Joint Modelling of Longitudinal and Survival Data: Tools to Evaluate Exposures and Predict Outcome Across the Lifespan

Eleni-Rosalina Andrinopoulou and Rhonda Szczesniak

Erasmus University Medical Center and Cincinnati Children's Hospital

material based on work by Dimitris Rizopoulos

ACE 2018

September 23, 2018, Cincinnati

Contents

3 Relative Risk Models 40

4 The Basic Joint Model 62

5 Extensions of Joint Models 91

6 Dynamic Predictions   130

7 Closing   140

What is this Course About

- Often in follow-up studies different types of outcomes are collected

- Explicit outcomes

  ▷ multiple longitudinal responses (e.g., biomarkers, blood values)

  ▷ time-to-event(s) of particular interest (e.g., death, treatment)

What is this Course About (cont'd)

- Methods for the separate analysis of such outcomes are well established in the literature

- **Longitudinal data:**

  ▷ mixed effects models, GEE, . . .

- **Survival data:**

  ▷ Cox model, accelerated failure time models, . . .

What is this Course About (cont'd)

Purpose of this workshop is to present the state of the art in

**Joint Modeling Framework
for Longitudinal and Survival Data**

Learning Objectives

- **Goals:** After this workshop participants will be able to

  ▷ identify settings in which a joint modeling approach is required,

  ▷ construct and fit an appropriate joint model, and

  ▷ correctly interpret the obtained results

- The workshop will be explanatory rather than mathematically rigorous

  ▷ emphasis is given on applications

Agenda

- ## **Part I:** Introduction

  ▷ Data sets that we will use throughout the course

  ▷ Categorization of possible research questions

- ## **Part II:** (brief) Review of Linear Mixed Models

  ▷ Features of repeated measurements data

  ▷ Linear mixed models

- **Part III:** (brief) Review of Relative Risk Models

  ▷ Features of survival data

  ▷ Relative risk models

  ▷ Time-dependent covariates

- **Part IV:** The Basic Joint Model

  ▷ Definition

  ▷ Estimation & Inference

Agenda (cont'd)

- **<u>Part V:</u>** Extensions of the Basic Joint Model

    ▷ Parameterizations

    ▷ Variable selection

    ▷ Time-varying effects

- **<u>Part VI:</u>** Dynamic Predictions

    ▷ Individualized predictions for the survival

Structure of the Course & Material

- Lectures & short software practicals using R package **JMbayes**

- Material (also available in

https://github.com/ERandrinopoulou/JointModelsWorkshopCincinnati2018):

  ▷ Course Notes (slides)

  ▷ R code in soft format (app)

  ▷ Practicals with solutions

References

- **Joint modeling** sources

  ▷ Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data, with Applications in R*. Boca Raton: Chapman & Hall/CRC.

  ▷ Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2009). *Longitudinal Data Analysis*. Handbooks of Modern Statistical Methods. Boca Raton: Chapman & Hall/CRC, Chapter 15.

  ▷ Wu, L. (2009). *Mixed Effects Models for Complex Data*. Boca Raton: Chapman & Hall/CRC, Chapter 8.

  ▷ Ibrahim, J., Chen, M.-H. and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer-Verlag, Chapter 7.

References (cont'd)

- Useful material for package **JMbayes**

  ▷ a paper describing the current capabilities of the package is available on JSS
  http://dx.doi.org/10.18637/jss.v072.i07

- Standard texts in **longitudinal** data analysis

    ▷ Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

    ▷ Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer-Verlag.

    ▷ Fitzmaurice, G., Laird, N., and Ware, J. (2004). *Applied Longitudinal Analysis*. Hoboken: Wiley.

    ▷ Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of Longitudinal Data*, 2nd edition. New York: Oxford University Press.

References (cont'd)

- Standard texts in **survival** analysis

  ▷ Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data, 2nd Ed.*. New York: Wiley.

  ▷ Therneau, T. and Grambsch, P. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.

  ▷ Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman & Hall.

  ▷ Andersen, P., Borgan, O., Gill, R. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

  ▷ Klein, J. and Moeschberger, M. (2003). *Survival Analysis - Techniques for Censored and Truncated Data*. New York: Springer-Verlag.

Chapter 1

Introduction

1.1 Motivating Studies

- AIDS: 467 HIV infected patients who had failed or were intolerant to zidovudine therapy (AZT) (Abrams et al., NEJM, 1994)

- The aim of this study was to compare the efficacy and safety of two alternative antiretroviral drugs, didanosine (ddI) and zalcitabine (ddC)

  ▷ Randomized treatment: 230 patients ddI and 237 ddC

- Outcomes of interest:

  ▷ **time-to-death**

  ▷ **CD4 cell count** at baseline, 2, 6, 12 and 18 months
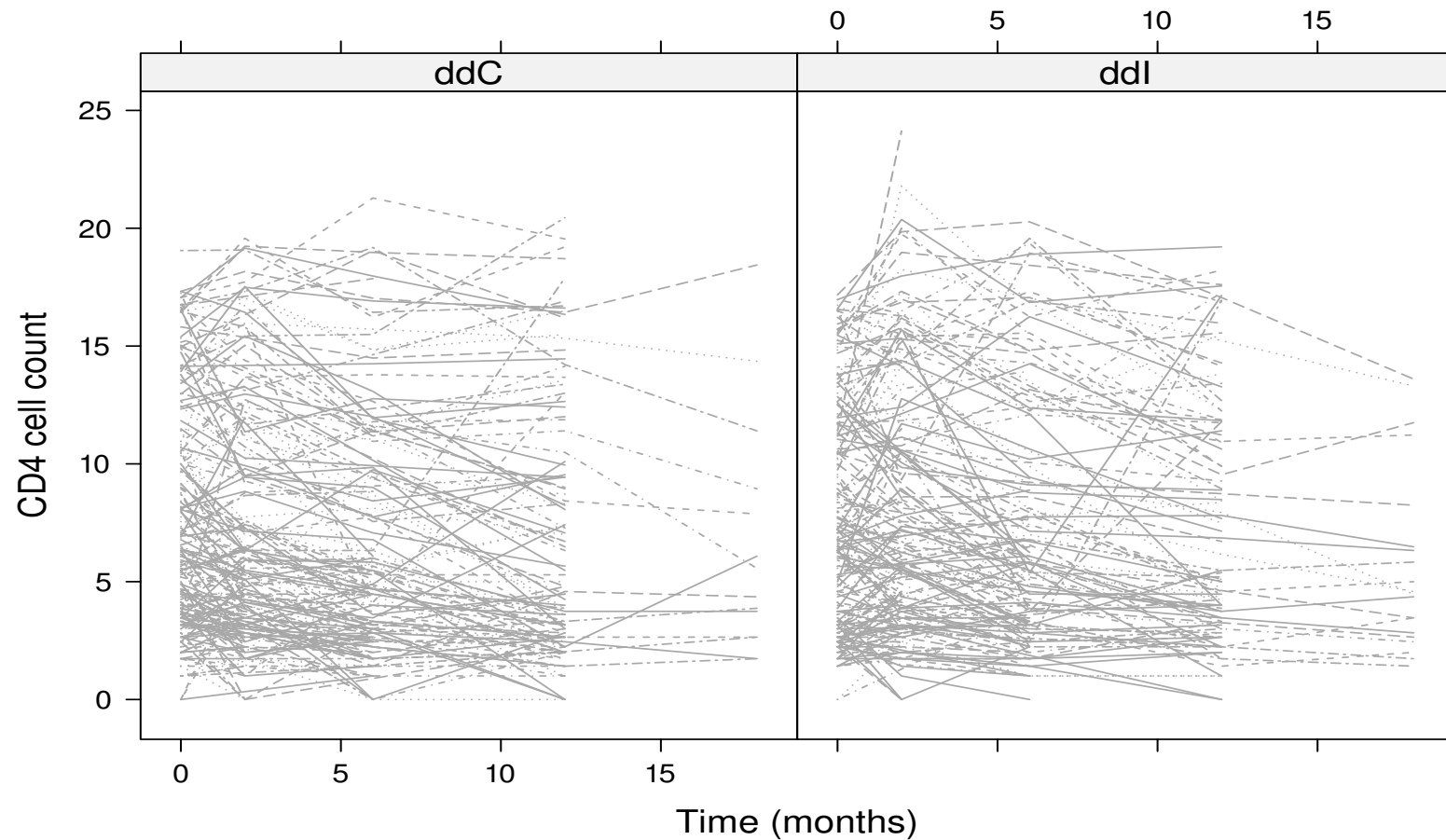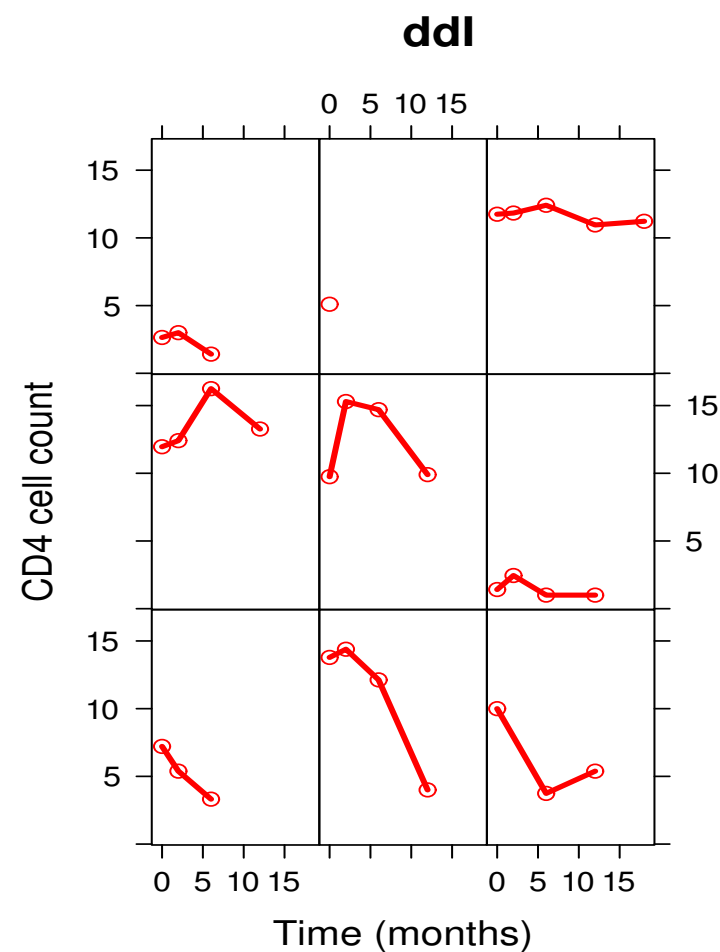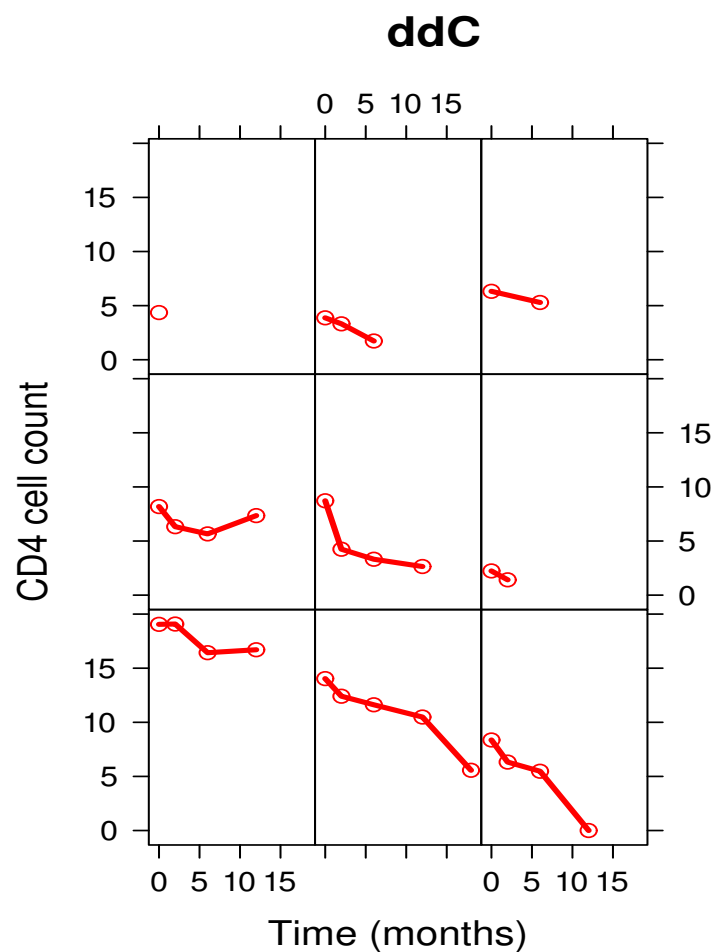
Kaplan–Meier Estimate

