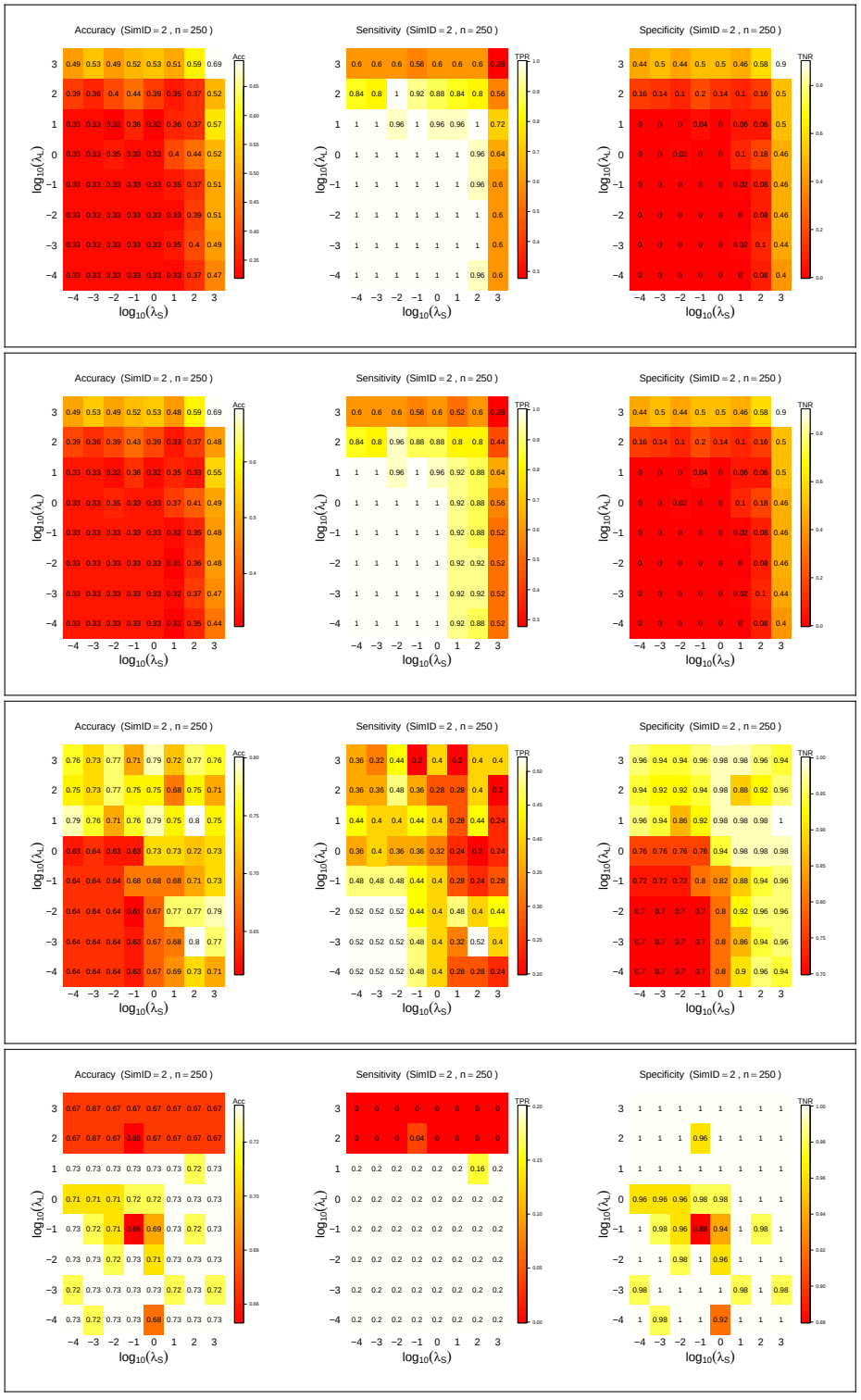


Figure B.3: Simulation 2, $M_{12}^{(S)}$. Performance metrics of the variable selection process. Top to bottom: Classifiers 1 to 4; left to right: Accuracy, Sensitivity and Specificity.

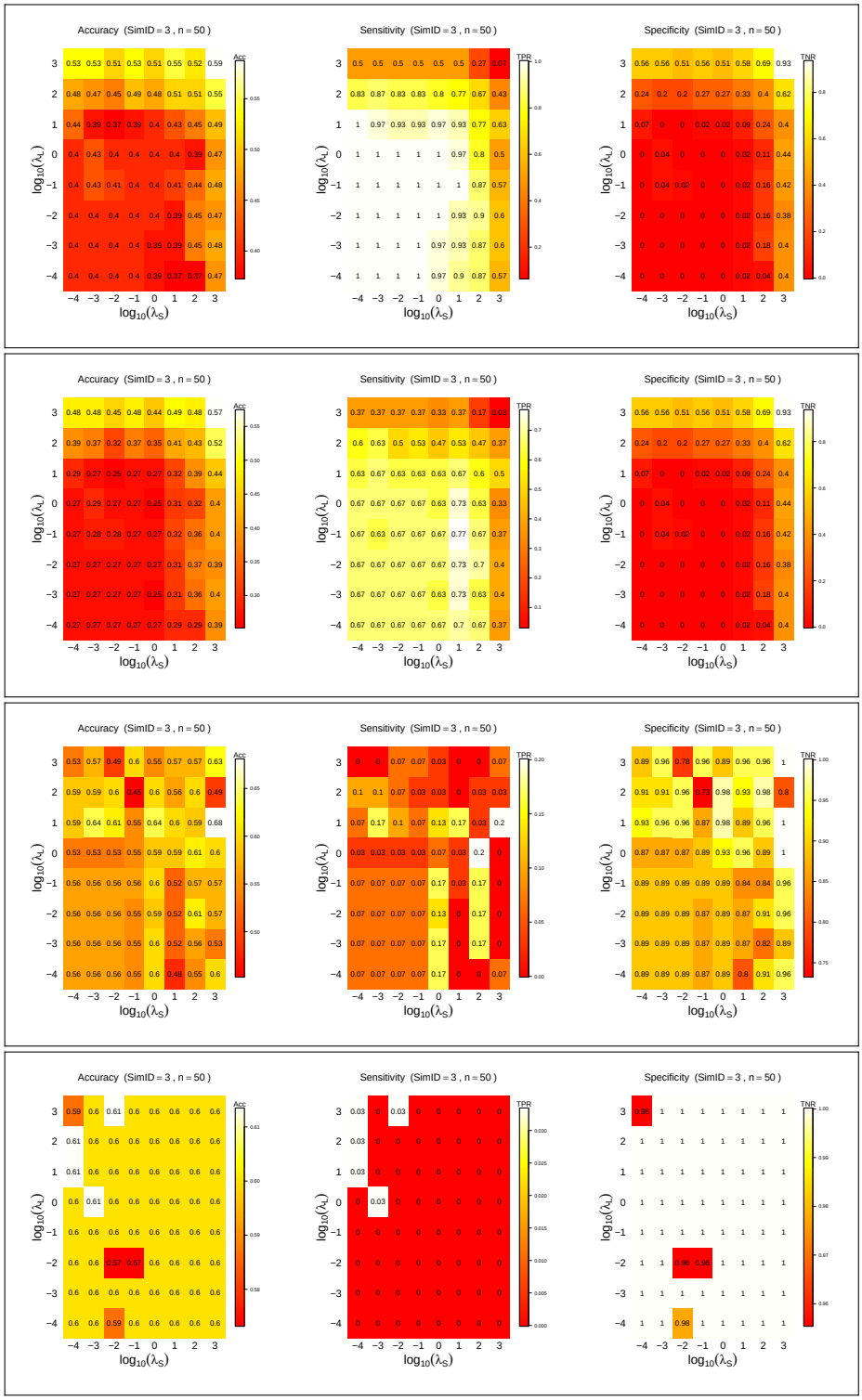


Figure B.4: Simulation 3, $M_{34}^{(S)}$. Performance metrics of the variable selection process. Top to bottom: Classifiers 1 to 4; left to right: Accuracy, Sensitivity and Specificity.

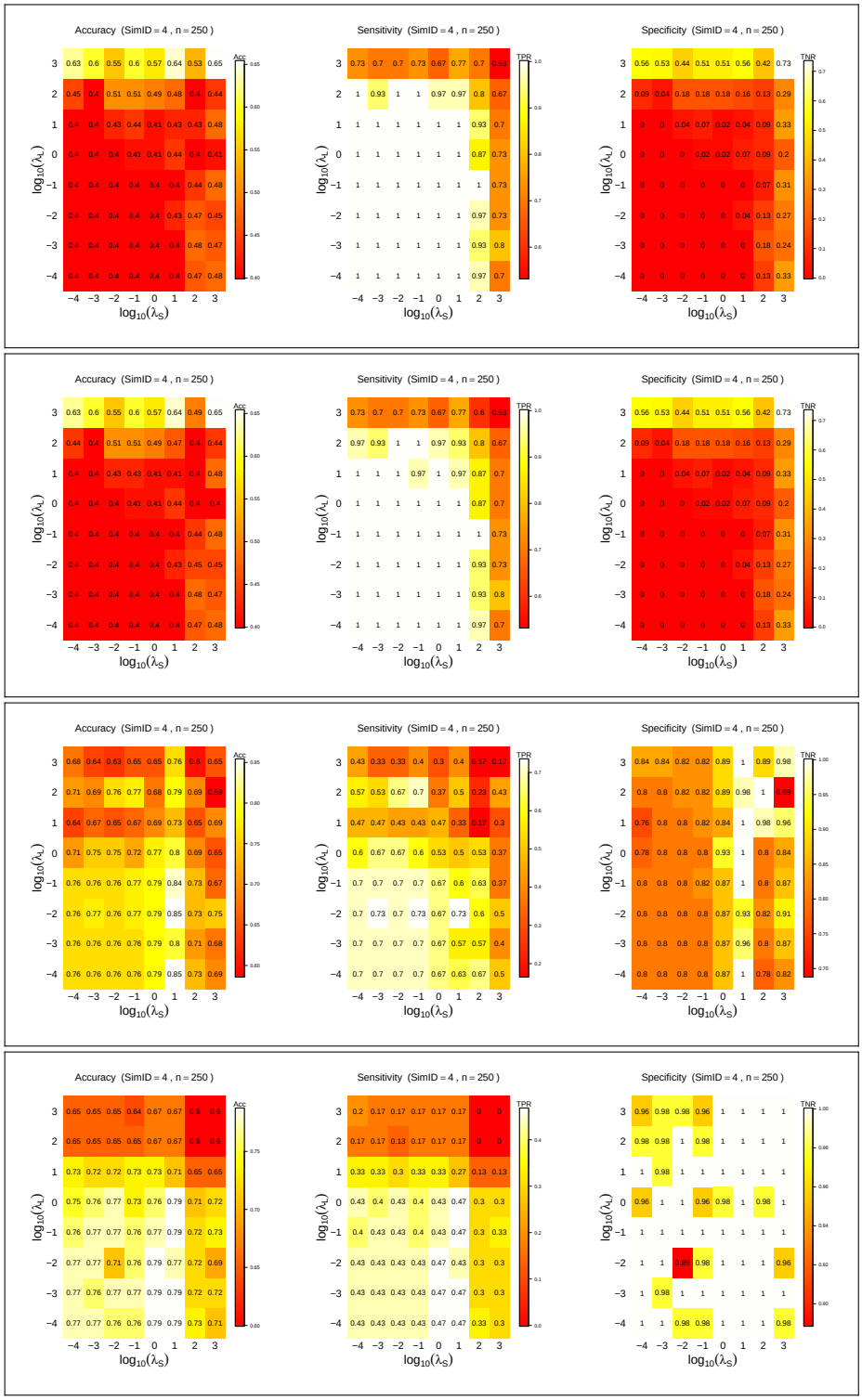


Figure B.5: Simulation 4, $M_{34}^{(S)}$. Performance metrics of the variable selection process. Top to bottom: Classifiers 1 to 4; left to right: Accuracy, Sensitivity and Specificity.

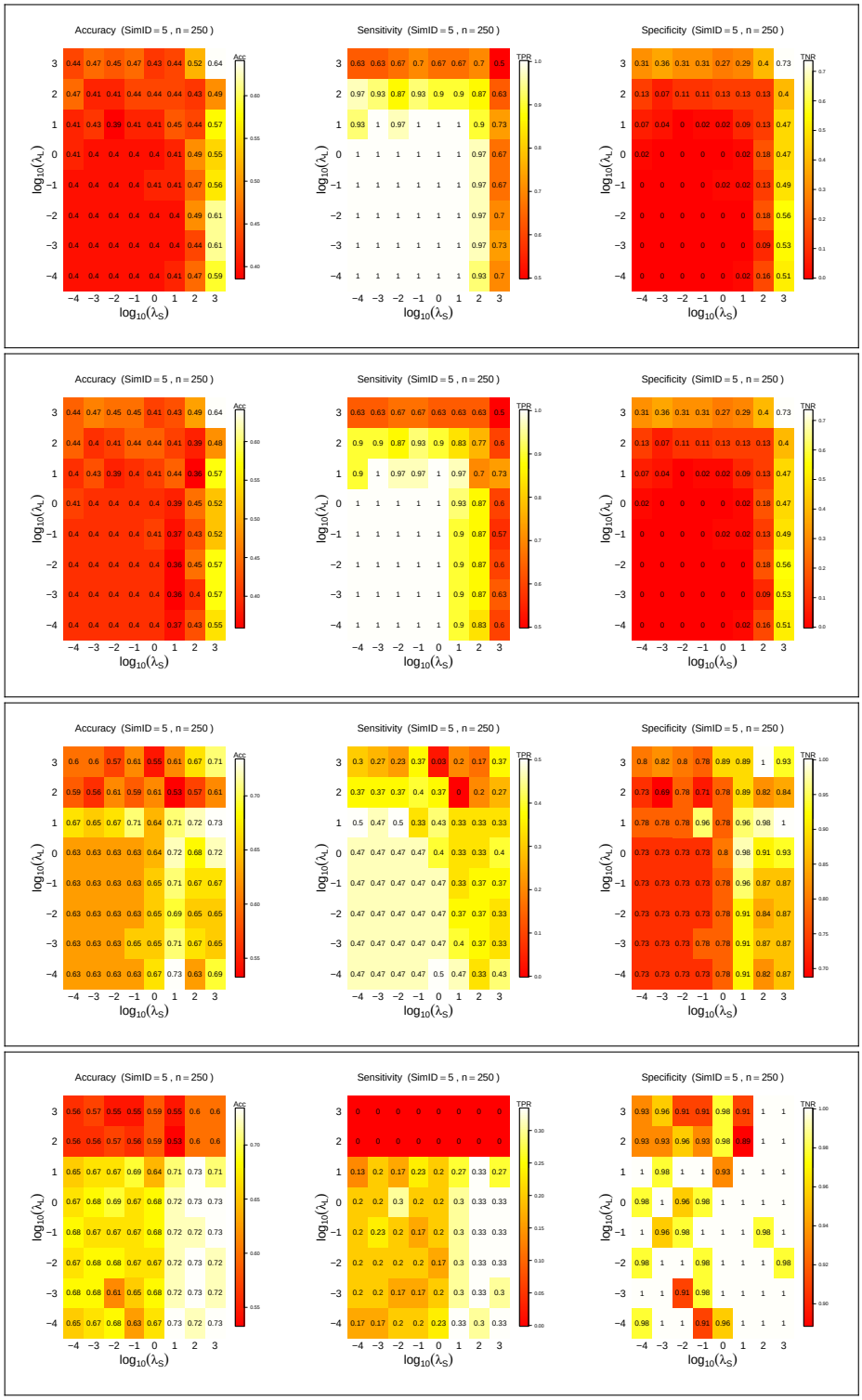


Figure B.6: Simulation 5, $M_5^{(S)}$. Performance metrics of the variable selection process. Top to bottom: Classifiers 1 to 4; left to right: Accuracy, Sensitivity and Specificity.

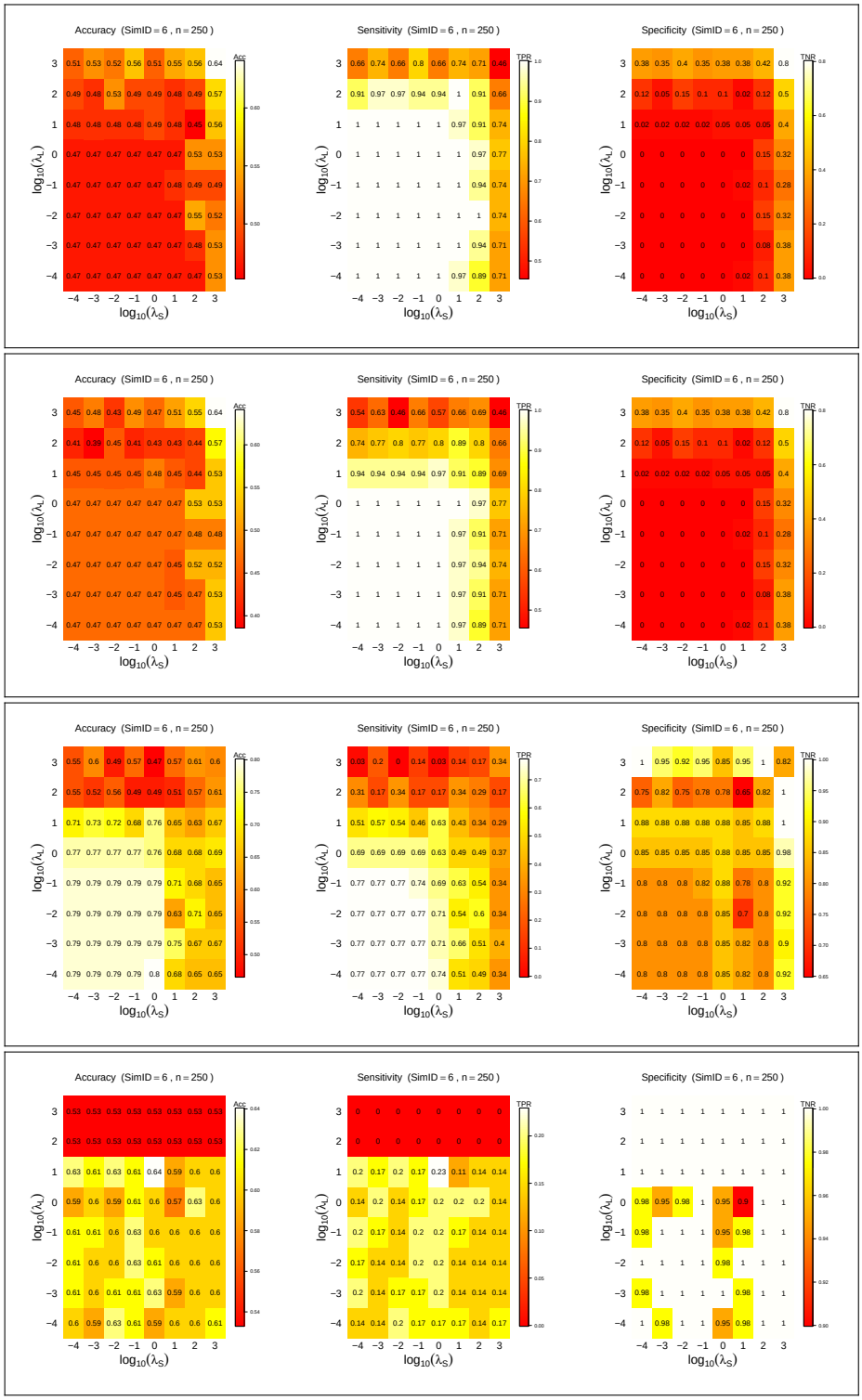


Figure B.7: Simulation 6, $M_6^{(S)}$. Performance metrics of the variable selection process. Top to bottom: Classifiers 1 to 4; left to right: Accuracy, Sensitivity and Specificity.

B.2.2 Regression coefficients of Simulation 6

Linear-mixed submodel $\left(\widehat{\beta}\right)$

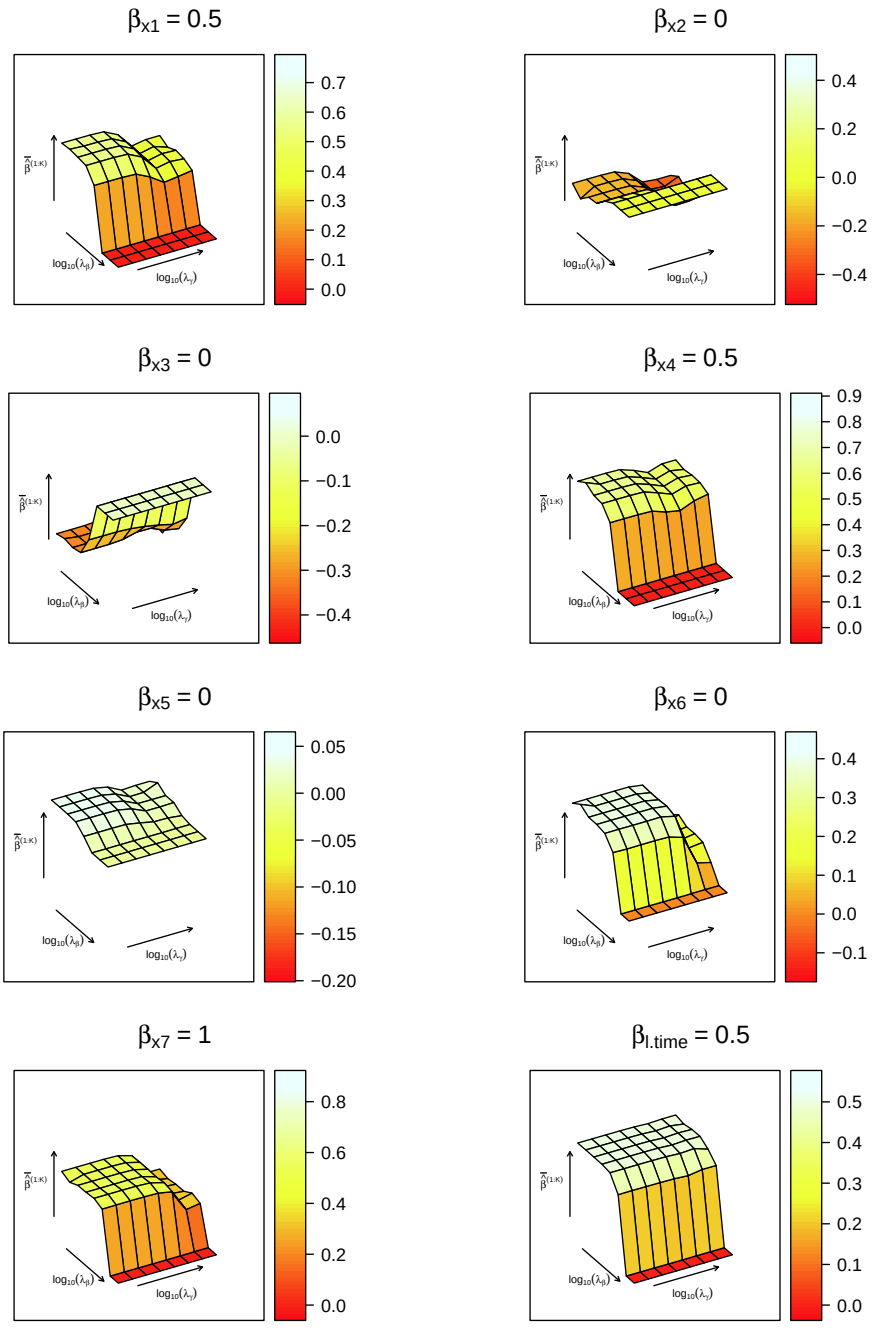


Figure B.8: Regression coefficients of the linear mixed submodel for each combination $\log_{10}\lambda_L, \log_{10}\lambda_S \in \{4, 3, 2, 1, 0, 1, 2, 3\}$.

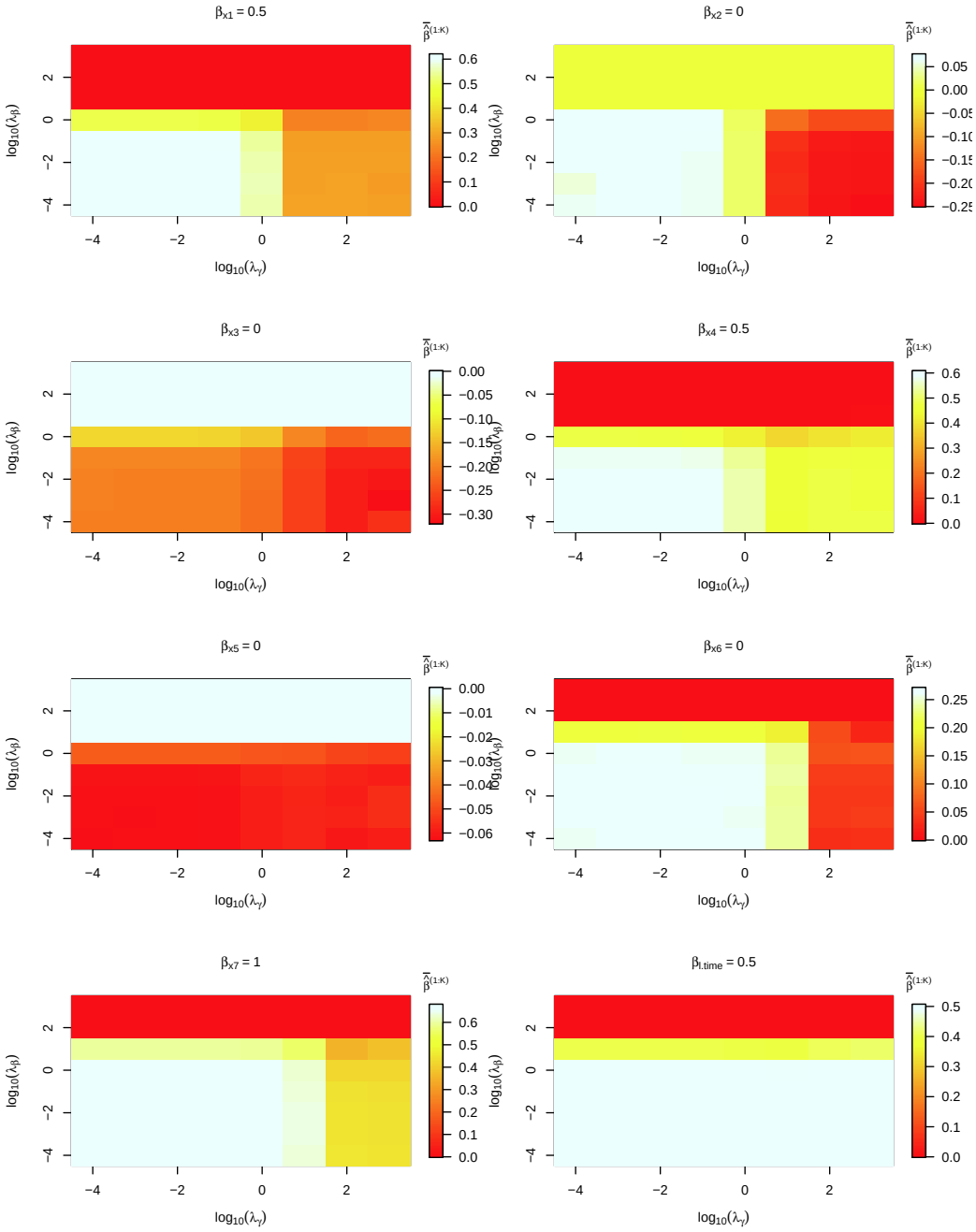


Figure B.9: Regression coefficients of the linear mixed submodel for each combination $\log_{10} \lambda_L, \log_{10} \lambda_S \in \{4, 3, 2, 1, 0, 1, 2, 3\}$.

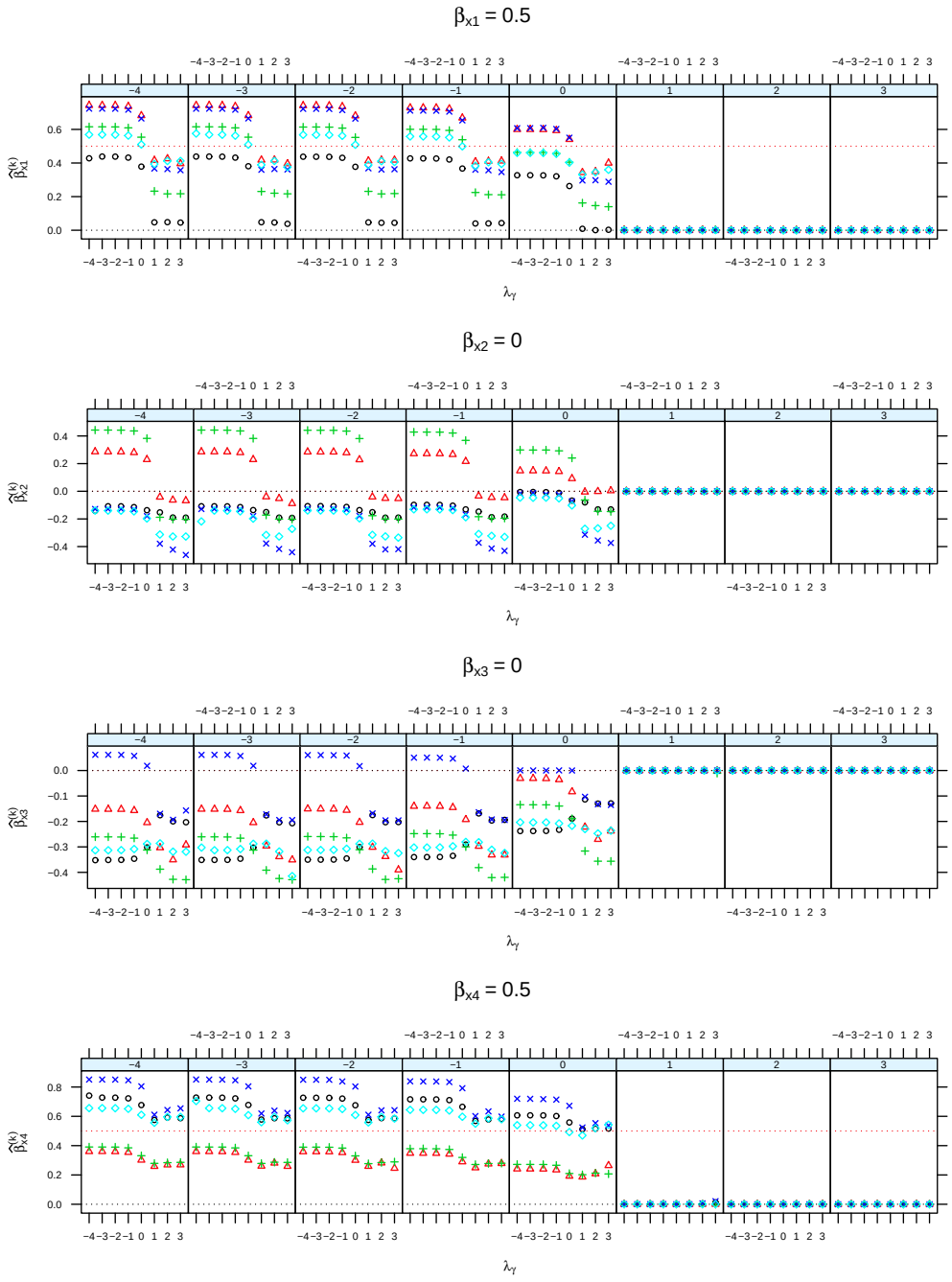


Figure B.10: Regression coefficients of the linear mixed submodel of the K times the model was tested for each penalty combination, $\log_{10}\lambda_L, \log_{10}\lambda_S \in \{4, 3, 2, 1, 0, 1, 2, 3\}$.