- **Research Questions:**

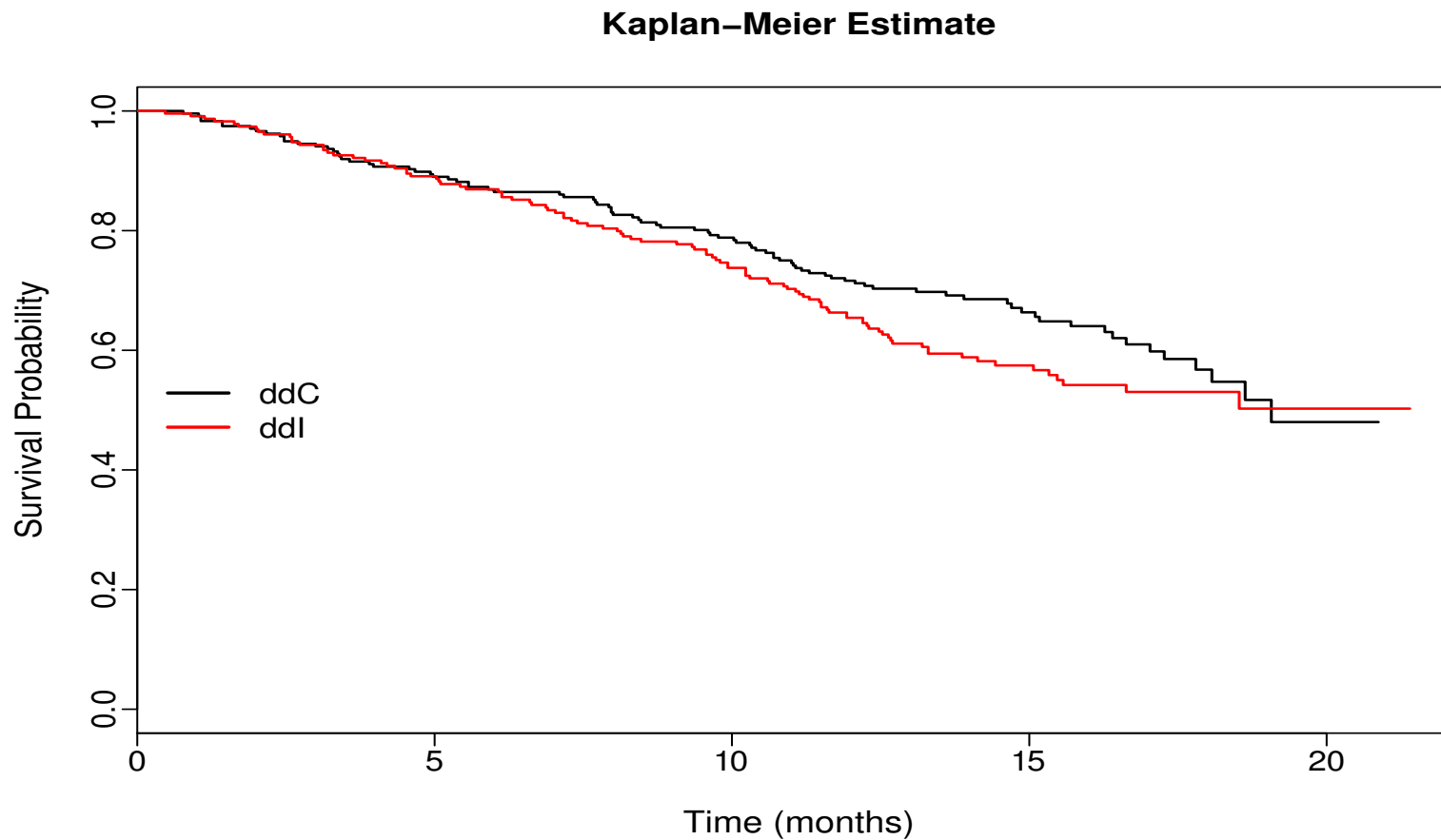   ▷ How strong is the association between CD4 cell count and the risk of death?

    ▷ Is CD4 cell count a good biomarker?
       * if treatment improves CD4 cell count, does it also improve survival?

1.1 Motivating Studies (cont'd)

- $\boxed{\text{PBC:}}$ 312 patients with Primary Biliary Cirrhosis which is a chronic, fatal but rare liver disease

- Data collected by Mayo Clinic from 1974 to 1984 (Murtaugh et al., Hepatology, 1994)
  - ▷ Randomized treatment: 158 patients received D-penicillamine and 154 placebo