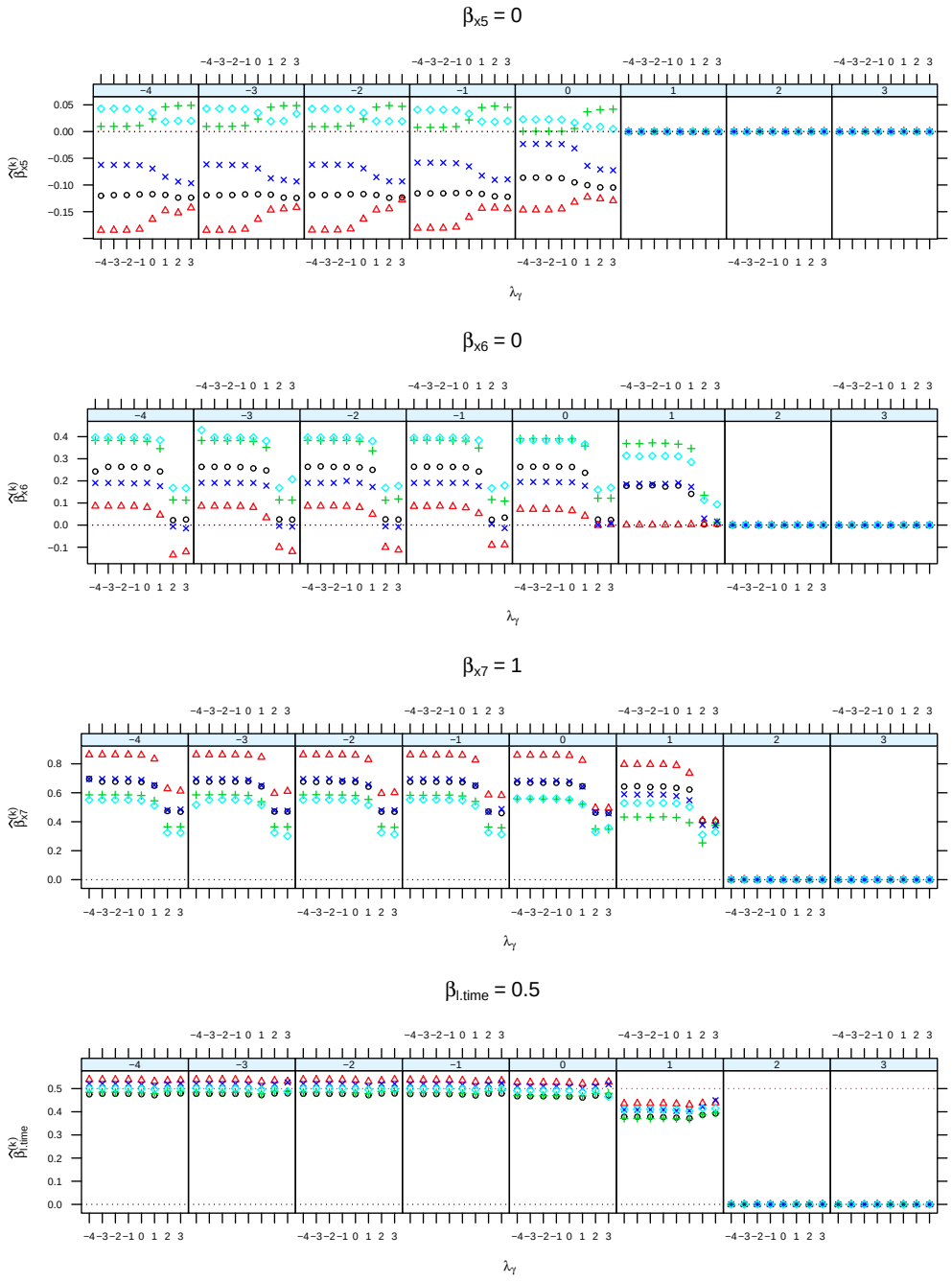


Figure B.11: Regression coefficients of the linear mixed submodel of the K times the model was tested for each penalty combination, $\log_{10}\lambda_L, \log_{10}\lambda_S \in \{4, 3, 2, 1, 0, 1, 2, 3\}$.

Time-to-event submodel $(\widehat{\gamma})$

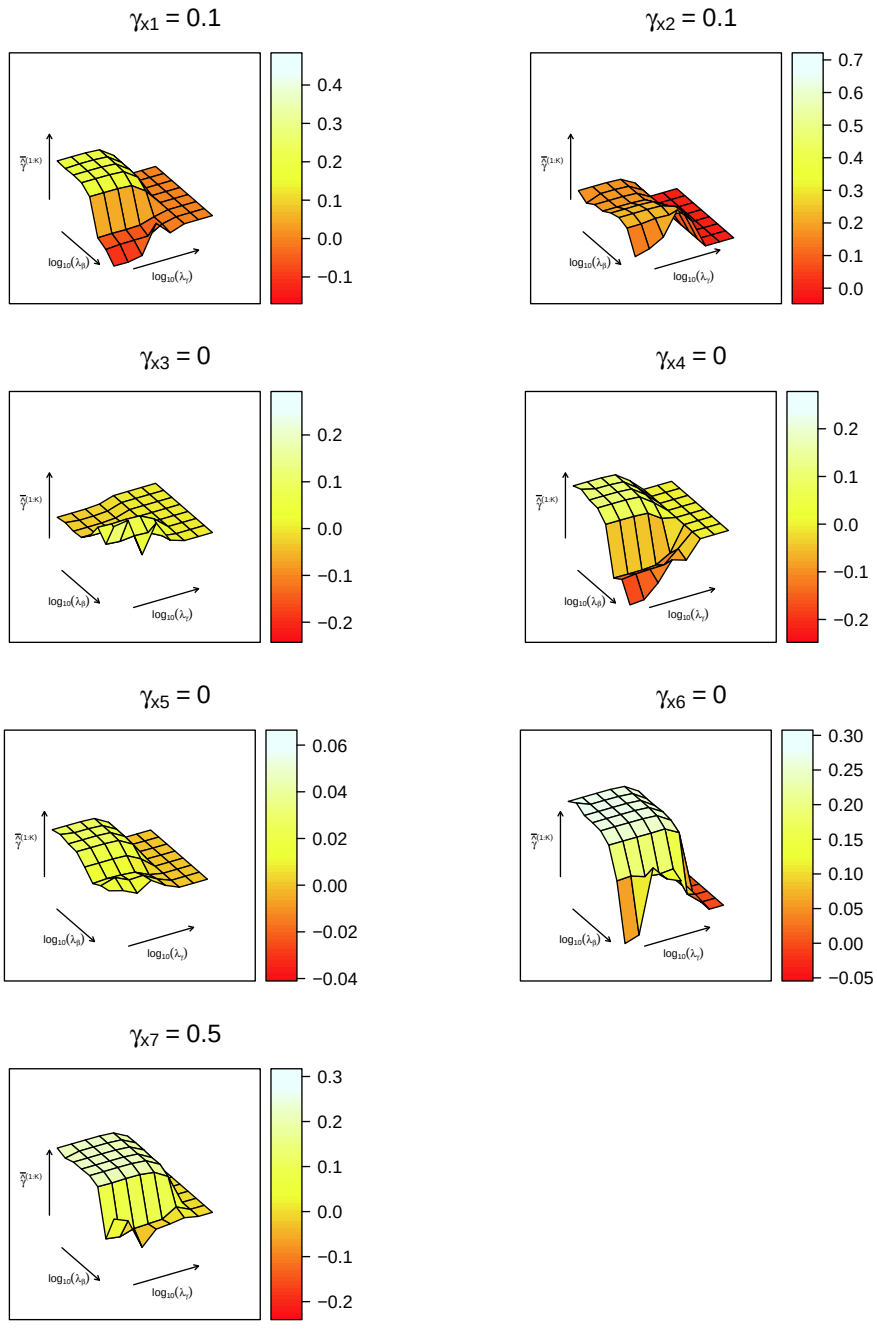


Figure B.12: Regression coefficients of the time-to-event submodel for each combination $\log_{10}\lambda_L, \log_{10}\lambda_S \in \{4, 3, 2, 1, 0, 1, 2, 3\}$.

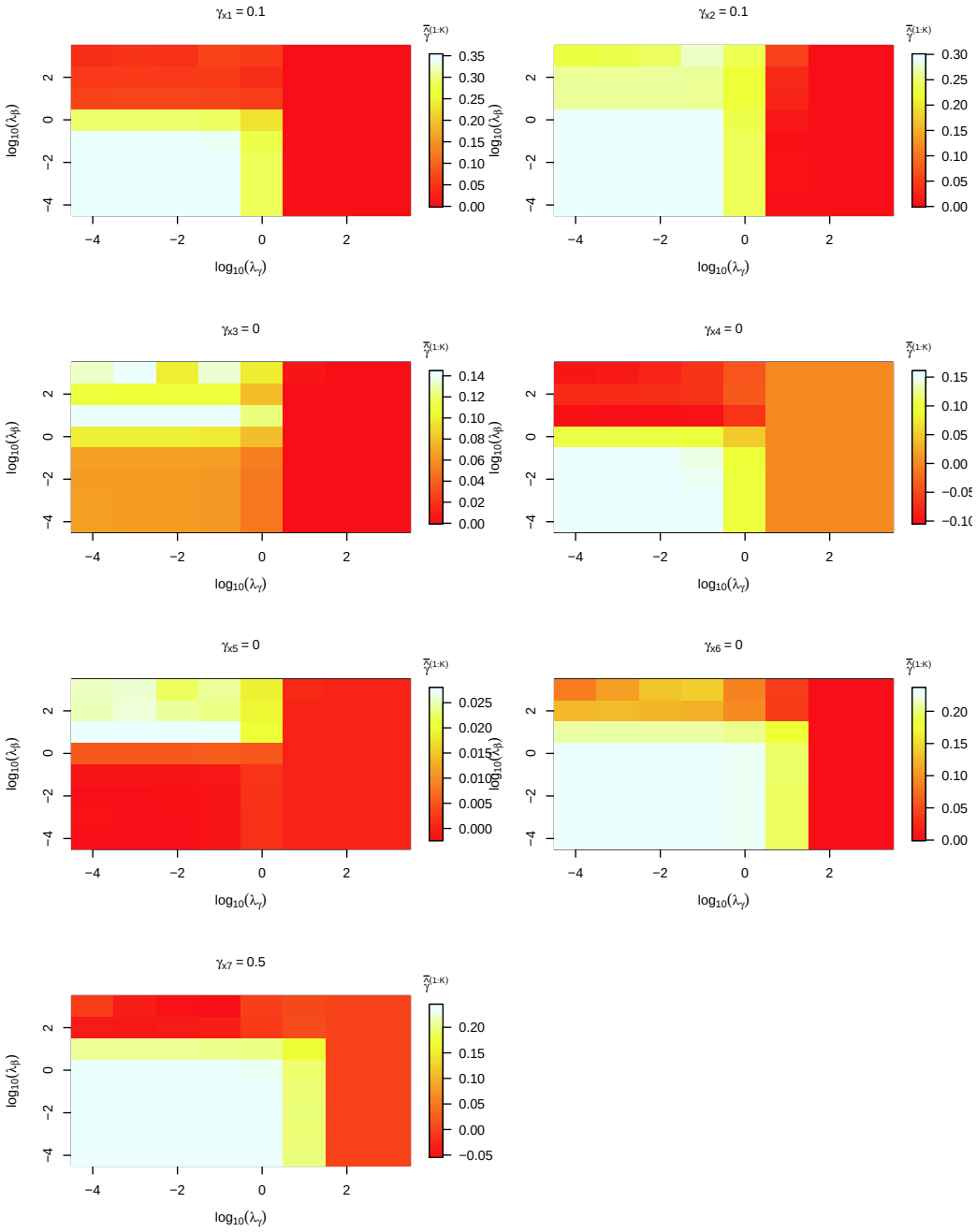


Figure B.13: Regression coefficients of the time-to-event submodel for each combination $\log_{10}\lambda_L, \log_{10}\lambda_S \in \{4, 3, 2, 1, 0, 1, 2, 3\}$.

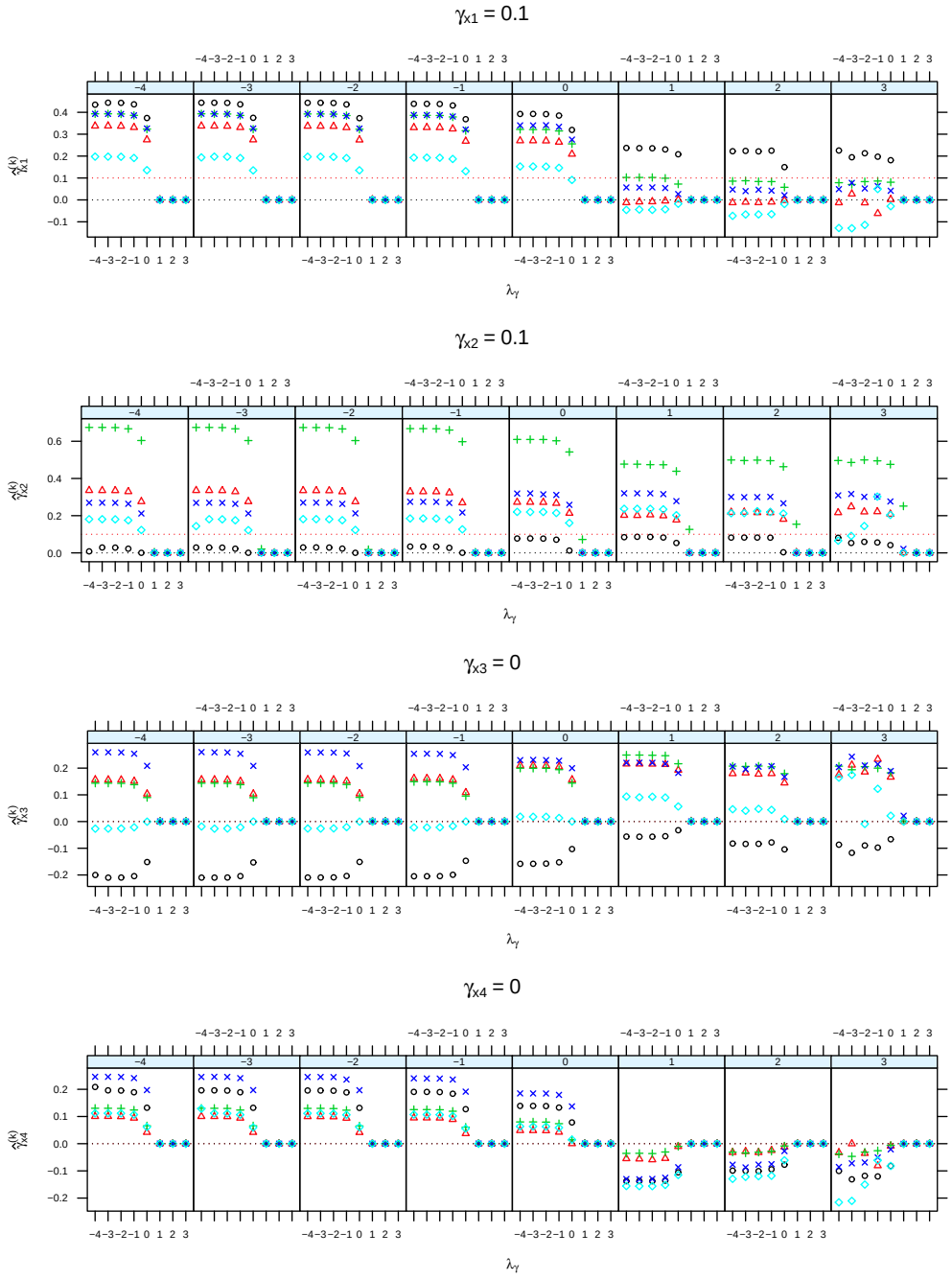


Figure B.14: Regression coefficients of the time-to-event submodel of the K times the model was tested for each penalty combination, $\log_{10}\lambda_L, \log_{10}\lambda_S \in \{4, 3, 2, 1, 0, 1, 2, 3\}$.

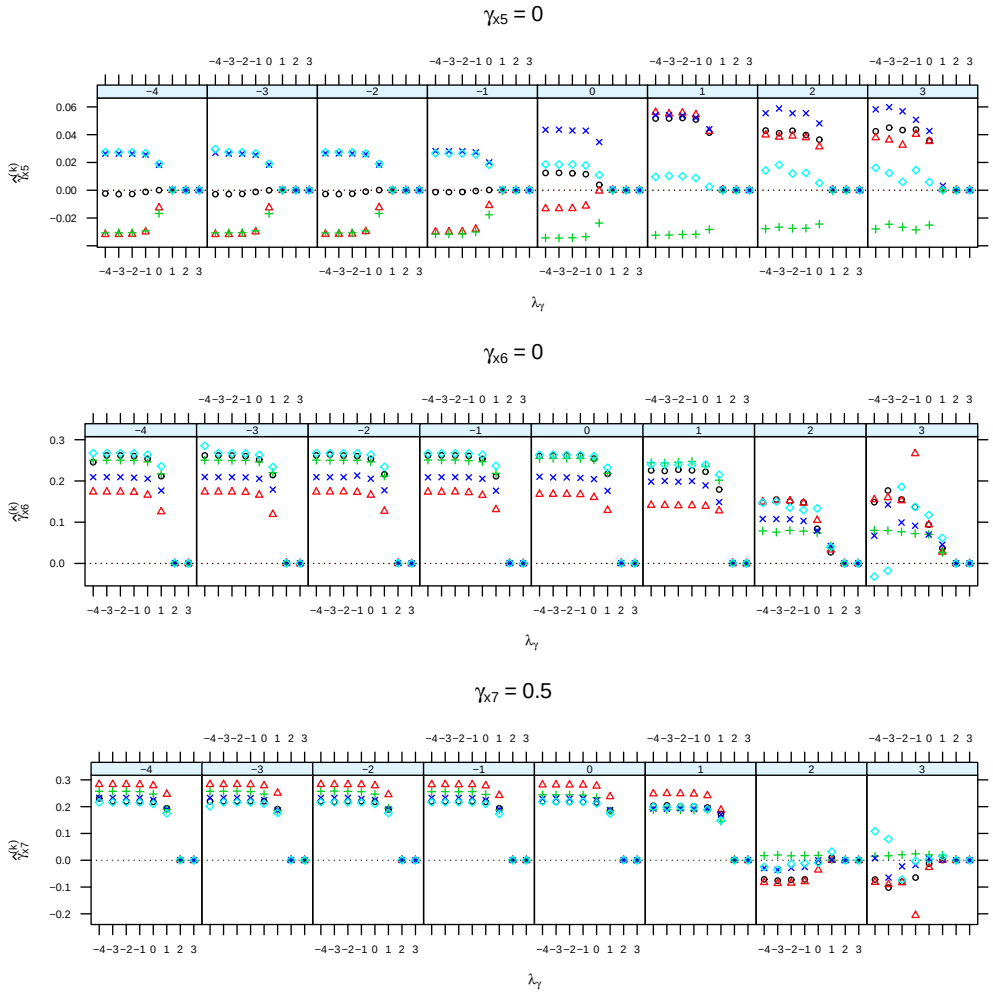


Figure B.15: Regression coefficients of the time-to-event submodel of the K times the model was tested for each penalty combination, $\log_{10}\lambda_L, \log_{10}\lambda_S \in \{4, 3, 2, 1, 0, 1, 2, 3\}$.

B.3 Simulation results of Chapter 5

B.3.1 Analysis data simulated from model M_3

M_3 fitted to D_3

M_2 fitted to D_3

Compare M_3 and M_2 fitted to D_3

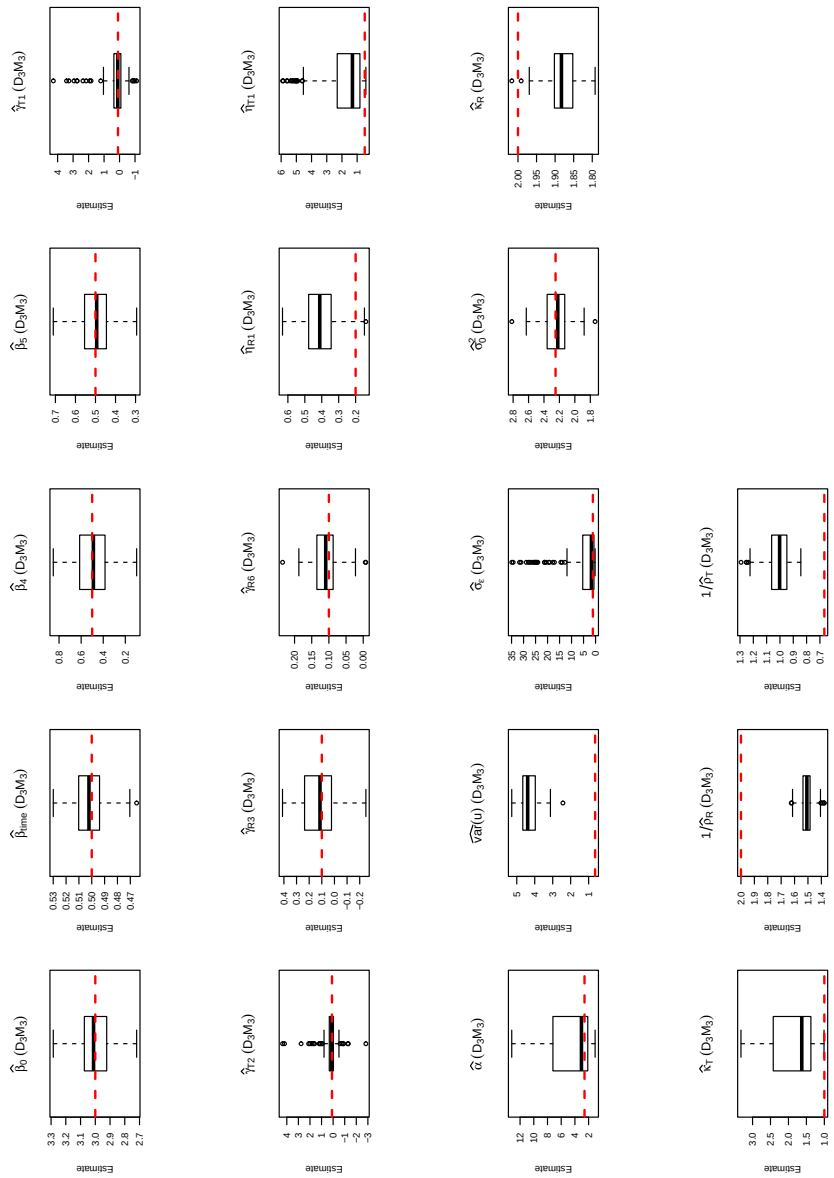


Figure B.16: Parameter estimates of model M_3 fitted to data D_3.

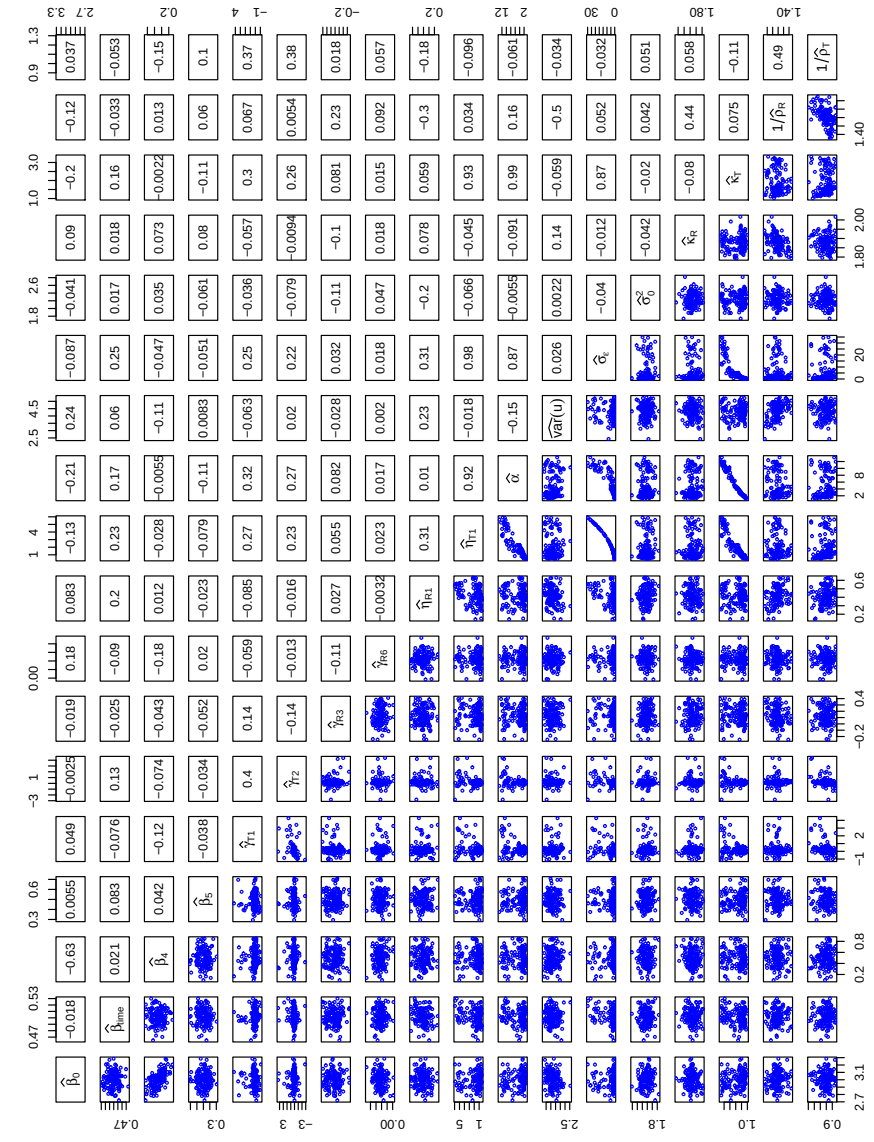


Figure B.17: Pairwise scatter plots of parameter estimates of model M_3 fitted to data D_3.

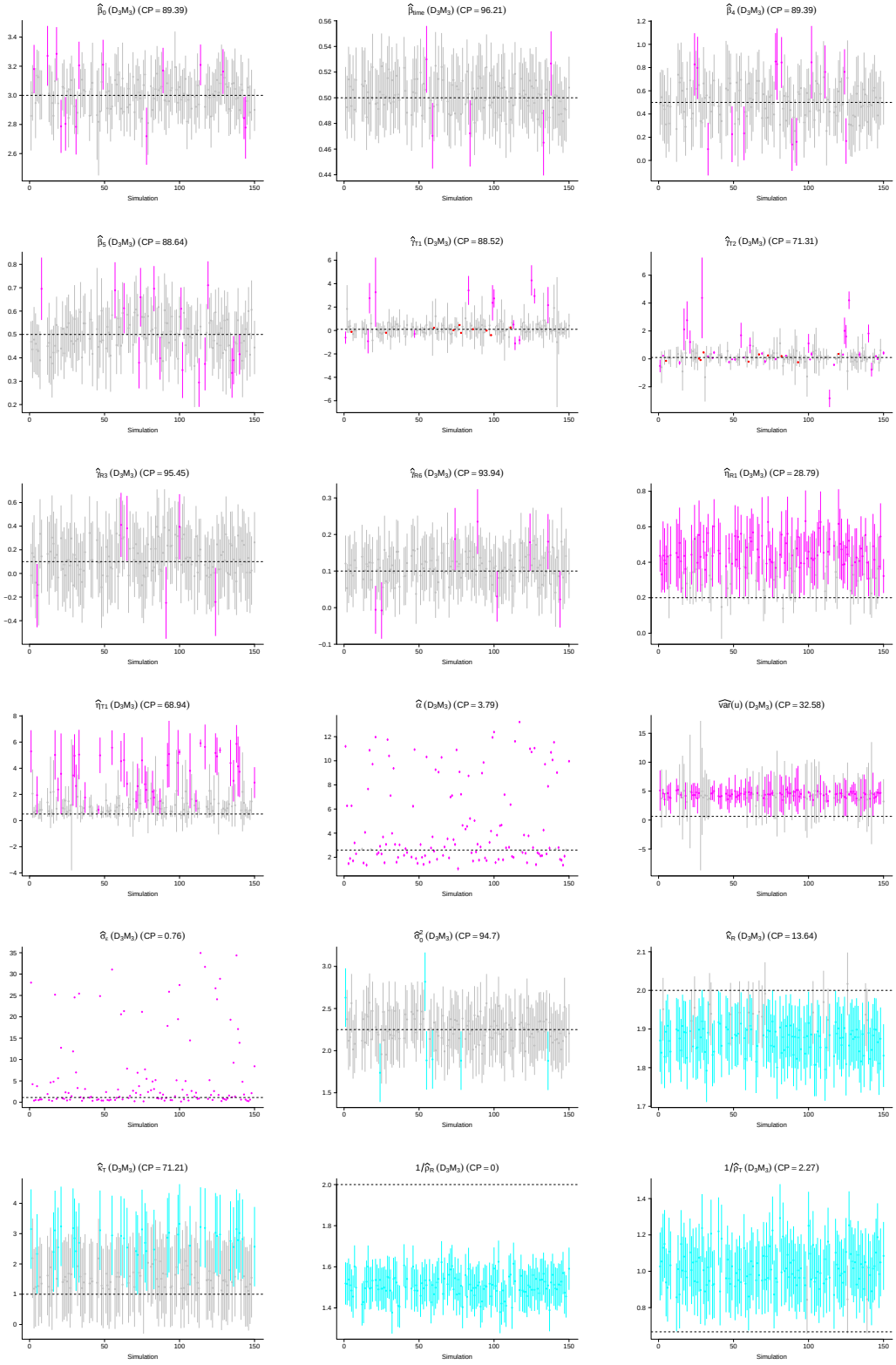


Figure B.18: Model 95% interval estimates of model M_3 fitted to data D_3. Dashed horizontal lines at the true parameter value. Vertical lines represent 95% interval estimates (gray if the range contains the true parameter, colored otherwise).