- Outcomes of interest:
  - ▷ **time-to-death** and/or **time-to-liver transplantation**
  - ▷ **longitudinal serum bilirubin levels**
  - ▷ **longitudinal serum cholesterol levels**
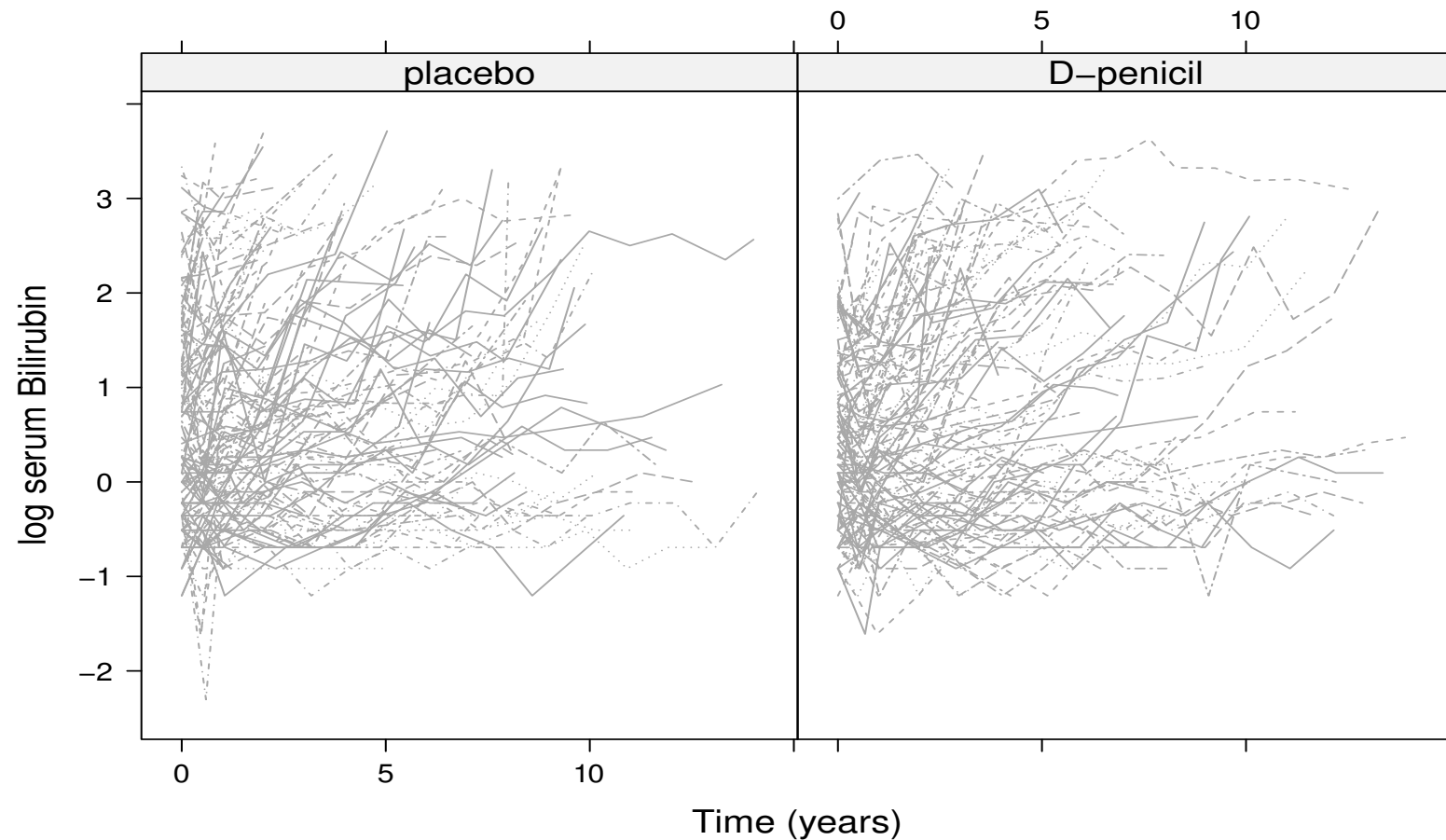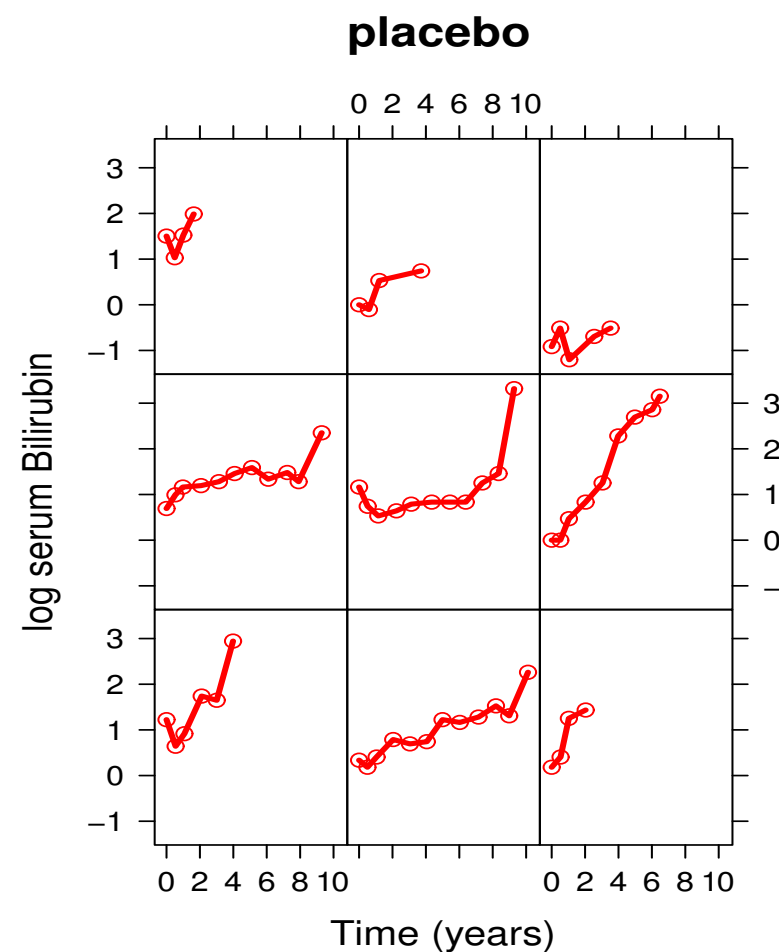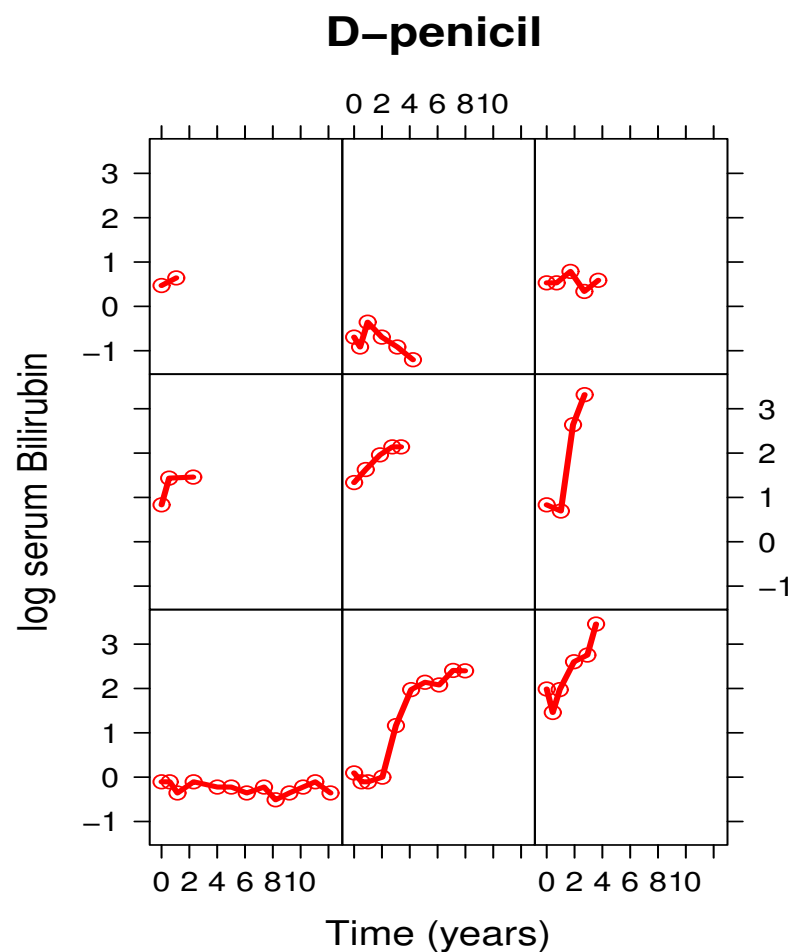
Kaplan–Meier Estimate
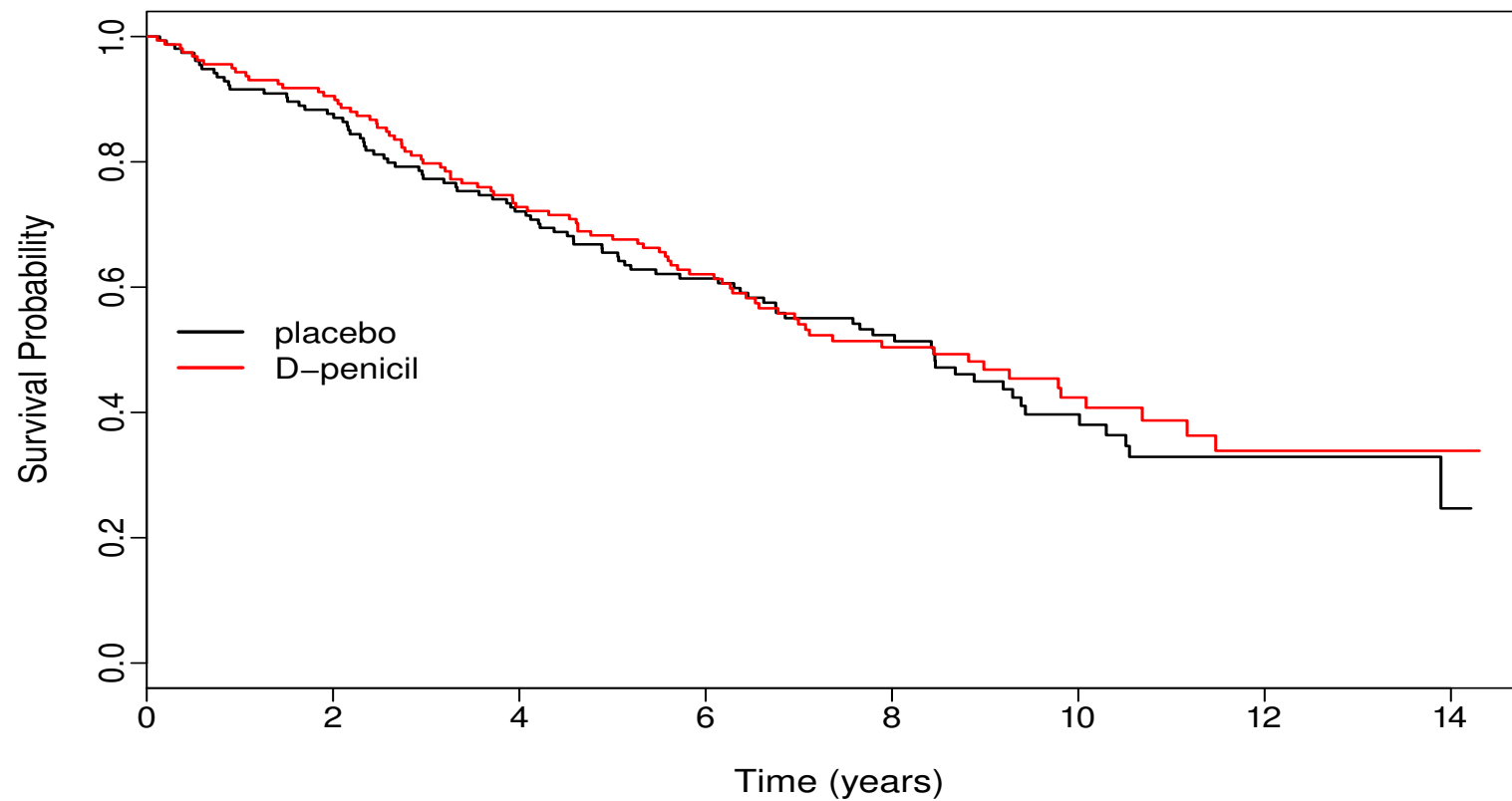
- Research Questions:

  ▷ How strong is the association between bilirubin and cholesterol with the risk of death?

  ▷ How could we utilize the observed serum bilirubin levels to provide predictions of survival probabilities?

  ▷ Can bilirubin and cholesterol discriminate between patients of low and high risk?

1.2 Research Questions

- Depending on the questions of interest, different types of statistical analysis are required

- We will distinguish between two general types of analysis

  ▷ separate analysis per outcome

  ▷ joint analysis of outcomes

- Focus on each outcome separately

  ▷ does treatment affect survival?

  ▷ are the average longitudinal evolutions different between the two treatments or between males and females?

- Focus on multiple outcomes

  ▷ How strong is the association between the longitudinal evolution of the biomarkers and the hazard rate of death?

  ▷ Association structures:
    * which feature of the biomarker(s) is associated with the hazard of death?
    * how marker-specific evolutions are related to each other

  ▷ Prediction: can we improve prediction for the time to death by considering all available information simultaneously?

1.3 Recent Developments

- Up to now emphasis has been

  ▷ *restricted* or *coerced* to separate analysis per outcome

  ▷ or given to naive types of joint analysis (e.g., last observation carried forward, mean or slope of the repeated covariate, . . .)

- Main reasons

  ▷ lack of appropriate statistical methodology

  ▷ lack of efficient computational approaches & software

- However, recently there has been an explosion in the statistics and biostatistics literature of joint modeling approaches

- Many different approaches have been proposed that

    ▷ can handle different types of outcomes

    ▷ can be utilized in pragmatic computing time

    ▷ can be rather flexible

    ▷ **most importantly:** can answer the questions of interest

Chapter 2

Linear Mixed-Effects Models

2.1 Features of Longitudinal Data

- Repeated evaluations of the same outcome in each subject over time

    ▷ CD4 cell count in HIV-infected patients

    ▷ serum bilirubin and cholesterol in PBC patients

- Longitudinal studies allow to investigate

    1. how treatment means differ at specific time points, e.g., at the end of the study (*cross-sectional effect*)
    2. how treatment means or differences between means of treatments change over time (*longitudinal effect*)

> **Measurements on the same subject are expected to be (positively) correlated**

- This implies that standard statistical tools, such as the t-test and simple linear regression that assume independent observations, are not optimal for longitudinal data analysis.

- The direct approach to model correlated data $\Rightarrow$ *linear regression*

$$y_i(t) = x_i^\top(t)\beta + \varepsilon_i(t), \quad \varepsilon_i(t) \sim \mathcal{N}(0, V_i),$$

where

- $\triangleright$ $y_i(t)$ the vector of responses for the ith subject

- $\triangleright$ $x_i^\top(t)$ design matrix describing structural component

- $\triangleright$ V_i covariance matrix describing the correlation structure

- There are several options for modeling V_i, e.g., compound symmetry, autoregressive process, exponential spatial correlation, Gaussian spatial correlation, ...