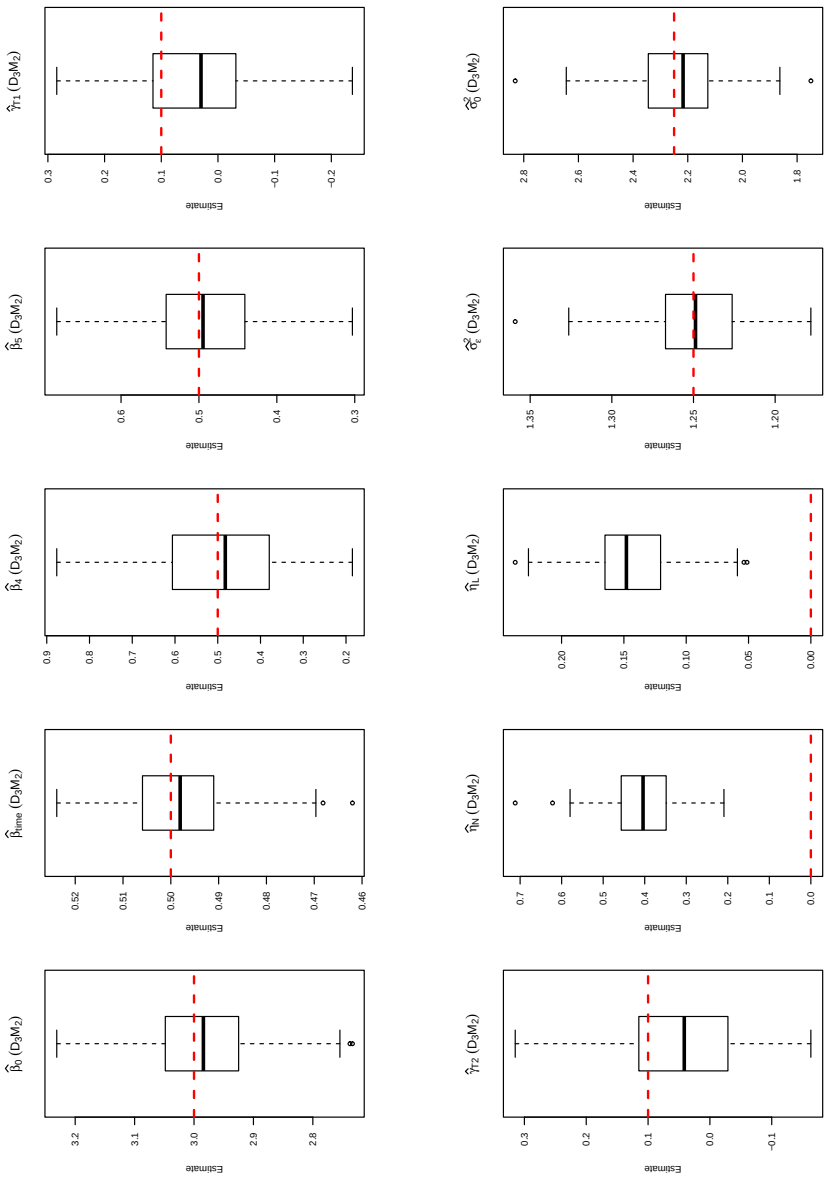


Figure B.19: Parameter estimates of model M_2 fitted to data D_3.

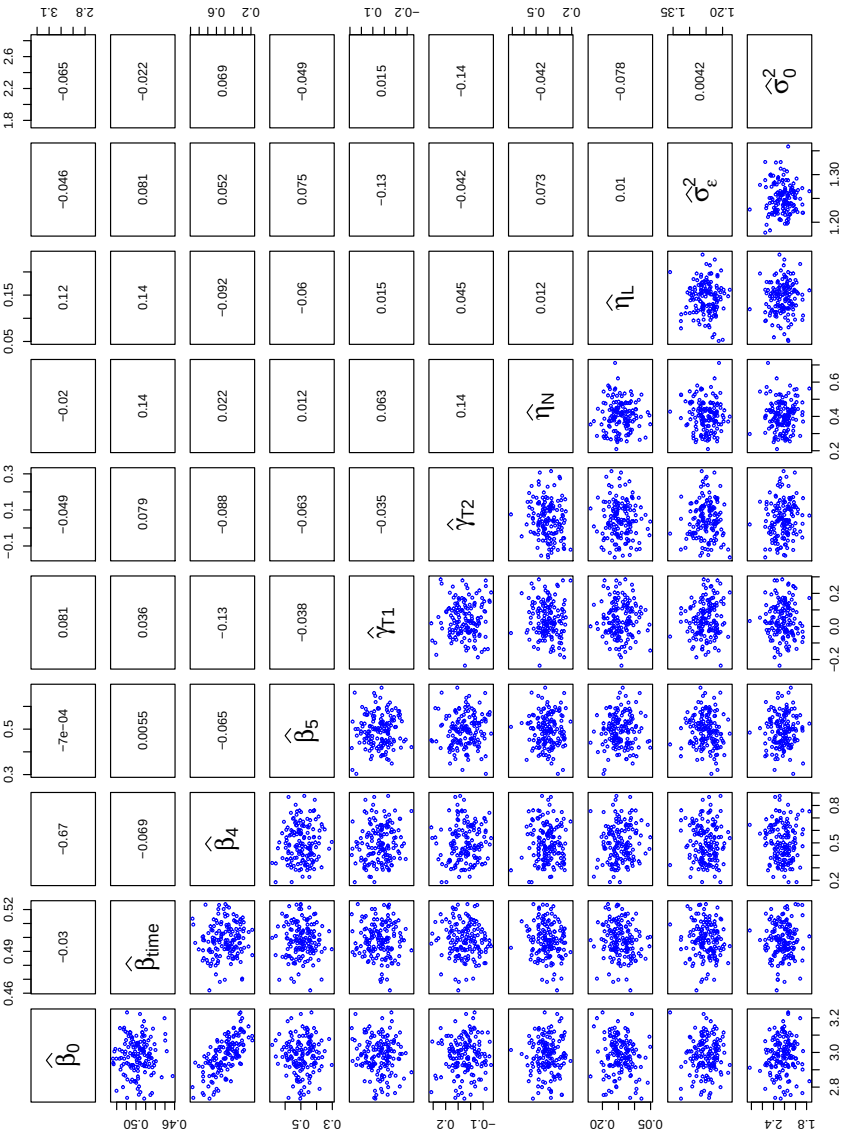


Figure B.20: Pairwise scatter plots of parameter estimates of model M_2 fitted to data D_3.

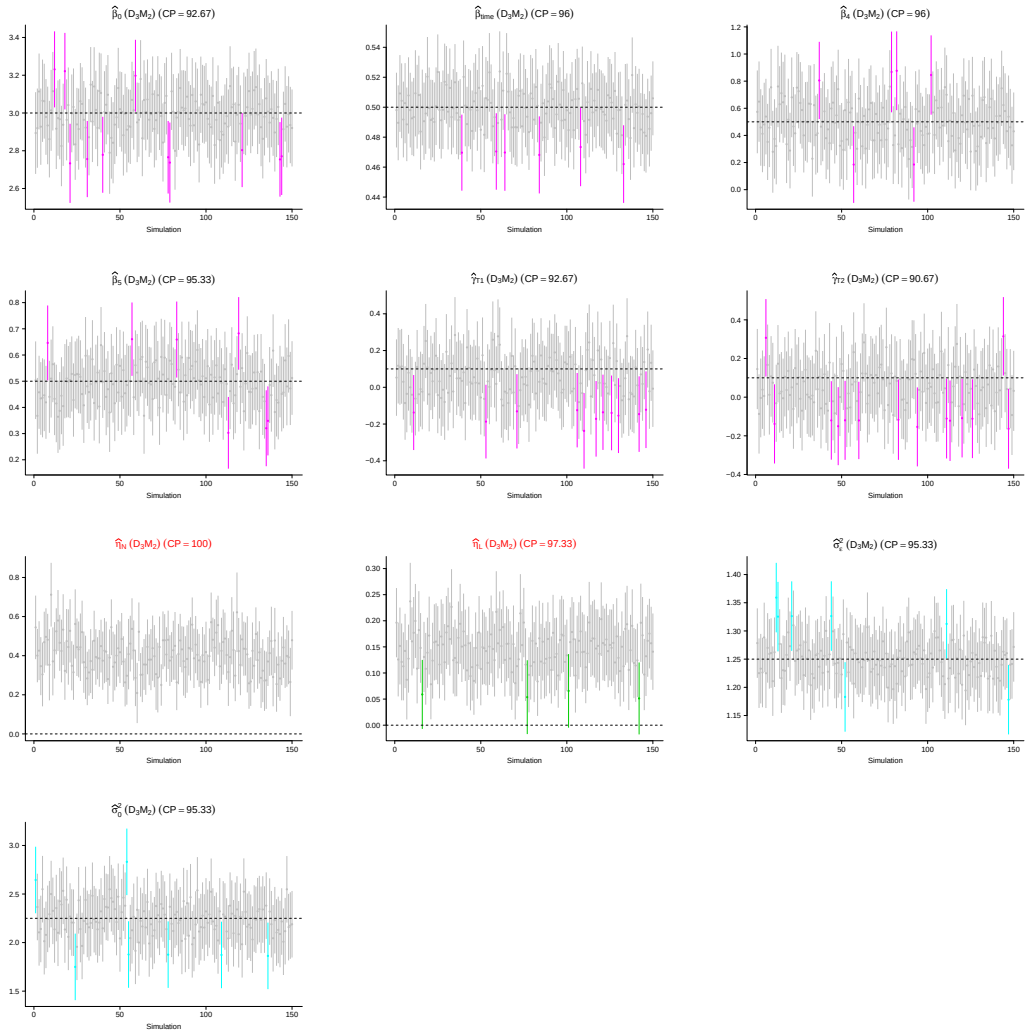


Figure B.21: Model M_2 point and 95% interval estimates for M_3. Dashed horizontal lines are drawn at the true parameter value. Vertical lines represent 95% interval estimates (gray if the range contains the true parameter, colored otherwise). A title in red means that the association parameter in model M_2 measures something different than in model M_3, and the number in CP refers to the relative frequency in the whole range of the interval lies agrees with the sign of the true value of the association parameter.

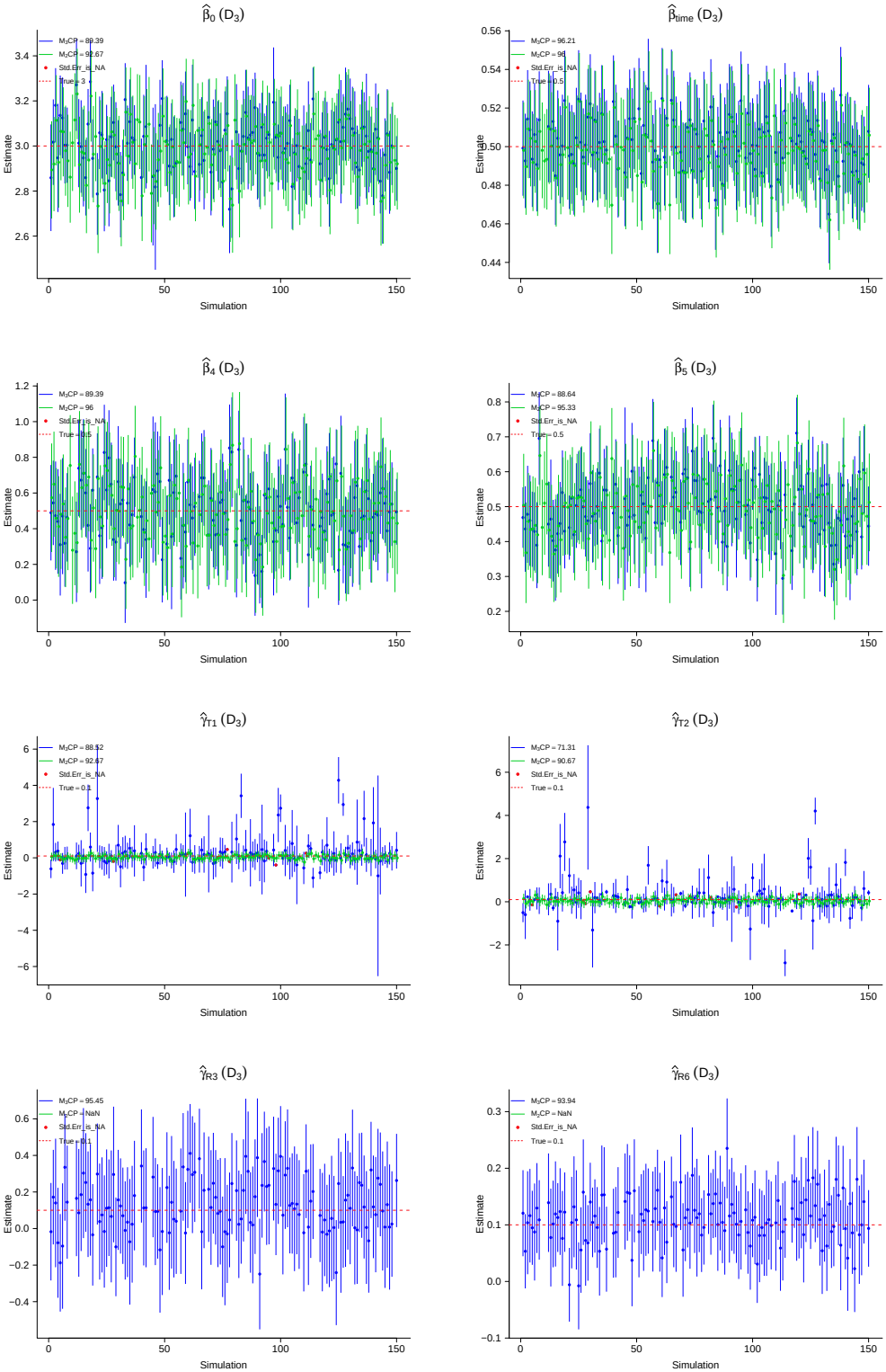


Figure B.22: Fixed effects regression coefficients of data produced with model M_3. Compare model fit M_3 and M_2 on M_3.

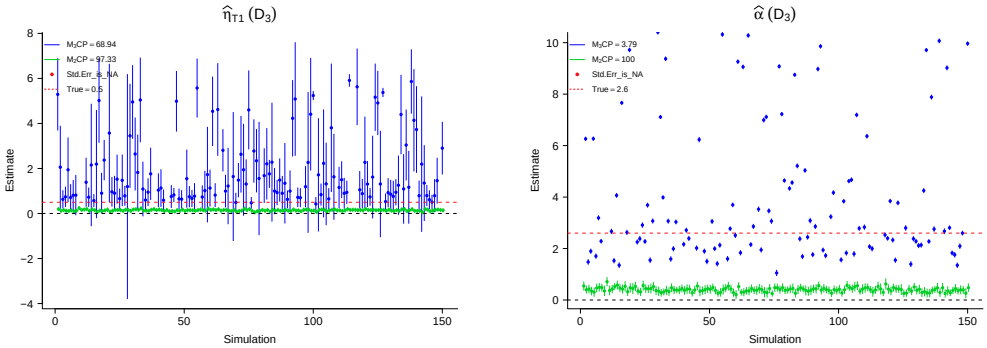


Figure B.23: Association parameters of data produced with model M_3. Compare model fit M_3 and M_2 on D_3.