- **Alternative intuitive approach:** Each subject in the population has her own subject-specific mean response profile over time
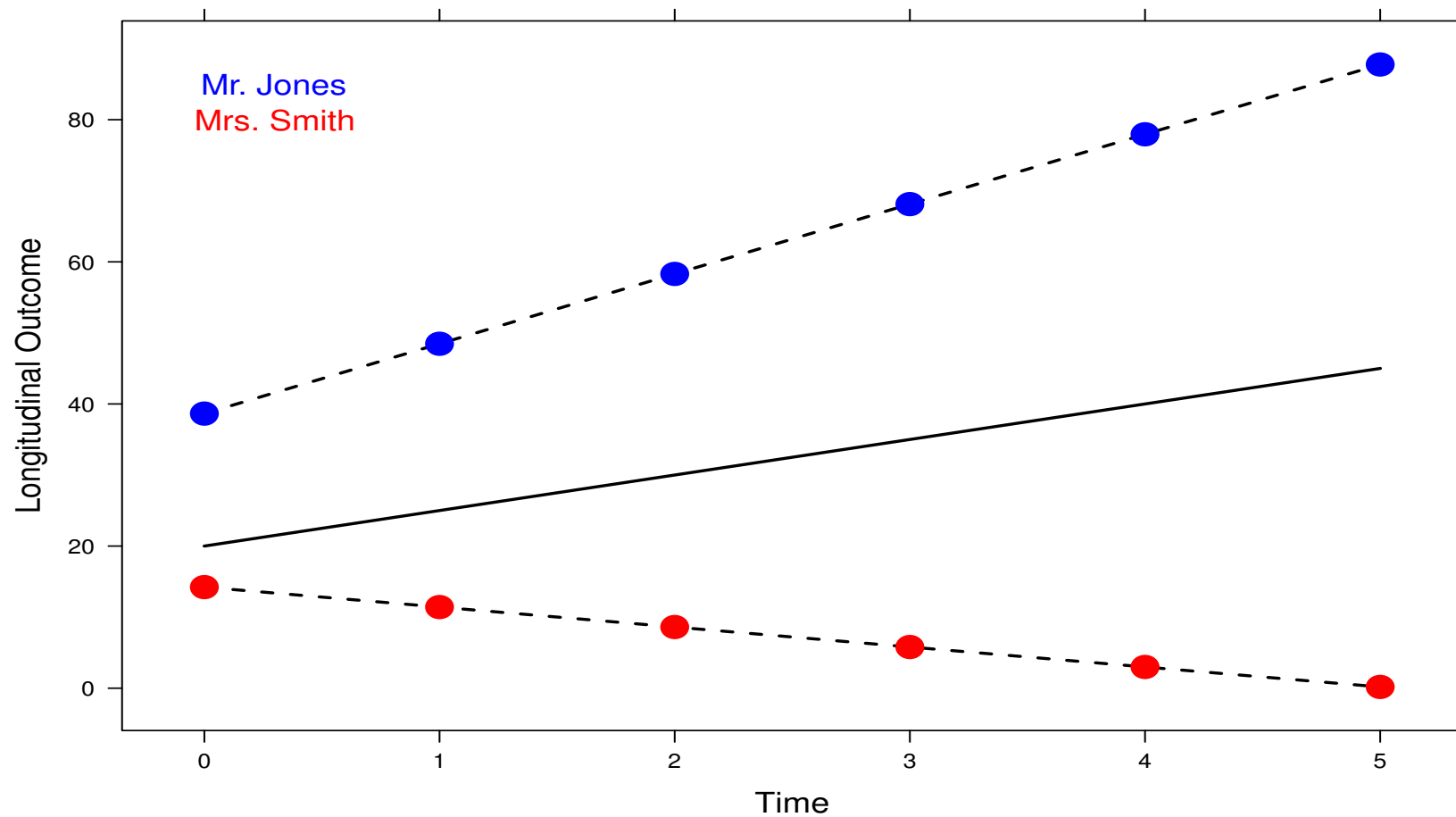
- The evolution of each subject over time can be described by a linear model

$$y_{ij} = \tilde{\beta}_{i0} + \tilde{\beta}_{i1} t_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

  where

  $\triangleright$ y_{ij} the jth response of the ith subject

  $\triangleright$ $\tilde{\beta}_{i0}$ is the intercept and $\tilde{\beta}_{i1}$ the slope for subject i

- We can reformulate the model as

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

  where

  $\triangleright$ βs are known as the *fixed effects*

  $\triangleright$ b_is are known as the *random effects*

- In accordance for the random effects we assume

$$b_i = \begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} \sim \mathcal{N}(0, D)$$
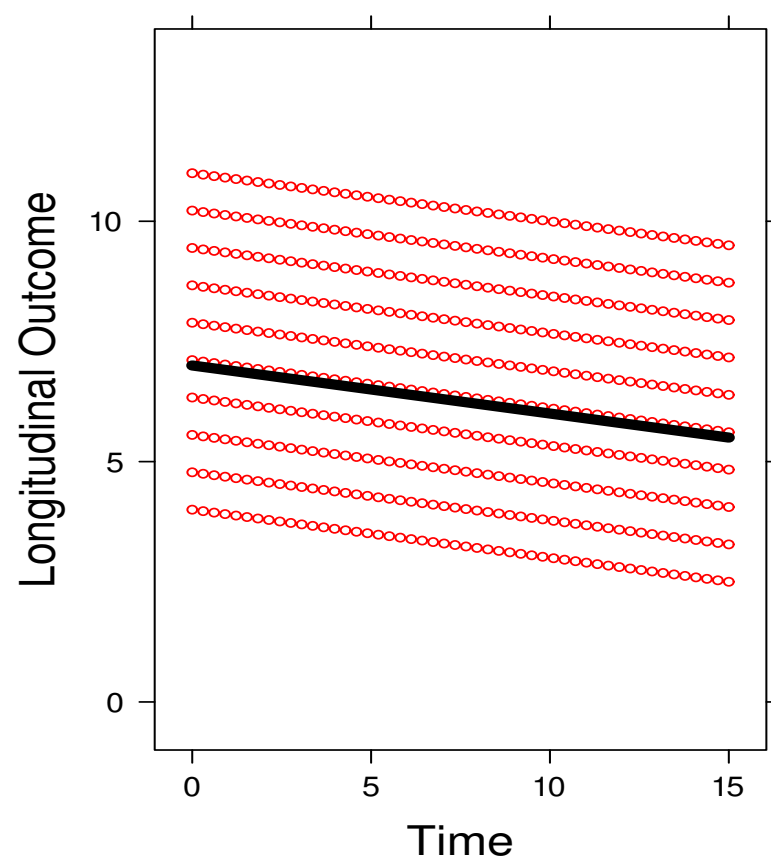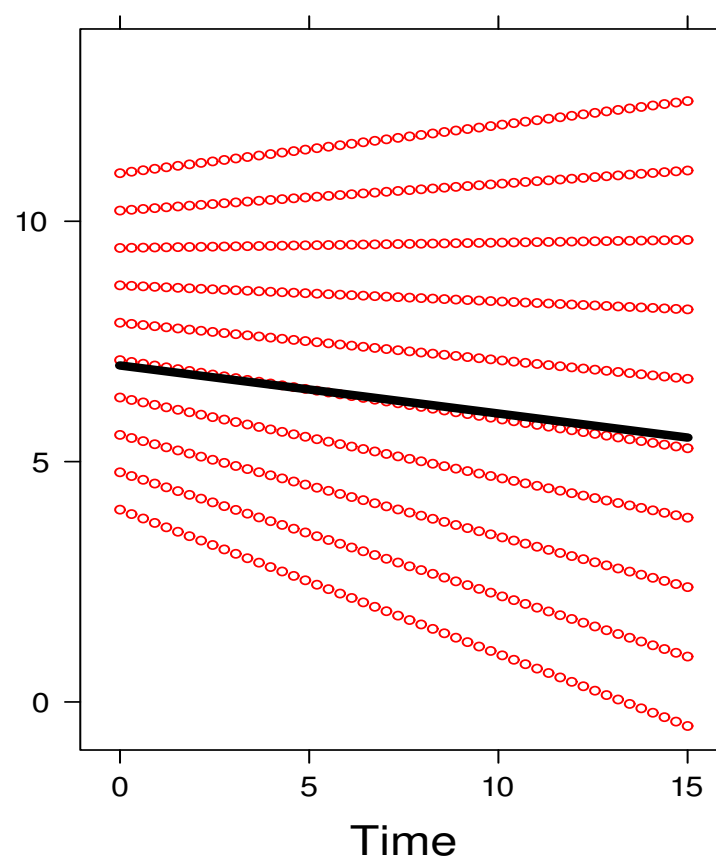
2.2 The Linear Mixed Model (cont'd)

- Put in a general form

$$
\begin{cases}
y_i(t) = x_i^\top(t)\beta + z_i^\top(t)b_i + \varepsilon_i(t), \\
\\
b_i \sim \mathcal{N}(0, D), \quad \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2 \mathsf{I}_{n_i}),
\end{cases}
$$

with

$\triangleright$ $x_i^\top(t)$ design matrix for the fixed effects β

$\triangleright$ $z_i^\top(t)$ design matrix for the random effects b_i

$\triangleright$ $b_i \perp\!\!\!\perp \varepsilon_i$

- Interpretation:

  ▷ β_j denotes the change in the average y_i when x_j is increased by one unit

  ▷ b_i are interpreted in terms of how a subset of the regression parameters for the ith subject deviates from those in the population

- Advantageous feature: population + subject-specific predictions

  ▷ β describes mean response changes in the population

  ▷ $\beta + b_i$ describes individual response trajectories

- Estimation $\rightarrow$ maximum likelihood (MLE)

- The log-likelihood of a linear mixed model takes the form

$$\ell(\theta) = \sum_{i=1}^{n} log \int p(y_i \mid b_i; \theta_y) p(b_i; \theta_b) db_i,$$

where $p(.)$ the density function

2.2 The Linear Mixed Model (cont'd)

- Example: We fit a linear mixed model for the AIDS dataset assuming

  ▷ different average longitudinal evolutions per treatment group (fixed part)

  ▷ random intercepts & random slopes (random part)

$$\begin{cases} y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 \{\texttt{ddI}_i \times t_{ij}\} + b_{i0} + b_{i1} t_{ij} + \varepsilon_{ij}, \\ \\ b_i \sim \mathcal{N}(0, D), \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

- Note: We did not include a main effect for treatment due to randomization

	Value	Std.Err.	t-value	p-value
β_0	7.189	0.222	32.359	< 0.001
β_1	−0.163	0.021	−7.855	< 0.001
β_2	0.028	0.030	0.952	0.342

- No evidence of differences in the average longitudinal evolutions between the two treatments

- Mixed-effetcs models consist of two parts, namely

  ▷ **fixed effects** that describe how specific covariates influence the average longitudinal evolutions

  ▷ **random effects** that describe how specific regression coefficients deviate from the overall mean described by the fixed effects

     * the random effects also model the correlations in the repeated measurements

- Interest can either be

  ▷ on the fixed-effects part alone (e.g., does treatment influence the average evolutions) or

  ▷ on both parts (e.g., to obtain subject specific predictions)

- The general model building strategy is:

  1. Put all the covariates of interest in the fixed-effects part, considering possible nonlinear terms and/or interactions between them - **do NOT** remove the ones that are not significant

  2. Then select an appropriate random-effects structure that adequately describes the correlations in the repeated measurements

     * typically we start from random intercepts and include each time an additional random effect term to see if we improve the fit (i.e., random slopes, quadratic random slopes, etc.)