B.3.2 Analysis data simulated from model M_2

M_2 fitted to D_2

M_3 fitted to D_2

Compare M_2 and M_3 fitted to D_2

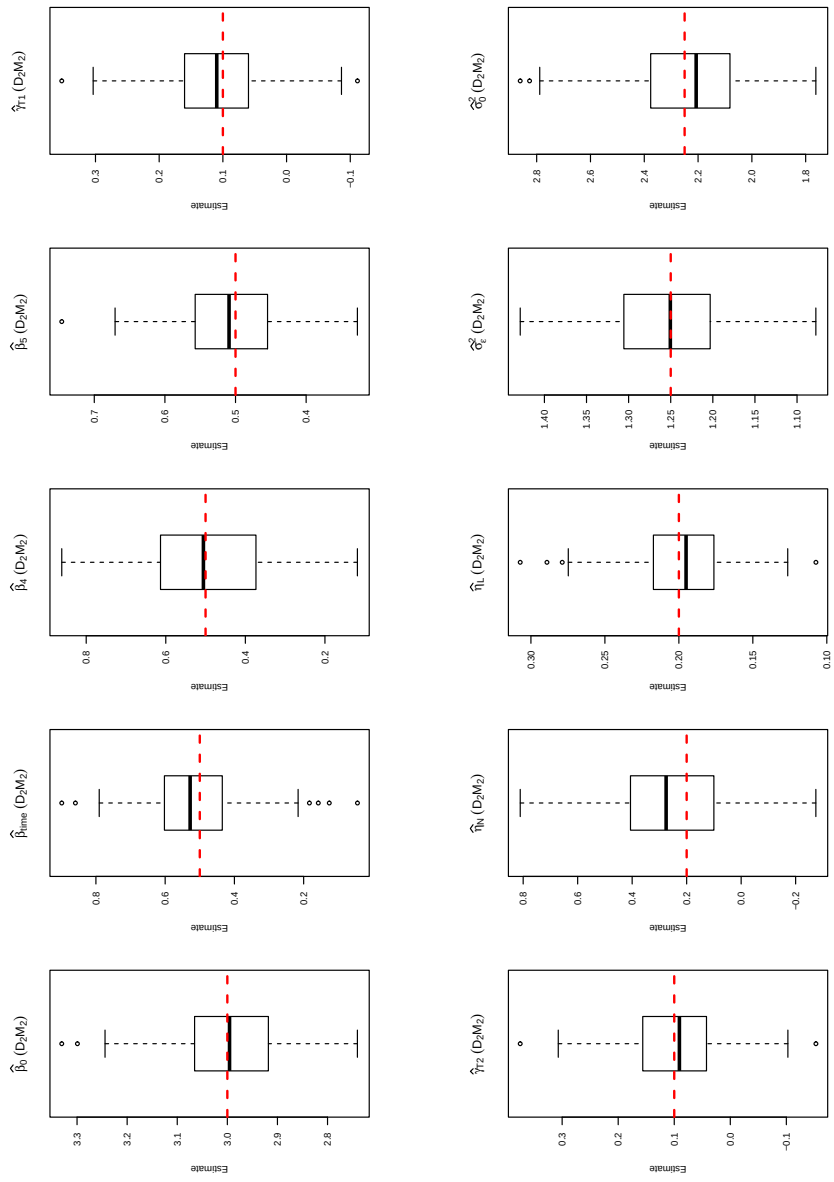


Figure B.24: Parameter estimates of model M_2 fitted to data D_2.

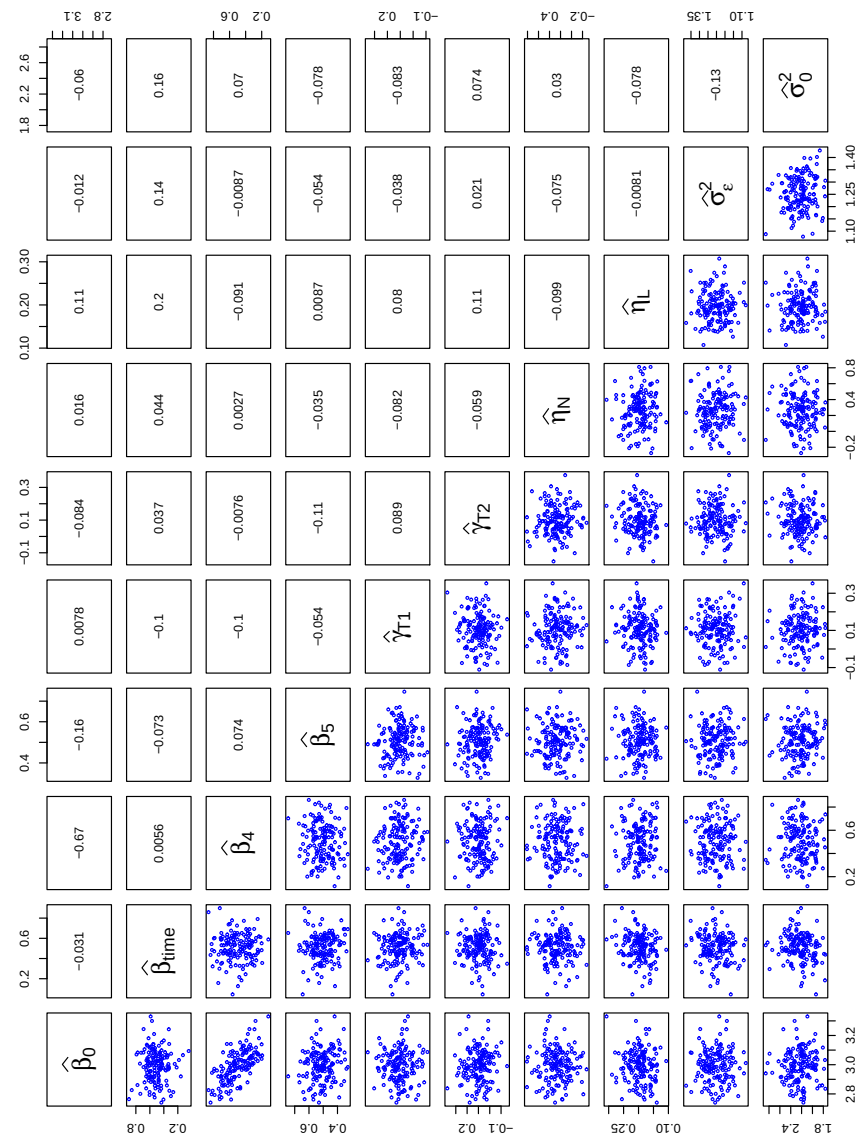


Figure B.25: Pairwise scatter plots of parameter estimates of model M_2 fitted to data D_2.

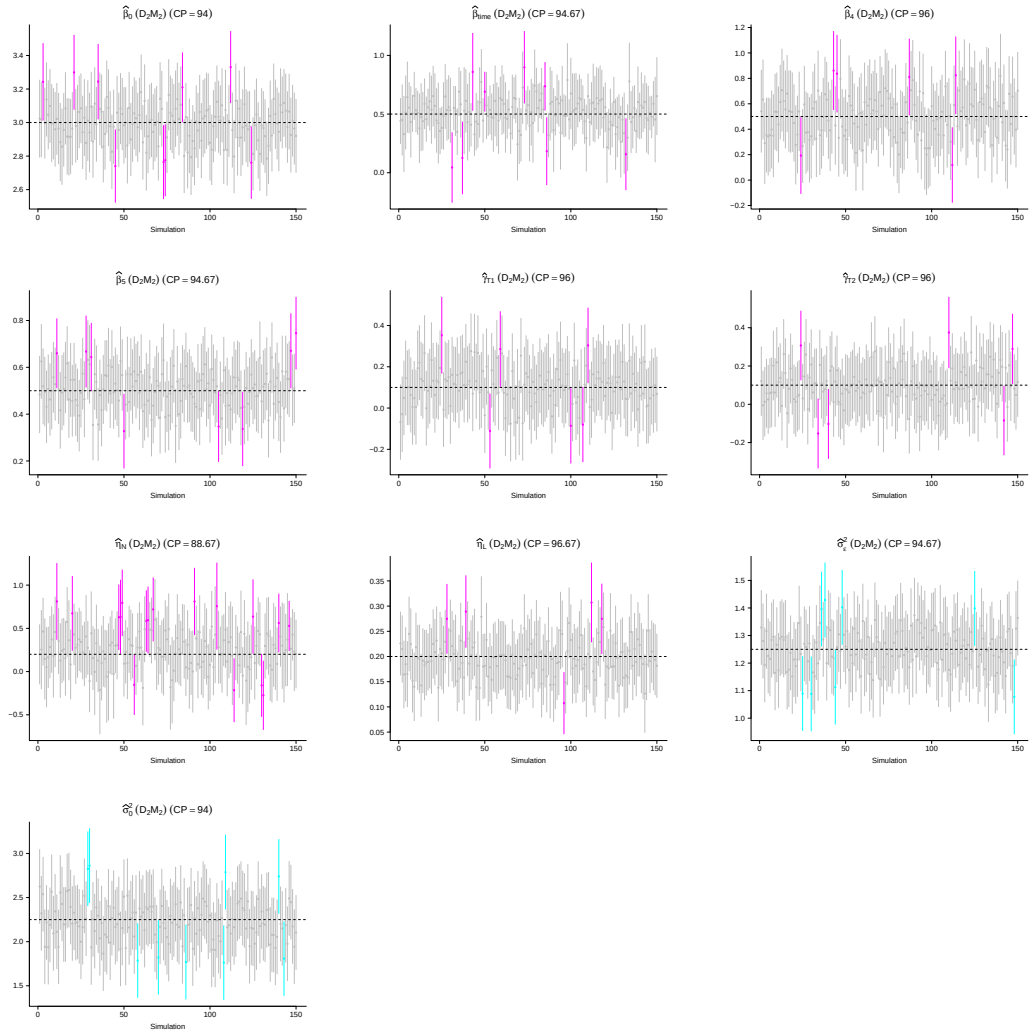


Figure B.26: Model M_2 point and 95% interval estimates for D_2. Dashed horizontal lines are drawn at the true parameter value. Vertical lines represent 95% interval estimates (gray if the range contains the true parameter, colored otherwise).

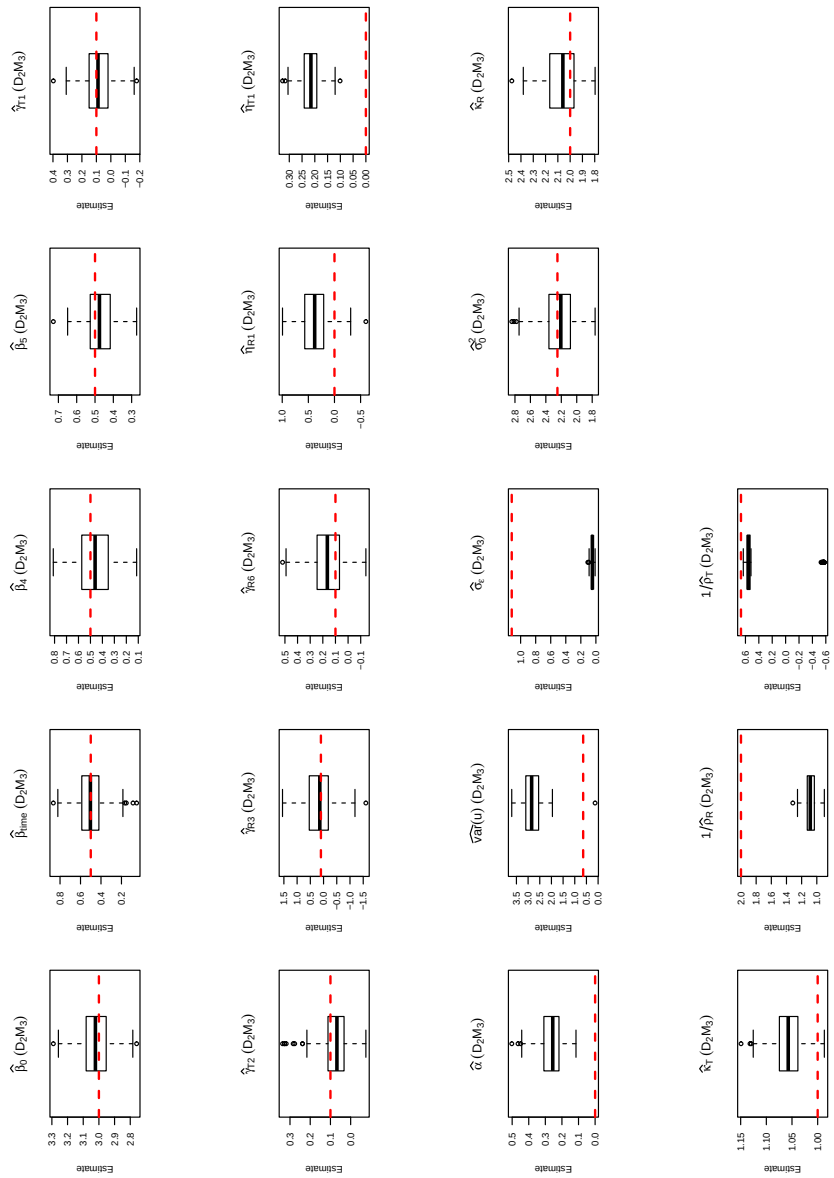


Figure B.27: Parameter estimates of model M_3 fitted to data D_2.

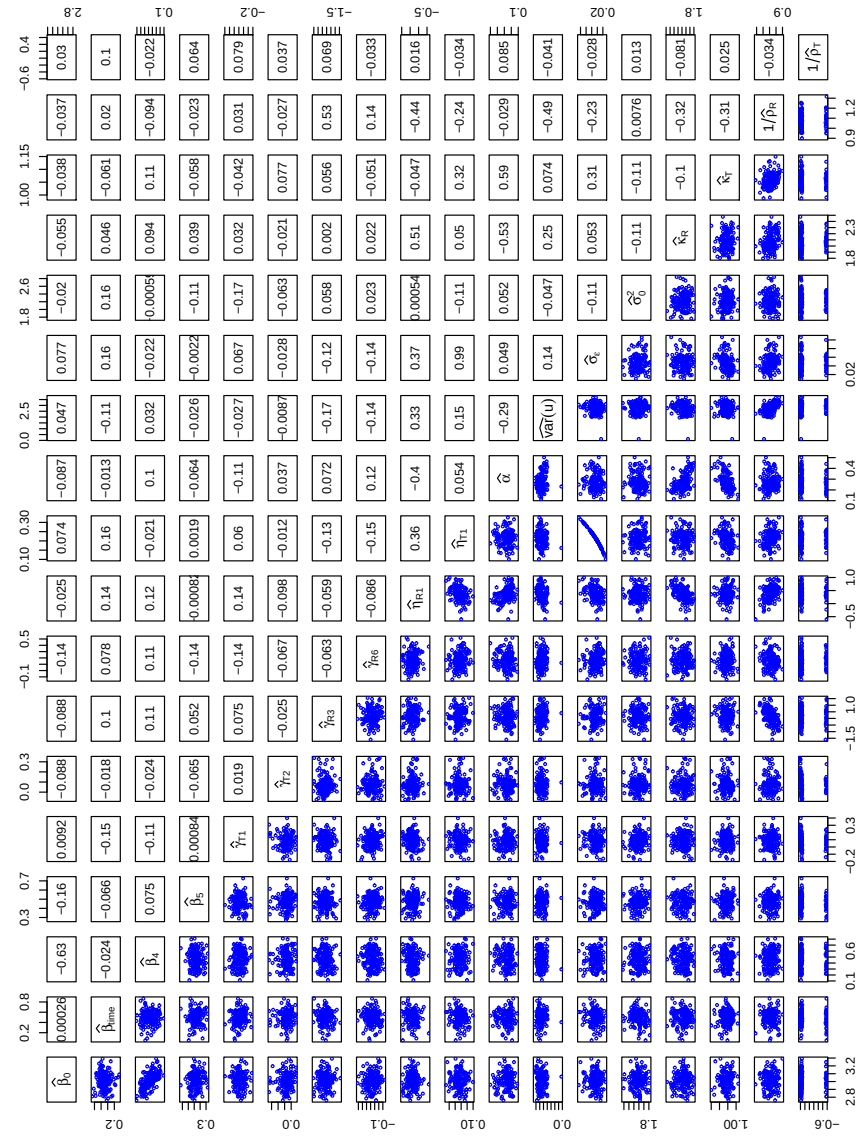


Figure B.28: Pairwise scatter plots of parameter estimates of model M_3 fitted to data D_2.