     * you should be a bit anti-conservative, i.e., do not favor a simpler covariance matrix if the p-value is just non-significant

3. Finally, return to the mean part and exclude non significant covariates

  * first start by testing the nonlinear & interaction terms

2.3 Model Building (cont'd)

- For **nested** models the preferable test for selecting a model is the likelihood ratio test

  * nested means that all terms of a smaller model occur in a larger model

- For **non-nested** models we choose the model that has the lowest AIC/BIC value

2.4 Mixed-Effects Models in R

R> There are two primary packages in R for mixed models analysis:

▷ Package **nlme**
  * fits linear & nonlinear mixed effects models, and marginal models for normal data
  * allows for both random effects & correlated error terms
  * several options for covariances matrices and variance functions

▷ Package **lme4**
  * fits linear, nonlinear & generalized mixed effects models
  * uses only random effects
  * allows for nested and crossed random-effects designs

2.4 Mixed-Effects Models in R (cont'd)

R> We will only use package **nlme** because package **JMbayes** accepts as an argument a linear mixed model fitted by **nlme**

R> The basic function to fit linear mixed models is `lme()` and has three basic arguments

   ▷ `fixed`: a formula specifying the response vector and the fixed-effects structure

   ▷ `random`: a formula specifying the random-effects structure

   ▷ `data`: a data frame containing all the variables

R> The data frame that contains all variables should be in the *long format*

Subject	y	time	gender	age
1	5.1	0.0	male	45
1	6.3	1.1	male	45
2	5.9	0.1	female	38
2	6.9	0.9	female	38
2	7.1	1.2	female	38
2	7.3	1.5	female	38
⋮	⋮	⋮	⋮	⋮

R> Using formulas in R

> ▷ CD4 = Time + Gender
>   ⇒ `cd4 ~ time + gender`

> ▷ CD4 = Time + Gender + Time*Gender
>   ⇒ `cd4 ~ time + gender + time:gender`
>   ⇒ `cd4 ~ time*gender` (the same)

> ▷ CD4 = Time + Time^2
>   ⇒ `cd4 ~ time + I(time^2)`

R> <u>Note</u>: the intercept term is included by default

2.4 Mixed-Effects Models in R (cont'd)

R> The code used to fit the linear mixed model for the AIDS dataset (p. 28) is as
follows

```
lmeFit <- lme(CD4 ~ obstime + obstime:drug, data = aids,
    random = ~ obstime | patient)

summary(lmeFit)
```

R> The same fixed-effects structure but only random intercepts

```
lme(CD4 ~ obstime + obstime:drug, data = aids,
    random = ~ 1 | patient)
```

R> The same fixed-effects structure, random intercepts & random slopes, with a diagonal covariance matrix (using the pdDiag() function)

```
lme(CD4 ~ obstime + obstime:drug, data = aids,
    random = list(patient = pdDiag(form = ~ obstime)))
```

Chapter 3

Relative Risk Models

3.1 Features of Survival Data

- The most important characteristic that distinguishes the analysis of time-to-event outcomes from other areas in statistics is **Censoring**

  ▷ the event time of interest is not fully observed for all subjects under study

- Implications of censoring:

  ▷ standard tools, such as the sample average, the t-test, and linear regression **cannot** be used

  ▷ inferences may be sensitive to misspecification of the distribution of the event times

3.1 Features of Survival Data (cont'd)

- Several types of censoring:

  ▷ Location of the true event time wrt the censoring time: *right*, *left* & *interval*

  ▷ Probabilistic relation between the true event time & the censoring time: *informative* & *non-informative*

  > Here we focus on non-informative right censoring

- <u>Note:</u> Survival times may often be truncated; analysis of truncated samples requires similar calculations as censoring

3.1 Features of Survival Data (cont'd)

- Notation (i denotes the subject)

  ▷ T_i^* 'true' time-to-event

  ▷ C_i the censoring time (e.g., the end of the study or a random censoring time)

- Available data for each subject

  ▷ observed event time: $T_i = \min(T_i^*, C_i)$

  ▷ event indicator: $\delta_i = 1$ if event; $\delta_i = 0$ if censored

> **Our aim is to make valid inferences for T_i^* but using only $\{T_i, \delta_i\}$**

3.2 Relative Risk Models

- **Relative Risk Models** assume a multiplicative effect of covariates on the hazard scale, i.e.,

$$h_i(t) = h_0(t) \exp(\gamma_1 w_{i1} + \gamma_2 w_{i2} + \ldots + \gamma_p w_{ip}) \Rightarrow$$

$$\log h_i(t) = \log h_0(t) + \gamma_1 w_{i1} + \gamma_2 w_{i2} + \ldots + \gamma_p w_{ip},$$

where

$\triangleright$ $h_i(t)$ denotes the hazard of an event for patient i at time t

$\triangleright$ $h_0(t)$ denotes the baseline hazard

$\triangleright$ $w_{i1}, \ldots, w_{ip}$ a set of covariates

- Standard MLE can be applied based on the log-likelihood function

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

  which also can be re-expressed in terms of the hazard function

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

where $p(.)$ density function; $S(.)$ survival function

<div style="border:1px solid black; text-align:center; color:red; font-weight:bold;">

Sensitivity to distributional assumptions due to censoring

</div>

3.2 Relative Risk Models (cont'd)

- **Cox Model:** We make no assumptions for the baseline hazard function

- Parameter estimates and standard errors are based on the log partial likelihood function

$$p\ell(\gamma) = \sum_{i=1}^{n} \delta_i \Big[\gamma^\top w_i - \log\Big\{ \sum_{j:T_j \geq T_i} \exp(\gamma^\top w_j) \Big\} \Big],$$

where only patients who had an event contribute

3.2 Relative Risk Models (cont'd)

- **Example:** For the PBC dataset we are interested in the treatment effect while correcting for sex and age effects

$$h_i(t) = h_0(t) \exp(\gamma_1 \mathtt{D\text{-}penic}_i + \gamma_2 \mathtt{Female}_i + \gamma_3 \mathtt{Age}_i)$$

	Value	HR	Std.Err.	z-value	p-value
γ_1	-0.138	0.871	0.156	-0.882	0.378
γ_2	-0.493	0.611	0.207	-2.379	0.017
γ_3	0.021	1.022	0.008	2.784	0.005

3.3 Relative Risk Models in R

R> The primary package in R for the analysis of survival data is the **survival** package

R> A key function in this package that is used to specify the available event time information in a sample at hand is `Surv()`

R> For right censored failure times (i.e., what we will see in this course) we need to provide the observed event times `time`, and the event indicator `status`, which equals 1 for true failure times and 0 for right censored times

```
Surv(time, status)
```

R> Cox models are fitted using function `coxph()`. For instance, for the PBC data the following code fits the Cox model that contains the main effects of 'drug', 'sex' and 'age':

```
coxFit <- coxph(Surv(years, status2) ~ drug + sex + age,
    data = pbc2.id)

summary(coxFit)
```

R> The two main arguments are a formula specifying the design matrix of the model and a data frame containing all the variables

3.4 Time Dependent Covariates

- Often interest in the association between a time-dependent covariate and the risk of an event

  ▷ treatment changes with time (e.g., dose)

  ▷ time-dependent exposure (e.g., smoking, diet)

  ▷ markers of disease or patient condition (e.g., blood pressure, PSA levels)

  ▷ . . .

- Example: In the PBC study, are the longitudinal bilirubin measurements associated with the hazard of death?

3.4 Time Dependent Covariates (cont'd)

- To answer our questions of interest we need to postulate a model that relates

  ▷ the serum bilirubin with

  ▷ the time-to-death

- The association between **baseline** marker levels and the risk of death can be estimated with standard statistical tools (e.g., Cox regression)

- When we move to the time-dependent setting, a more **careful consideration** is required

3.4 Time Dependent Covariates (cont'd)

- There are two types of time-dependent covariates

  (Kalbfleisch and Prentice, 2002, Section 6.3)

  ▷ Exogenous (aka external): the future path of the covariate up to any time $t > s$ is not affected by the occurrence of an event at time point s, i.e.,

$$\Pr\big\{\mathcal{Y}_i(t) \mid \mathcal{Y}_i(s), T_i^* \geq s\big\} = \Pr\big\{\mathcal{Y}_i(t) \mid \mathcal{Y}_i(s), T_i^* = s\big\},$$

  where $0 < s \leq t$ and $\mathcal{Y}_i(t) = \{y_i(s), 0 \leq s < t\}$

  ▷ Endogenous (aka internal): not Exogenous

3.4 Time Dependent Covariates (cont'd)

- It is **very important** to distinguish between these two types of time-dependent covariates, because the type of covariate dictates the appropriate type of analysis

- In our motivating examples all time-varying covariates are **Biomarkers** $\Rightarrow$ These are always **endogenous** covariates

  ▷ measured with error (i.e., biological variation)

  ▷ the complete history is not available

  ▷ existence directly related to failure status

Subject 127
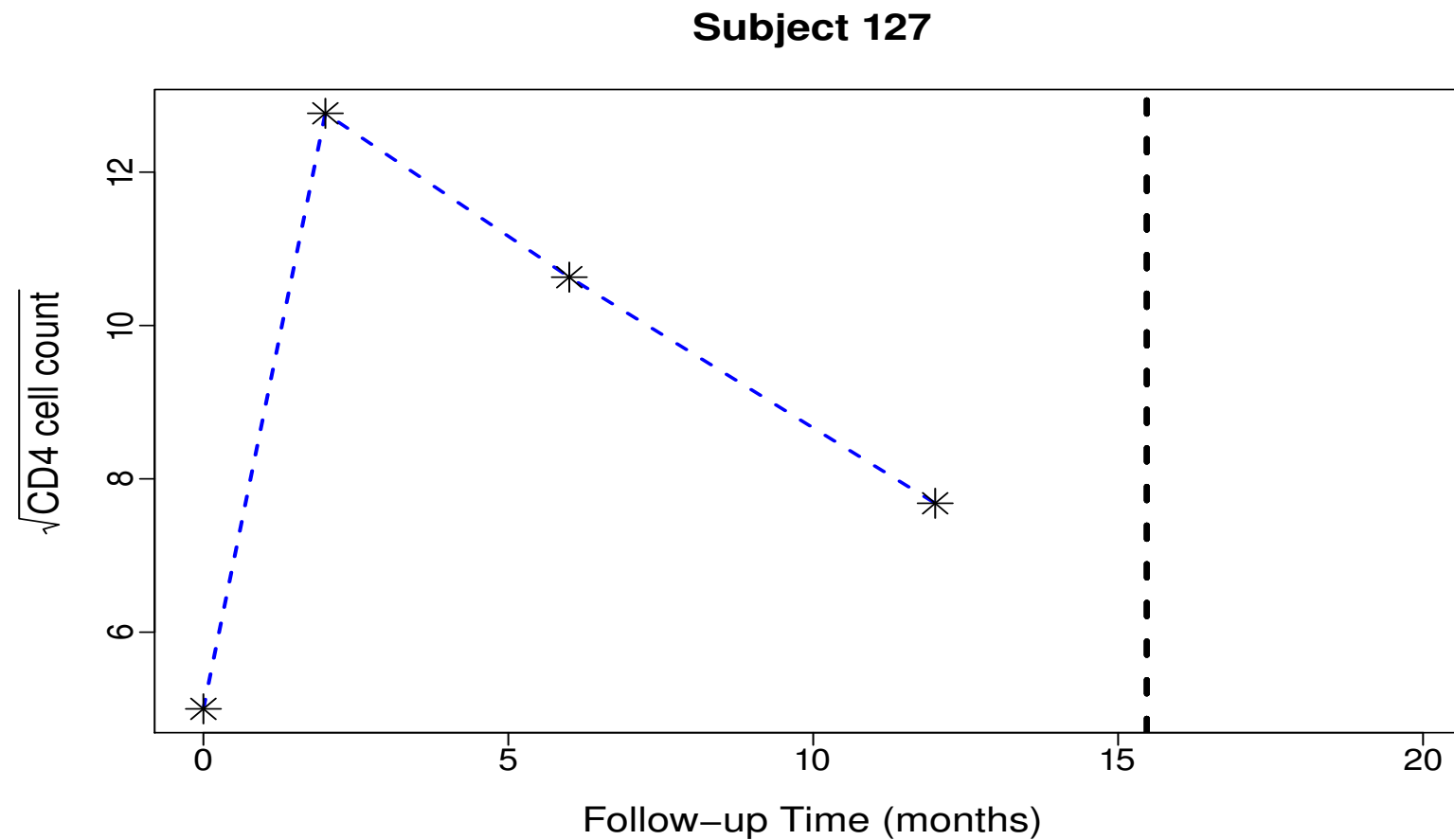
3.5 Extended Cox Model

- The Cox model presented earlier can be extended to handle time-dependent covariates using the counting process formulation

$$h_i(t \mid \mathcal{Y}_i(t), w_i) = h_0(t) \exp\{\gamma^\top w_i + \alpha y_i(t)\},$$

where

▷ $y_i(t)$ denotes the value of the time-varying covariate at t

- Interpretation:

$$h_i(t \mid \mathcal{Y}_i(t), w_i) = h_0(t) \exp\{\gamma^\top w_i + \alpha y_i(t)\}$$