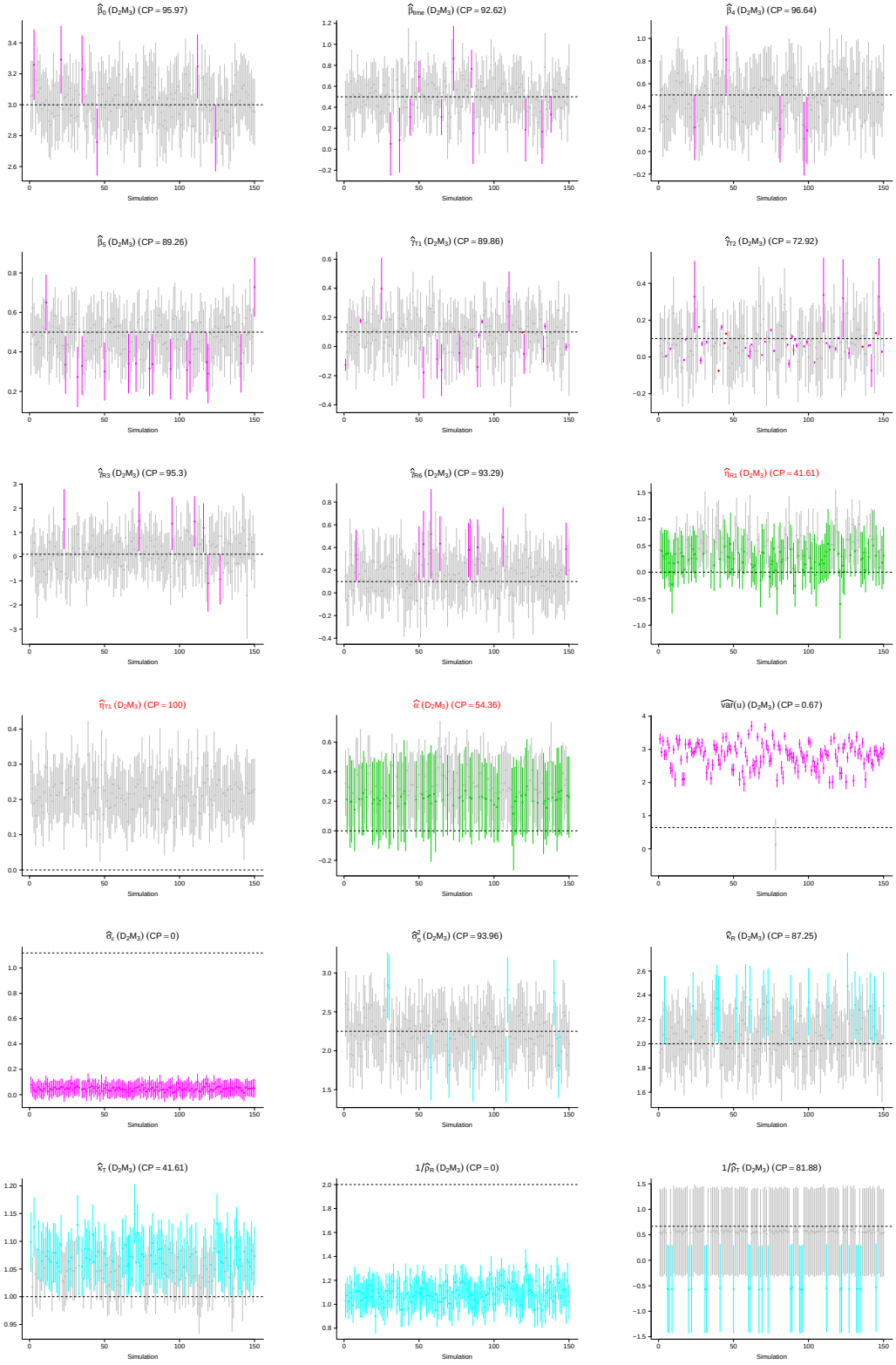


Figure B.29: Model M_3 point and 95% interval estimates for D_2. Dashed horizontal lines are drawn at the true parameter value. Vertical lines represent 95% interval estimates (gray if the range contains the true parameter, colored otherwise).

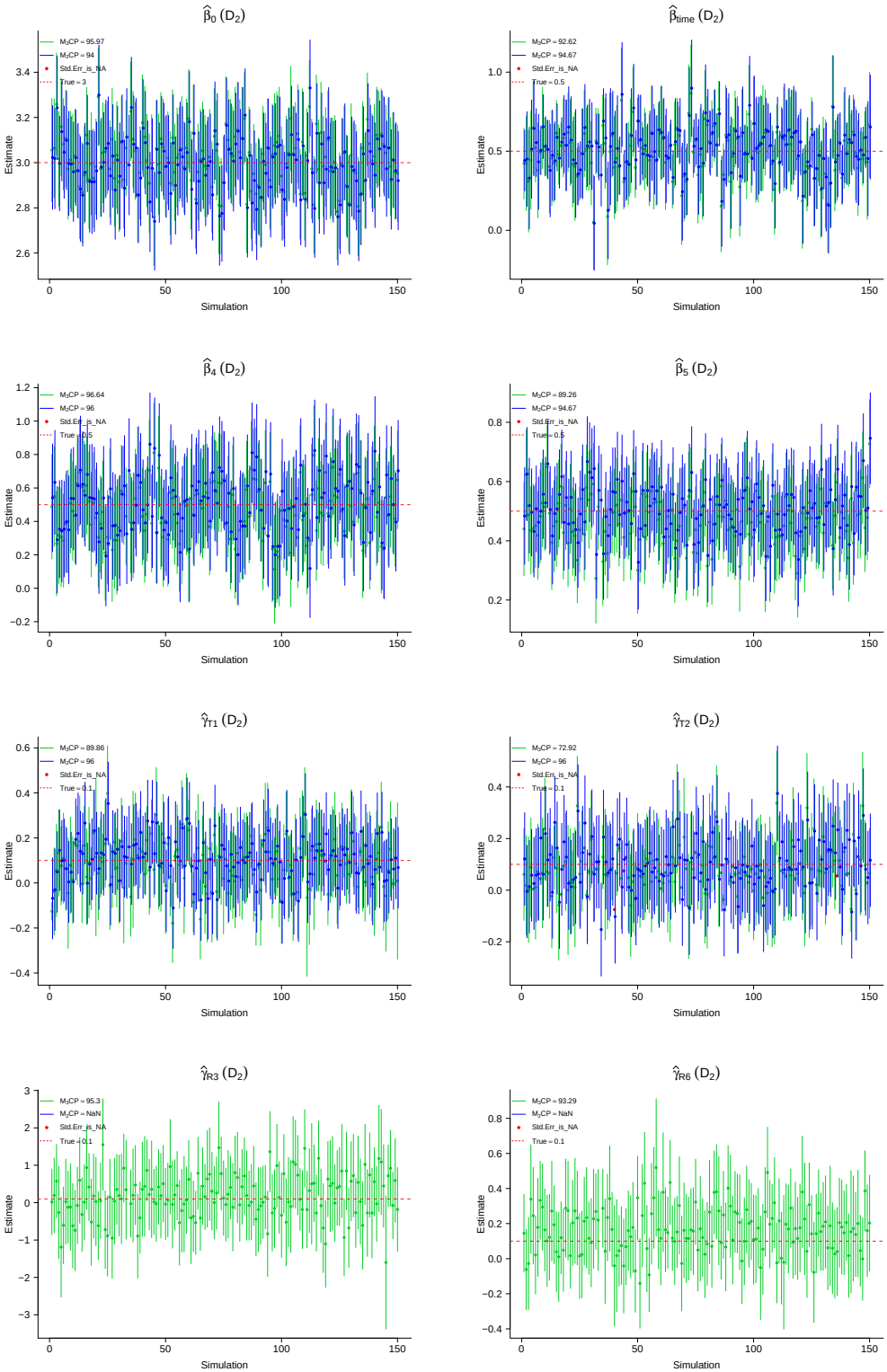


Figure B.30: Fixed effects regression coefficients of data produced with model M_2. Compare model fit M_3 and M_2 on D_2.

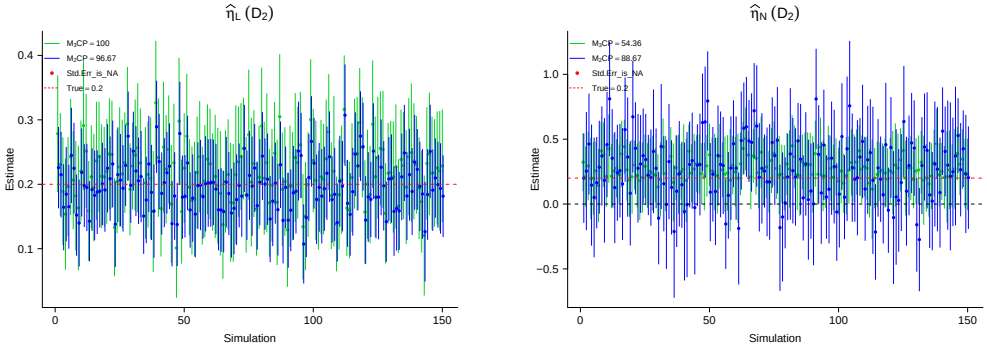


Figure B.31: Association parameters of data produced with model M_2. Compare model fit M_3 and M_2 on D_2.

References

ABRAMS, D., GOLDMAN, A., LAUNER, C., KORVICK, J., NEATON, J., CRANE, L., GRODESKY, M., WAKEFIELD, S., MUTH, K., KORNEGAY, S. *et al.* (1994). Comparative trial of didanosine and zalcitabine in patients with human immunodeficiency virus infection who are intolerant of or have failed zidovudine therapy. *New England Journal of Medicine*, **330**, 657–662. 115

AITCHISON, J. & DUNSMORE, I.R. (1975). *Statistical Prediction Analysis*. Cambridge University Press, Cambridge. 49

ALSEFRI, M., SUDELL, M., GARCÍA-FIÑANA, M. & KOLAMUNNAGE-DONA, R. (2020). Bayesian joint modelling of longitudinal and time to event data: a methodological review. *BMC Medical Research Methodology*, **20**, 1–17. 122

ANDERSEN, P.K. & GILL, R.D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics*, **10**, 1100–1120. 42

ANDERSEN, P.K., BORGAN, O., GILL, R.D. & KEIDING, N. (2012). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York. 42

ANDERSON, D.R. (2007). *Model Based Inference in the Life Sciences: A Primer on Evidence*. Springer-Verlag, New York. 165

BARKER, P. & HENDERSON, R. (2005). Small sample bias in the gamma frailty model for univariate survival. *Lifetime Data Analysis*, **11**, 265–284. 228

BARLOW, W.E. & PRENTICE, R.L. (1988). Residuals for relative risk regression. *Biometrika*, **75**, 65–74. 152

BENEDETTI, R. (2010). Scoring rules for forecast verification. *Monthly Weather Review*, **138**, 203–211. 187

BENNER, A., ZUCKNICK, M., HIELSCHER, T., ITTRICH, C. & MANSMANN, U. (2010). High-dimensional Cox models: the choice of penalty as part of the model building process. *Biometrical Journal*, **52**, 50–69. 164

BOLLEN, K.A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, **53**, 605–634. 4

BRIER, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3. 52, 54

BURNHAM, K.P. & ANDERSON, D.R. (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York. 165

BURNHAM, K.P. & ANDERSON, D.R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, **33**, 261–304. 165

CHANG, S.F. & LIN, P.L. (2015). Frail phenotype and mortality prediction: a systematic review and meta-analysis of prospective cohort studies. *International Journal of Nursing Studies*, **52**, 1362–1374. 126

CHEN, S.S., DONOHO, D.L. & SAUNDERS, M.A. (2001). Atomic decomposition by basis pursuit. *SIAM Review*, **43**, 129–159. 168

CHEN, Y. & WANG, Y. (2017). Variable selection for joint models of multivariate longitudinal measurements and event time data. *Statistics in Medicine*, **36**, 3820–3829. 164

CHENG, M.H. & CHANG, S.F. (2017). Frailty as a risk factor for falls among community dwelling people: Evidence from a meta-analysis. *Journal of Nursing Scholarship*, **49**, 529–536. 126

CHOODARI-OSKOOEI, B., ROYSTON, P. & PARMAR, M.K. (2012a). A simulation study of predictive ability measures in a survival model i: explained variation measures. *Statistics in Medicine*, **31**, 2627–2643. 53, 54

CHOODARI-OSKOOEI, B., ROYSTON, P. & PARMAR, M.K. (2012b). A simulation study of predictive ability measures in a survival model ii: explained randomness and predictive accuracy. *Statistics in Medicine*, **31**, 2644–2659. 53, 187

CLEGG, A., YOUNG, J., ILIFFE, S., RIKKERT, M.O. & ROCKWOOD, K. (2013). Frailty in elderly people. *The Lancet*, **381**, 752–762. 3, 124

CLEGG, A., BATES, C., YOUNG, J., RYAN, R., NICHOLS, L., ANN TEALE, E., MOHAMMED, M.A., PARRY, J. & MARSHALL, T. (2016). Development and validation of an electronic frailty index using routine primary care electronic health record data. *Age and Ageing*, **45**, 353–360. 5

COLE, S.R., CHU, H. & GREENLAND, S. (2014). Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *American Journal of Epidemiology*, **179**, 252–260. 62

COLLETT, D. (2015). *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC, Boca Raton, FL. 35, 36, 37, 42

COMMENGES, D. & JACQMIN-GADDA, H. (2015). *Dynamical Biostatistical Models*. Chapman and Hall/CRC, Boca Raton, FL. 19, 36, 90, 93, 122

COMMENGES, D., LIQUET, B. & PROUST-LIMA, C. (2012). Choice of prognostic estimators in joint models by estimating differences of expected conditional Kullback–Leibler risks. *Biometrics*, **68**, 380–387. 108

COOK, R.J. & LAWLESS, J. (2007). *The Statistical Analysis of Recurrent Events*. Springer-Verlag, New York. 43, 214

COX, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B (Methodological)*, **34**, 187–202. 27, 28

DAS, K., LI, R., HUANG, Z., GAI, J. & WU, R. (2012). A Bayesian framework for functional mapping through joint modeling of longitudinal and time-to-event data. *International Journal of Plant Genomics*, **2012**, Article ID 680634, 1–12. 96

DEMPSTER, A.P., LAIRD, N.M. & RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 1–38. 32, 96

DOBSON, A. & HENDERSON, R. (2003). Diagnostics for joint longitudinal and dropout time modeling. *Biometrics*, **59**, 741–751. 152, 153, 156, 161

DUCHATEAU, L. & JANSSEN, P. (2007). *The Frailty Model*. Springer-Verlag, New York. 31, 32

ENSRUD, K.E., EWING, S.K., TAYLOR, B.C., FINK, H.A., STONE, K.L., CAULEY, J.A., TRACY, J.K., HOCHBERG, M.C., RODONDI, N. & CAWTHON, P.M. (2007). Frailty and risk of falls, fracture, and mortality in older women: the study of osteoporotic fractures. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, **62**, 744–751. 126, 198

EVERETT, B. (2013). *An Introduction to Latent Variable Models*. Springer-Verlag, New York. 4

FALLER, J.W., PEREIRA, D.D.N., DE SOUZA, S., NAMPO, F.K., ORLANDI, F.D.S. & MATUMOTO, S. (2019). Instruments for the detection of frailty syndrome in older adults: A systematic review. *PLoS ONE*, **14**, 1–23. 5

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360. 164

FAN, J. & LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics*, **30**, 74–99. 164

FAWCETT, C. & THOMAS, D. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates. *Statistics in Medicine*, **15**, 1663–1685. 96

FAWCETT, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861–874. 54, 178

FITZMAURICE, G.M., LAIRD, N.M. & WARE, J.H. (2012). *Applied Longitudinal Analysis*. John Wiley & Sons, Hoboken, New Jersey. 10, 15, 17, 19

GERDS, T.A. & SCHUMACHER, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, **48**, 1029–1040. 57

GÖNEN, M. & HELLER, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, **92**, 965–970. 54

GRAF, E., SCHMOOR, C., SAUERBREI, W. & SCHUMACHER, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, **18**, 2529–2545. 54, 55, 57