$\exp(\alpha)$ denotes the relative increase in the risk of an event at time t that results from one unit increase in $y_i(t)$ at the same time point
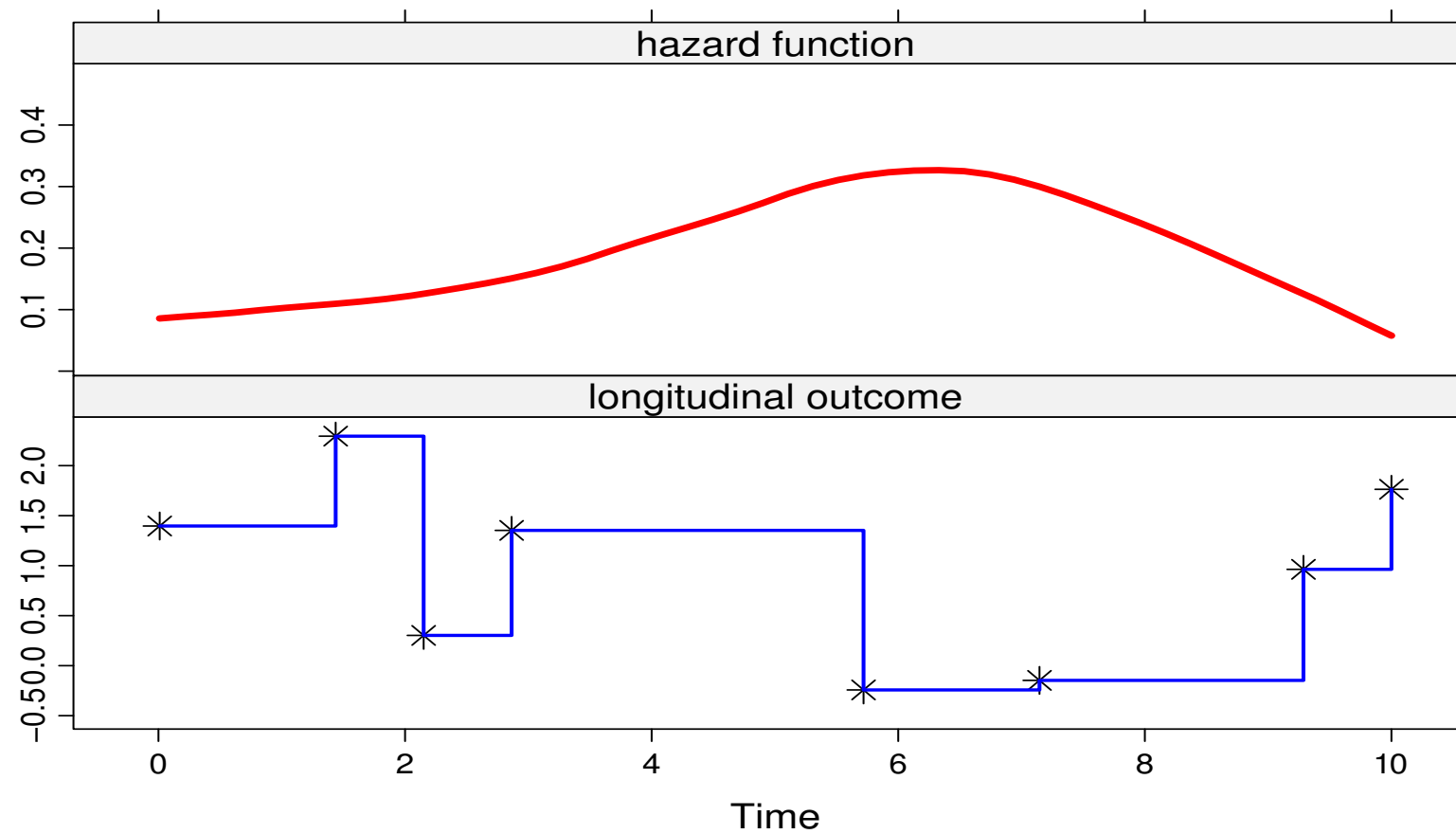
- Typically, data must be organized in the long format

Patient	Start	Stop	Event	$y_i(t)$	Age
1	0	135	1	5.5	45
2	0	65	0	2.2	38
2	65	120	0	3.1	38
2	120	155	1	4.1	38
3	0	115	0	2.5	29
3	115	202	0	2.9	29
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- How does the extended Cox model handle time-varying covariates?

  ▷ assumes no measurement error

  ▷ step-function path

  ▷ existence of the covariate is not related to failure status

3.5 Extended Cox Model (cont'd)

* Therefore, the extended Cox model is only valid for exogenous time-dependent covariates

<div style="border:1px solid black">

Treating endogenous covariates as exogenous may produce spurious results!

</div>

Quiz 1: Is treatment an endogenous or an exogenous time-dependent variable?

Chapter 4

The Basic Joint Model
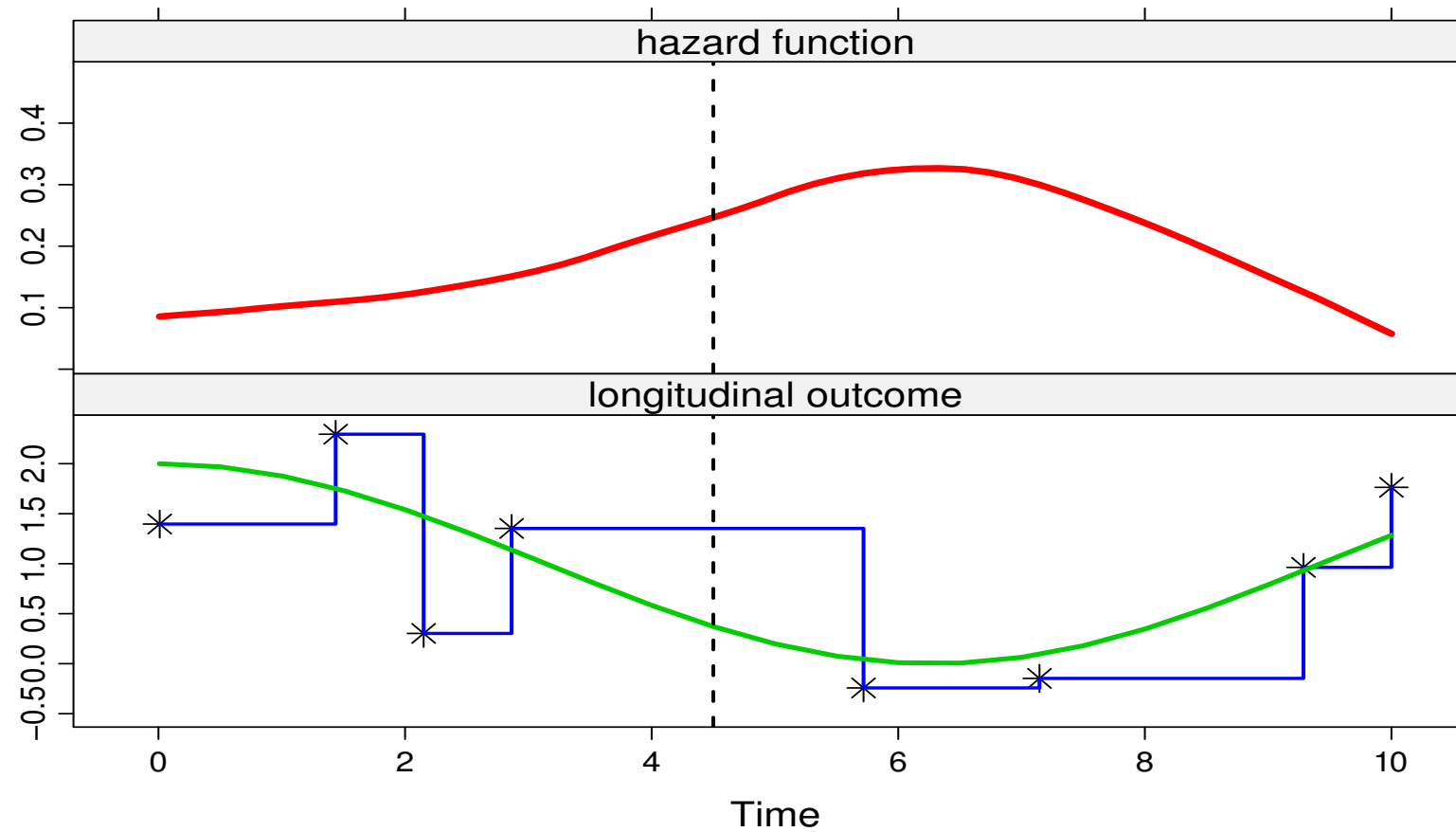
4.1 Joint Modeling Framework

- To account for the special features of endogenous covariates a new class of models has been developed

<div style="text-align:center">

Joint Models for Longitudinal and Time-to-Event Data

</div>

- Intuitive idea behind these models

  1. use an appropriate model to describe the evolution of the marker over time for each patient
  2. the estimated evolutions are then used in a Cox model

- Feature: Marker levels are **not** assumed constant between visits

- Some notation

    ▷ T_i^*: True event time for patient i

    ▷ T_i: Observed event time for patient i

    ▷ δ_i: Event indicator, i.e., equals 1 for true events

    ▷ y_i: Longitudinal responses

- We will formulate the joint model in 3 steps – in particular, . . .

4.1 Joint Modeling Framework (cont'd)

- $\boxed{\text{Step 1:}}$ From the observed longitudinal response $y_i(t)$ reconstruct the covariate history for each subject

- Then, we can define a mixed effects model (we focus, for now, on continuous markers)

$$y_i(t) \;=\; \textcolor{red}{m_i(t)} + \varepsilon_i(t)$$

$$= x_i^\top(t)\beta + z_i^\top(t)b_i + \varepsilon_i(t), \qquad \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2),$$

where

$\triangleright$ $x_i(t)$ and β: Fixed-effects part

$\triangleright$ $z_i(t)$ and b_i: Random-effects part, $b_i \sim \mathcal{N}(0, D)$

- $\boxed{\text{Step 2:}}$ Let's assume that we know $m_i(t)$, i.e., the *true* & *unobserved* value of the marker at time t

- Then, we can define a standard relative risk model

$$h_i\{t \mid \mathcal{M}_i(t)\} = h_0(t) \exp\{\gamma^\top w_i + \alpha m_i(t)\},$$

  where

  $\triangleright$ $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ longitudinal history

  $\triangleright$ α quantifies the strength of the association between the marker and the risk of an event

  $\triangleright$ w_i baseline covariates

4.1 Joint Modeling Framework (cont'd)

- The two processes are associated $\Rightarrow$ define a model for their joint distribution

- Joint Models (JM) for such joint distributions are of the following form
  (Tsiatis & Davidian, Stat. Sinica, 2004)

$$p(y_i, T_i, \delta_i) = \int p(y_i \mid b_i) \left\{ h(T_i \mid b_i)^{\delta_i} S(T_i \mid b_i) \right\} p(b_i) \, db_i,$$

where

  $\triangleright$ b_i a vector of random effects that explains the interdependencies

  $\triangleright$ $p(\cdot)$ density function; $S(\cdot)$ survival function

4.1 Joint Modeling Framework (cont'd)

- Key assumption: **Full Conditional Independence** $\Rightarrow$ random effects explain all interdependencies