GRAHAM, J.W. (2012). *Missing Data: Analysis and Design*. Springer-Verlag, New York. 21

HAN, D., SU, X., SUN, L., ZHANG, Z. & LIU, L. (2020). Variable selection in joint frailty models of recurrent and terminal events. *Biometrics*. 190

HANAGAL, D.D. (2011). *Modeling Survival Data using Frailty Models*. Springer-Verlag, New York. 30, 32

HANLEY, J.A. & MCNEIL, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, **143**, 29–36. 108

HAO, Q., ZHOU, L., DONG, B., YANG, M., DONG, B. & WEIL, Y. (2019). The role of frailty in predicting mortality and readmission in older adults in acute care wards: a prospective study. *Scientific Reports*, **9**, 1–8. 198

HARRELL, F.E., LEE, K.L. & MARK, D.B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361–387. 54

HARTHOLT, K.A., VAN BEECK, E.F. & VAN DER CAMMEN, T.J. (2018). Mortality from falls in Dutch adults 80 years and older, 2000–2016. *Journal of the American Medical Association*, **319**, 1380–1382. 127

HARTHOLT, K.A., LEE, R., BURNS, E.R. & VAN BEECK, E.F. (2019). Mortality from falls among US adults aged 75 years or older, 2000-2016. *Journal of the American Medical Association*, **321**, 2131–2133. 127

HARVILLE, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320–338. 18, 19

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J.H.J.H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, 2nd edn. 66, 165

HE, Z., TU, W., WANG, S., FU, H. & YU, Z. (2015). Simultaneous variable selection for joint models of longitudinal and survival outcomes. *Biometrics*, **71**, 178–187. 7, 164, 165

HECKMAN, J.J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, **5**, 475–492. 23

HEDEKER, D. & GIBBONS, R.D. (2006). *Longitudinal Data Analysis*. John Wiley & Sons, New York. 15

HEINZE, G., WALLISCH, C. & DUNKLER, D. (2018). Variable selection–a review and recommendations for the practicing statistician. *Biometrical Journal*, **60**, 431–449. 165, 190

HENDERSON, R. & OMAN, P. (1999). Effect of frailty on marginal regression estimates in survival analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**, 367–379. 228

HENDERSON, R., DIGGLE, P. & DOBSON, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1**, 465–480. 233

HICKEY, G.L., PHILIPSON, P., JORGENSEN, A. & KOLAMUNNAGE-DONA, R. (2016). Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues. *BMC Medical Research Methodology*, **16**, 1–15. 90, 98, 196

HICKEY, G.L., PHILIPSON, P., JORGENSEN, A. & KOLAMUNNAGE-DONA, R. (2018). Joint models of longitudinal and time-to-event data with more than one event time outcome: A review. *The International Journal of Biostatistics*, **14**. 90, 98, 196

HINKLEY, D.V. & COX, D. (1979). *Theoretical Statistics*. John Wiley & Sons, New York. 101

HOERL, A.E. & KENNARD, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67. 164

HOUGAARD, P. (1995). Frailty models for survival data. *Lifetime Data Analysis*, **1**, 255–273. 30

HSIEH, F., TSENG, Y.K. & WANG, J.L. (2006). Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics*, **62**, 1037–1043. 94

IBRAHIM, J.G., CHEN, M.H. & SINHA, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag, New York. 122

IBRAHIM, J.G., CHEN, M.H. & SINHA, D. (2004). Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. *Statistica Sinica*, 863–883. 96, 122

IBRAHIM, J.G., CHU, H. & CHEN, L.M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, **28**, 2796. 2, 89, 108, 122

JOLY, P., COMMENGES, D. & LETENNEUR, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia. *Biometrics*, 185–194. 45

KALBFLEISCH, J.D. & PRENTICE, R.L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons. 24, 25, 29, 36, 37

KAPLAN, E.L. & MEIER, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481. 58

KASS, R.E. & RAFTERY, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795. 87

KLEIN, J.P. & MOESCHBERGER, M.L. (2006). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, New York. 25, 29

KRÓL, A., FERRER, L., PIGNON, J.P., PROUST-LIMA, C., DUCREUX, M., BOUCHÉ, O., MICHIELS, S. & RONDEAU, V. (2016). Joint model for left-censored longitudinal data, recurrent events and terminal event: Predictive abilities of tumor burden for cancer evolution with application to the FFCD 2000–05 trial. *Biometrics*, **72**, 907–916. 122, 148

KRÓL, A., MAUGUEN, A., MAZROUI, Y., LAURENT, A., MICHIELS, S. & RONDEAU, V. (2017). Tutorial in joint modeling and prediction: A statistical software for correlated longitudinal outcomes, recurrent events and a terminal event. *Journal of Statistical Software*, **81**. 7

LAIRD, N.M. & WARE, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, 963–974. 13, 16, 19

LAVRAKAS, P.J., ed. (2008). *Encyclopedia of Survey Research Methods*, vol. 2. Sage Publications. 4

LAWLESS, J.F. (2011). *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, Hoboken, New Jersey. 37

LI, K. & LUO, S. (2017). Functional joint model for longitudinal and time-to-event data: an application to alzheimer's disease. *Statistics in Medicine*, **36**, 3560–3572. 98

LITTLE, R.J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125–134. 23

LITTLE, R.J. & RUBIN, D.B. (2019). *Statistical analysis with missing data*. John Wiley & Sons. 19, 22

LIU, F. & LI, Q. (2016). A Bayesian model for joint analysis of multivariate repeated measures and time to event data in crossover trials. *Statistical Methods in Medical Research*, **25**, 2180–2192. 122

LIU, L. & HUANG, X. (2009). Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **58**, 65–81. 122

LIU, L., HUANG, X. & O'QUIGLEY, J. (2008). Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics*, **64**, 950–958. 116, 122

MANZOLI, L., VILLARI, P., PIRONE, G.M. & BOCCIA, A. (2007). Marital status and mortality in the elderly: a systematic review and meta-analysis. *Social Science & Medicine*, **64**, 77–94. 136

MARQUARDT, D.W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, **11**, 431–441. 45, 96, 122, 147

MASUD, T. & MORRIS, R.O. (2001). Epidemiology of falls. *Age and Ageing*, **30**, 3–7. 6, 126, 156, 235

MOLENBERGHS, G. & KENWARD, M. (2007). *Missing Data in Clinical Studies*. John Wiley & Sons, New York. 22, 23

MOLENBERGHS, G. & VERBEKE, G. (2000). *Linear mixed models for longitudinal data*. Springer. 10, 15, 150

MONACO, J.V., GORFINE, M. & HSU, L. (2018). General semiparametric shared frailty model: Estimation and simulation with frailtysurv. *Journal of Statistical Software*, **86**. 34, 219

MUNDA, M., ROTOLO, F., LEGRAND, C. *et al.* (2012). parfm: Parametric frailty models in r. *Journal of Statistical Software*, **51**, 1–20. 32, 45

NOWAK, A. & HUBBARD, R.E. (2009). Falls and frailty: lessons from complex systems. *Journal of the Royal Society of Medicine*, **102**, 98–102. 124, 126

OBEL, N. (2012). Cd4 cell count and the risk of aids or death in hiv-infected adults on combination antiretroviral therapy with a suppressed viral load: a longitudinal cohort study from cohere. *PLOS Medicine*, **9**. 115

ORMEROD, J.T. & WAND, M.P. (2010). Explaining variational approximations. *The American Statistician*, **64**, 140–153. 239

OZGA, A.K., KIESER, M. & RAUCH, G. (2018). A systematic comparison of recurrent event models for application to composite endpoints. *BMC Medical Research Methodology*, **18**, 1–12. 43

PADRÓN-MONEDERO, A., DAMIÁN, J., MARTIN, M.P. & FERNÁNDEZ-CUENCA, R. (2017). Mortality trends for accidental falls in older people in Spain, 2000-2015. *BMC Geriatrics*, **17**, 276. 127

PATTERSON, H.D. & THOMPSON, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554. 18

PEARL, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge. 70, 71

PEARL, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics*, **6**, 1–62. 86

PEARL, J., GLYMOUR, M. & JEWELL, N.P. (2016). *Causal Inference in Statistics: A Primer*. John Wiley & Sons, Chichester. 68, 83

PENCINA, M.J. & D'AGOSTINO, R.B. (2004). Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, **23**, 2109–2123. 53, 54

PENG, H. & LU, Y. (2012). Model selection in linear mixed effect models. *Journal of Multivariate Analysis*, **109**, 109 – 129. 164

PENNY, W.D. (2012). Comparing dynamic causal models using aic, bic and free energy. *Neuroimage*, **59**, 319–330. 87

PENNY, W.D., STEPHAN, K.E., MECHELLI, A. & FRISTON, K.J. (2004). Comparing dynamic causal models. *Neuroimage*, **22**, 1157–1172. 87

PINHEIRO, J. & BATES, D. (2006). *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag, New York. 15

PRENTICE, R.L., WILLIAMS, B.J. & PETERSON, A.V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, **68**, 373–379. 42

PROUST-LIMA, C. & TAYLOR, J.M. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach. *Biostatistics*, **10**, 535–549. 122

R. BROWN, E. & G. IBRAHIM, J. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, **59**, 221–228. 96

RABE-HESKETH, S. & SKRONDAL, A. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC, Boca Raton, FL. 4

RAHMAN, M.S., AMBLER, G., CHOODARI-OSKOOEI, B. & OMAR, R.Z. (2017). Review and evaluation of performance measures for survival prediction models in external validation settings. *BMC Medical Research Methodology*, **17**, 60. 53, 54

RAWLINGS, J.O., PANTULA, S.G. & DICKEY, D.A. (2001). *Applied Regression Analysis: A Research Tool*. Springer-Verlag, New York. 163

RIZOPOULOS, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, **67**, 819–829. 7, 51, 101, 105

RIZOPOULOS, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Chapman and Hall/CRC, New York. 10, 22, 37, 91, 94, 96, 98, 99, 100, 108, 122, 220

RIZOPOULOS, D. (2014). The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *Journal of Statistical Software*, **72**. 96

RIZOPOULOS, D. & GHOSH, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*, **30**, 1366–1380. 122

ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical Modelling*, **7**, 1393–1512. 86

ROLFSON, D.B., MAJUMDAR, S.R., TSUYUKI, R.T., TAHIR, A. & ROCKWOOD, K. (2006). Validity and reliability of the Edmonton Frail Scale. *Age and Ageing*, **35**, 526–529. 5

RONDEAU, V., COMMENGES, D. & JOLY, P. (2003). Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime Data Analysis*, **9**, 139–153. 31, 32, 45, 96, 147

RONDEAU, V., MATHOULIN-PELISSIER, S., JACQMIN-GADDA, H., BROUSTE, V. & SOUBEYRAN, P. (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics*, **8**, 708–721. 96

RONDEAU, V., MAZROUI, Y. & GONZALEZ, J.R. (2012). frailtypack: An R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software*, **47**, 1–28. 32, 96, 122, 148

ROYSTON, P. & SAUERBREI, W. (2004). A new measure of prognostic separation in survival data. *Statistics in Medicine*, **23**, 723–748. 54

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592. 21

SAMPER-TERNENT, R., KARMARKAR, A., GRAHAM, J., REISTETTER, T. & OTTENBACHER, K. (2012). Frailty as a predictor of falls in older Mexican Americans. *Journal of Aging and Health*, **24**, 641–653. 126

SEBER, G.A. & LEE, A.J. (2012). *Linear Regression Analysis*. John Wiley & Sons, New York. 61, 163

SHMUELI, G. *et al.* (2010). To explain or to predict? *Statistical Science*, **25**, 289–310. 7, 160, 236

SOURDET, S., ROUGÉ-BUGAT, M., VELLAS, B. & FORETTE, F. (2012). Frailty and aging. *The Journal of Nutrition, Health & Aging*, **16**, 283–284. 126

STEYERBERG, E.W., VICKERS, A.J., COOK, N.R., GERDS, T., GONEN, M., OBUCHOWSKI, N., PENCINA, M.J. & KATTAN, M.W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, **21**, 128. 52, 54, 55

SU, X., WIJAYASINGHE, C.S., FAN, J. & ZHANG, Y. (2016). Sparse estimation of Cox proportional hazards models via approximated information criteria. *Biometrics*, **72**, 751–759. 190

THERNEAU, T.M., GRAMBSCH, P.M. & FLEMING, T.R. (1990). Martingale-based residuals for survival models. *Biometrika*, **77**, 147–160. 152

THOMADAKIS, C., MELIGKOTSIDOU, L., PANTAZIS, N. & TOULOUMI, G. (2019). Longitudinal and time-to-drop-out joint models can lead to seriously biased estimates when the drop-out mechanism is at random. *Biometrics*, **75**, 58–68. 100

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 267–288. 164, 168

TOM, S.E., ADACHI, J.D., ANDERSON JR, F.A., BOONEN, S., CHAPURLAT, R.D., COMPSTON, J.E., COOPER, C., GEHLBACH, S.H., GREENSPAN, S.L., HOOVEN, F.H. *et al.* (2013). Frailty and fracture, disability, and falls: a multiple country study from the global longitudinal study of osteoporosis in women. *Journal of the American Geriatrics Society*, **61**, 327–334. 126

TREVISAN, C., VERONESE, N., MAGGI, S., BAGGIO, G., DE RUI, M., BOLZETTA, F., ZAMBON, S., SARTORI, L., PERISSINOTTO, E., CREPALDI, G. *et al.* (2016). Marital status and frailty in older people: gender differences in the progetto veneto anziani longitudinal study. *Journal of Women's Health*, **25**, 630–637. 136

TREVISAN, C., GRANDE, G., VETRANO, D.L., MAGGI, S., SERGI, G., WELMER, A.K. & RIZZUTO, D. (2020). Gender differences in the relationship between marital status and the development of frailty: A Swedish longitudinal population-based study. *Journal of Women's Health*. 136

VAIDA, F. & XU, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine*, **19**, 3309–3324. 32

VAN HOUWELINGEN, H. & PUTTER, H. (2011). *Dynamic Prediction in Clinical Survival Analysis*. Chapman and Hall/CRC, Boca Raton, FL. 55, 57

VAN HOUWELINGEN, H.C. (2000). Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine*, **19**, 3401–3415. 54

VAN HOUWELINGEN, H.C. (2014). From model building to validation and back: a plea for robustness. *Statistics in Medicine*, **33**, 5223–5238. 233

VANDERWEELE, T.J. (2011). Causal mediation analysis with survival data. *Epidemiology (Cambridge, Mass.)*, **22**, 582. 204

VERBEKE, G. (1997). Linear mixed models for longitudinal data. In *Linear Mixed Models in Practice. A SAS-Oriented Approach*, 63–153, Springer-Verlag, New York. 10

WAND, M.P. (2017). Fast approximate inference for arbitrarily large semiparametric regression models via message passing. *Journal of the American Statistical Association*, **112**, 137–168. 239

WEI, L.J., LIN, D.Y. & WEISSFELD, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, **84**, 1065–1073. 42

WEST, B.T., WELCH, K.B. & GALECKI, A.T. (2014). *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman and Hall/CRC, Boca Raton FL. 11

WIENKE, A. (2010). *Frailty Models in Survival Analysis*. Chapman and Hall/CRC, Boca Raton, FL. 31, 32

WULFSOHN, M.S. & TSIATIS, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339. 96

ZHANG, H.H. & LU, W. (2007). Adaptive LASSO for Cox's proportional hazards model. *Biometrika*, **94**, 691–703. 164

ZHENG, Y. & HEAGERTY, P.J. (2007). Prospective accuracy for longitudinal markers. *Biometrics*, **63**, 332–341. 7

ZHOU, X.H., MCCLISH, D.K. & OBUCHOWSKI, N.A. (2009). *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, New York. 179

ZIEGEL, E. (1987). *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, Cambridge. 34, 174, 219

ZOU, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429. 164