  $\triangleright$ the longitudinal outcome is independent of the time-to-event outcome

  $\triangleright$ the repeated measurements in the longitudinal outcome are independent of each other

$$p(y_i, T_i, \delta_i \mid b_i) = p(y_i \mid b_i)\, p(T_i, \delta_i \mid b_i)$$

$$p(y_i \mid b_i) = \prod_j p(y_{ij} \mid b_i)$$

- The survival function, which is a part of the likelihood of the model, depends on the whole longitudinal history

$$S_i(t \mid b_i) = \exp\left(-\int_0^t h_0(s) \exp\{\gamma^\top w_i + \alpha m_i(s)\} \, ds\right)$$

- Therefore, care in the definition of the design matrices of the mixed model

  ▷ when subjects have nonlinear profiles ⇒

  ▷ use splines or polynomials to model them flexibly

- Assumptions for the baseline hazard function $h_0(t)$

  ▷ parametric $\Rightarrow$ possibly restrictive

  ▷ unspecified $\Rightarrow$ within JM framework underestimates standard errors

- It is advisable to use parametric but flexible models for $h_0(t)$

  ▷ splines

$$\log h_0(t) = \gamma_{h_0,0} + \sum_{q=1}^{Q} \gamma_{h_0,q} B_q(t, v),$$

where

* $B_q(t, v)$ denotes the q-th basis function of a B-spline with knots $v_1, \ldots, v_Q$

* γ_{h_0} a vector of spline coefficients

- The censoring and visiting* processes are assumed non-informative:

- Decision to withdraw from the study or appear for the next visit

  ▷ **may depend** on observed past history (baseline covariates + observed longitudinal responses)

  ▷ **no additional dependence** on underlying, latent subject characteristics associated with prognosis

*The visiting process is defined as the mechanism (stochastic or deterministic) that generates the time points at which longitudinal measurements are collected.

- Mainly maximum likelihood but also Bayesian approaches

- The log-likelihood contribution for subject i:

$$\ell_i(\theta) = \log \int \left\{ \prod_{j=1}^{n_i} p(y_{ij} \mid b_i; \theta) \right\} \left\{ p(T_i, \delta_i \mid b_i; \theta) \right\} p(b_i; \theta) \, db_i =$$

$$\log \int \left\{ \prod_{j=1}^{n_i} p(y_{ij} \mid b_i; \theta) \right\} \left\{ h(T_i \mid b_i; \theta)^{\delta_i} \, S_i(T_i \mid b_i; \theta) \right\} p(b_i; \theta) \, db_i,$$

where

$$S_i(t \mid b_i; \theta) = \exp\left(- \int_0^t h_0(s; \theta) \exp\{\gamma^\top w_i + \alpha m_i(s)\} \, ds \right)$$

- Bayes theorem

$$p(\mathsf{B} \mid \mathsf{A}) = \frac{p(\mathsf{A} \mid \mathsf{B}) * p(\mathsf{B})}{p(\mathsf{A})}$$

- Bayes theorem

$$p(\textbf{hypothesis} \mid \textbf{data}) = \frac{p(\textbf{data} \mid \textbf{hypothesis}) * p(\textbf{hypothesis})}{p(\textbf{data})}$$

where hypothesis is typically something unobserved or unknown. It's what you want to learn about using the data.

For regression models, the "hypothesis" is a parameter (intercept, slopes or error terms).

Bayes theorem tells you the probability of the hypothesis given the data.

- How plausible is some hypothesis given the data?

$$p(\mathbf{hypothesis} \mid \mathbf{data}) = \frac{p(\mathbf{data} \mid \mathbf{hypothesis}) * p(\mathbf{hypothesis})}{p(\mathbf{data})}$$

$$\propto p(\mathbf{data} \mid \mathbf{hypothesis}) * p(\mathbf{hypothesis})$$

- Renaming...

$$\text{posterior} \propto \text{data} * \text{prior}$$

Bayes' theorem provides a systematic way to update our knowledge as we encounter new data.

Quiz 2: What do you see in the following picture? Are you a Bayesian or not?

- Bayesian estimation

  ▷ under the Bayesian paradigm both θ and b_i $\{i = 1, \ldots, n\}$ are regarded as parameters

- Inference is based on the full posterior distribution

$$p(\theta, b \mid T, \delta, y) = \frac{\prod_{i=1}^{n} p(y_i \mid b_i; \theta) \, p(T_i, \delta_i \mid b_i; \theta) \, p(b_i; \theta) \, p(\theta)}{\prod_i p(T_i, \delta_i, y_i)}$$

$$\propto \prod_{i=1}^{n} \left\{ p(y_i \mid b_i; \theta) \, p(T_i, \delta_i \mid b_i; \theta) \, p(b_i; \theta) \right\} p(\theta)$$

4.4 Bayesian Estimation (cont'd)

- For the standard joint model we have define thus far, the majority of the parameters can be updated using Gibbs sampling (or slice sampling) $\Rightarrow$ **MCMC**

  ▷ when no close-form posterior conditionals are available, we can use the Metropolis-Hastings algorithm

- Good proposal distributions can be obtained from the separate fits of the two submodels

• Inference then proceeds in the usual manner from the MCMC output, e.g.,

▷ posterior means, variances, and standard errors

▷ credible intervals

▷ Bayes factors

▷ DIC, CPO

▷ . . .

4.4 Bayesian Estimation (cont'd)

- for **frequentists**, a probability is a measure of the frequency of repeated events, so the interpretation is that parameters are fixed (but unknown), and data are random

- for **Bayesians**, a probability is a measure of the degree of certainty about values, so the interpretation is that parameters are random and data are fixed

- Example: To illustrate the virtues of joint modeling, we compare it with the standard time-dependent Cox model for the AIDS data

$$
\begin{cases}
y_i(t) &= m_i(t) + \varepsilon_i(t) \\
&= \beta_0 + \beta_1 t + \beta_2\{t \times \mathtt{ddI}_i\} + b_{i0} + b_{i1}t + \varepsilon_i(t), \qquad \varepsilon_i(t) \sim \mathcal{N}(0, \sigma^2), \\[2ex]
h_i(t) &= h_0(t)\exp\{\gamma\mathtt{ddI}_i + \alpha m_i(t)\},
\end{cases}
$$

where

$\triangleright$ $h_0(t)$ is assumed P-splines

	JMbayes	Cox
	log HR (std.err)	log HR (std.err)
Treat	0.342 (0.008)	0.309 (0.147)
CD4$^{1/2}$	-0.297 (0.002)	-0.193 (0.024)

- Clearly, there is a considerable effect of ignoring the measurement error, especially for the CD4 cell counts

• A unit decrease in $CD4^{1/2}$, results in a

  ▷ **Joint Model**: 1.3-fold increase in risk (95% CI: 1.24; 1.47)

  ▷ **Time-Dependent Cox**: 1.2-fold increase in risk (95% CI: 1.16; 1.27)

• **Which one to believe?**

  ▷ a lot of theoretical and simulation work has shown that the Cox model underestimates the true association size of markers

4.6 Joint Models in R

R> Joint models are fitted using function `jointModelBayes()` from package **JMbayes**. This function accepts as main arguments a linear mixed model and a Cox PH model based on which it fits the corresponding joint model

```
lmeFit <- lme(CD4 ~ obstime + obstime:drug,
    random = ~ obstime | patient, data = aids)

coxFit <- coxph(Surv(Time, death) ~ drug, data = aids.id, x = TRUE)

jointFit <- jointModelBayes(lmeFit, coxFit, timeVar = "obstime",
    baseHaz = c("P-splines"))

summary(jointFit)
```

4.6 Joint Models in R (cont'd)

R> As before, the data frame given in `lme()` should be in the long format, while the data frame given to `coxph()` should have one line per subject*

  ▷ the ordering of the subjects needs to be the same

R> In the call to `coxph()` you need to set `x = TRUE` (or `model = TRUE`) such that the design matrix used in the Cox model is returned in the object fit

R> Argument `timeVar` specifies the time variable in the linear mixed model

* Unless you want to include exogenous time-varying covariates or handle competing risks

R> Argument `baseHaz` specifies the type of relative risk model

Available options are:

▷ `"regression-splines"`: B-spline basis function

▷ `"P-splines"`: B-spline basis function with penatlies (P-splines)

R> details about **JMbayes**:

▷ directly implements the MCMC

▷ allows for categorical longitudinal data as well

▷ allows for general transformation functions

▷ penalized B-splines for the baseline hazard function

▷ . . .

4.6 Joint Models in R (cont'd)

R> In both packages methods are available for the majority of the standard generic functions + extras

- ▷ `summary()`, `anova()`, `vcov()`, `logLik()`

- ▷ `coef()`, `fixef()`, `ranef()`

- ▷ `fitted()`, `residuals()`

- ▷ `plot()`

- ▷ `xtable()` (you need to load package **xtable** first)

Quiz 3: What is the interpretation of the parameters in the survival submodel from the output of the basic joint model (app)?
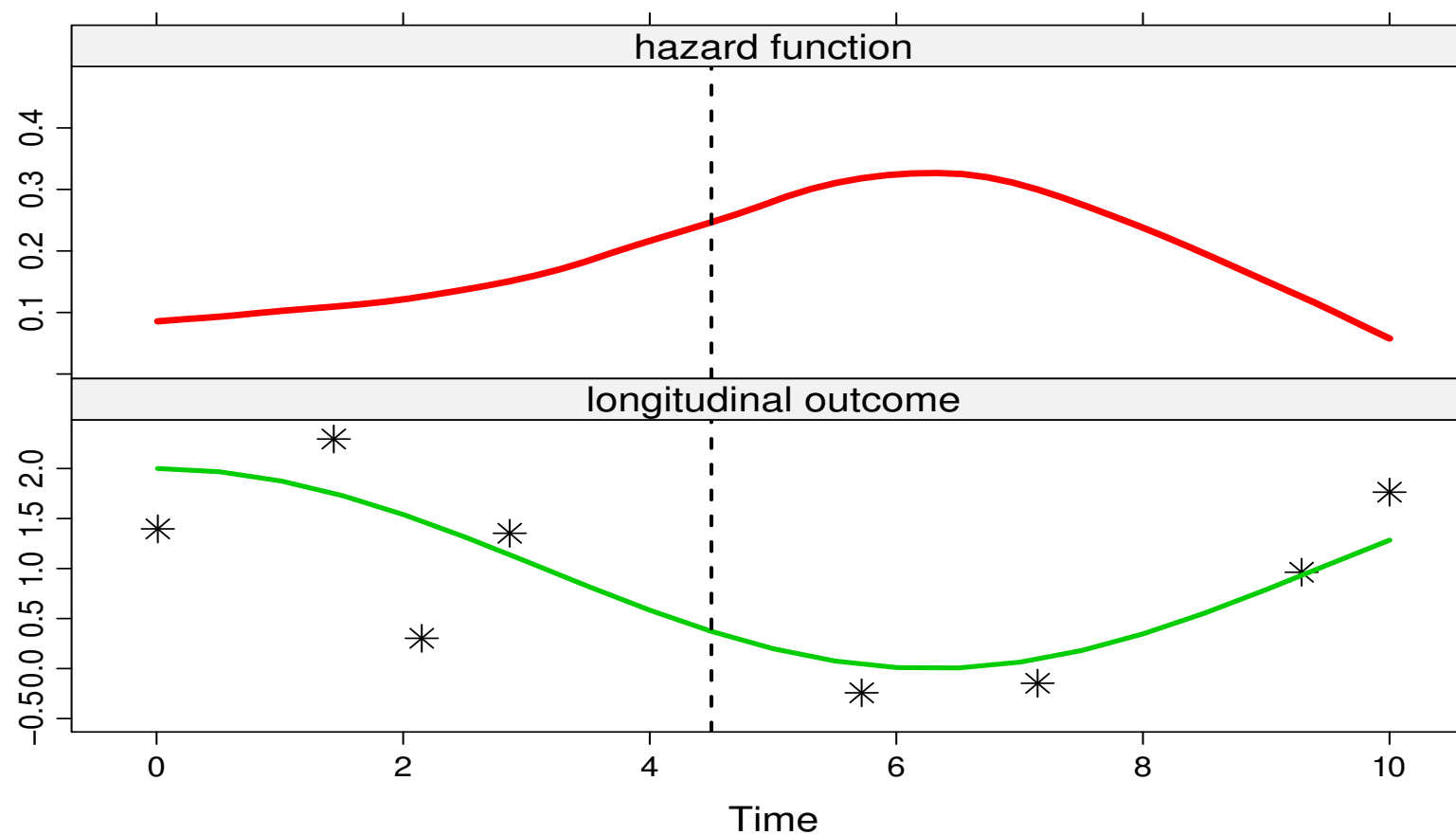
Chapter 5

Extensions of Joint Models

- The standard joint model

$$
\begin{cases}
h_i\{t \mid \mathcal{M}_i(t)\} \;=\; h_0(t)\exp\{\gamma^\top w_i + \alpha m_i(t)\}, \\[2em]
y_i(t) \qquad\;=\; m_i(t) + \varepsilon_i(t) \\[1em]
\qquad\qquad\;=\; x_i^\top(t)\beta + z_i^\top(t)b_i + \varepsilon_i(t),
\end{cases}
$$

where $\mathcal{M}_i(t) = \{m_i(s), 0 \le s < t\}$

- The standard joint model

$$
\begin{cases}
h_i\{t \mid \mathcal{M}_i(t)\} &= h_0(t)\exp\{\gamma^\top w_i + \alpha m_i(t)\}, \\
\\
y_i(t) &= m_i(t) + \varepsilon_i(t) \\
\\
&= x_i^\top(t)\beta + z_i^\top(t)b_i + \varepsilon_i(t),
\end{cases}
$$

where $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$

Is this the only option? Is this the most optimal choice?

5.1 Parameterizations (cont'd)

- Note: Inappropriate modeling of time-dependent covariates may result in surprising results

- Example: Cavender et al. (1992, J. Am. Coll. Cardiol.) conducted an analysis to test the effect of cigarette smoking on survival of patients who underwent coronary artery surgery

  ▷ the estimated effect of current cigarette smoking was positive on survival although not significant (i.e., patients who smoked had higher probability of survival)

  ▷ most of those who had died were smokers but many stopped smoking at the last follow-up before their death

> **We need to carefully consider the functional form of time-dependent covariates**

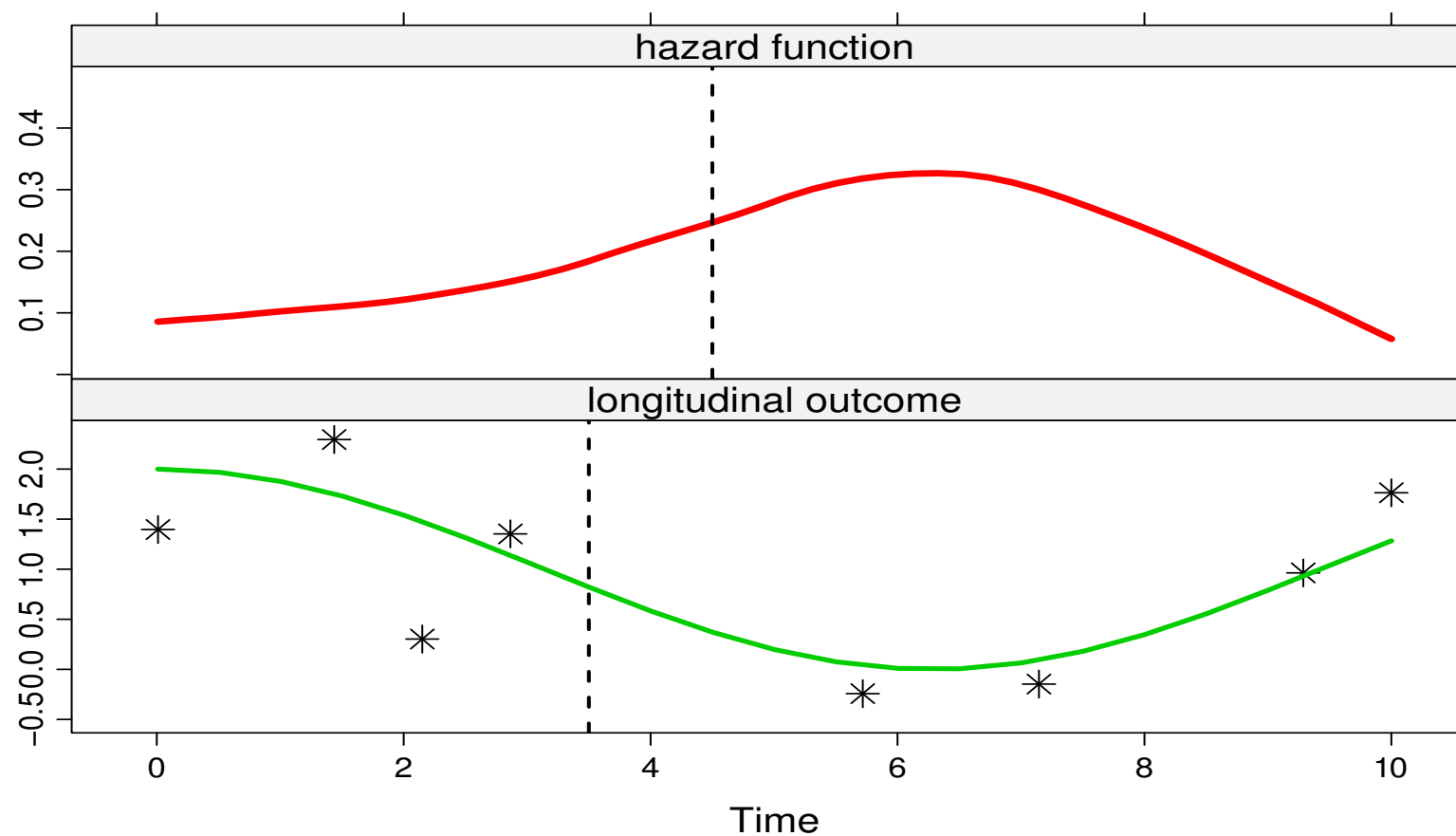- Let's see some possibilities. . .

- *Lagged Effects:* The hazard of an event at t is associated with the level of the marker at a previous time point:

$$h_i\{t \mid \mathcal{M}_i(t)\} \;=\; h_0(t) \exp\{\gamma^\top w_i + \alpha m_i(t_+^c)\},$$

where

$$t_+^c = \max(t - c, 0)$$

- *Time-dependent Slopes:* The hazard of an event at t is associated with both the current value and the slope of the trajectory at t (Ye et al., 2008, Biometrics):

$$h_i\{t \mid \mathcal{M}_i(t)\} = h_0(t) \exp\{\gamma^\top w_i + \alpha_1 m_i(t) + \alpha_2 m_i'(t)\},$$

where

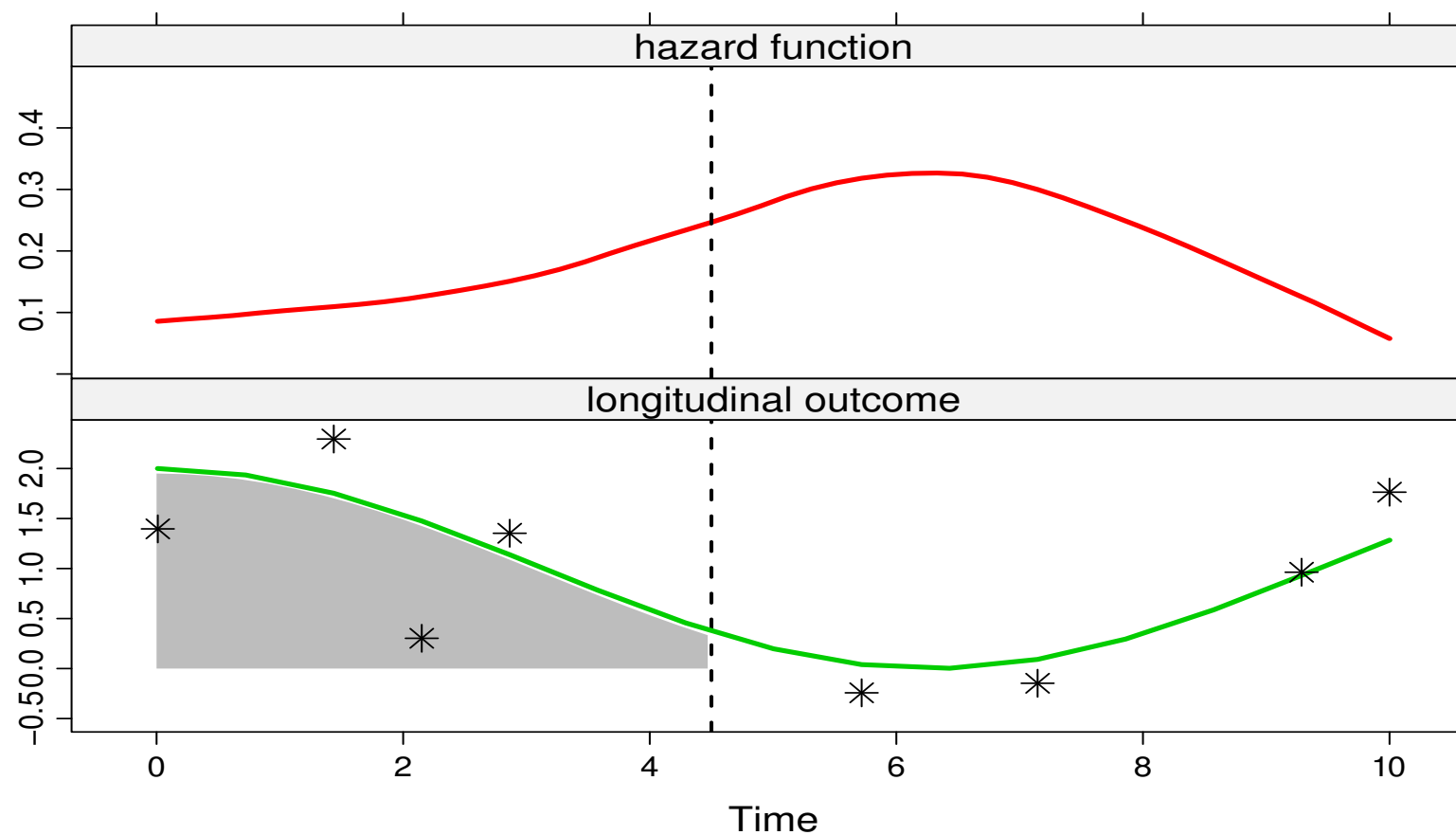$$m_i'(t) = \frac{d}{dt}\{x_i^\top(t)\beta + z_i^\top(t)b_i\}$$
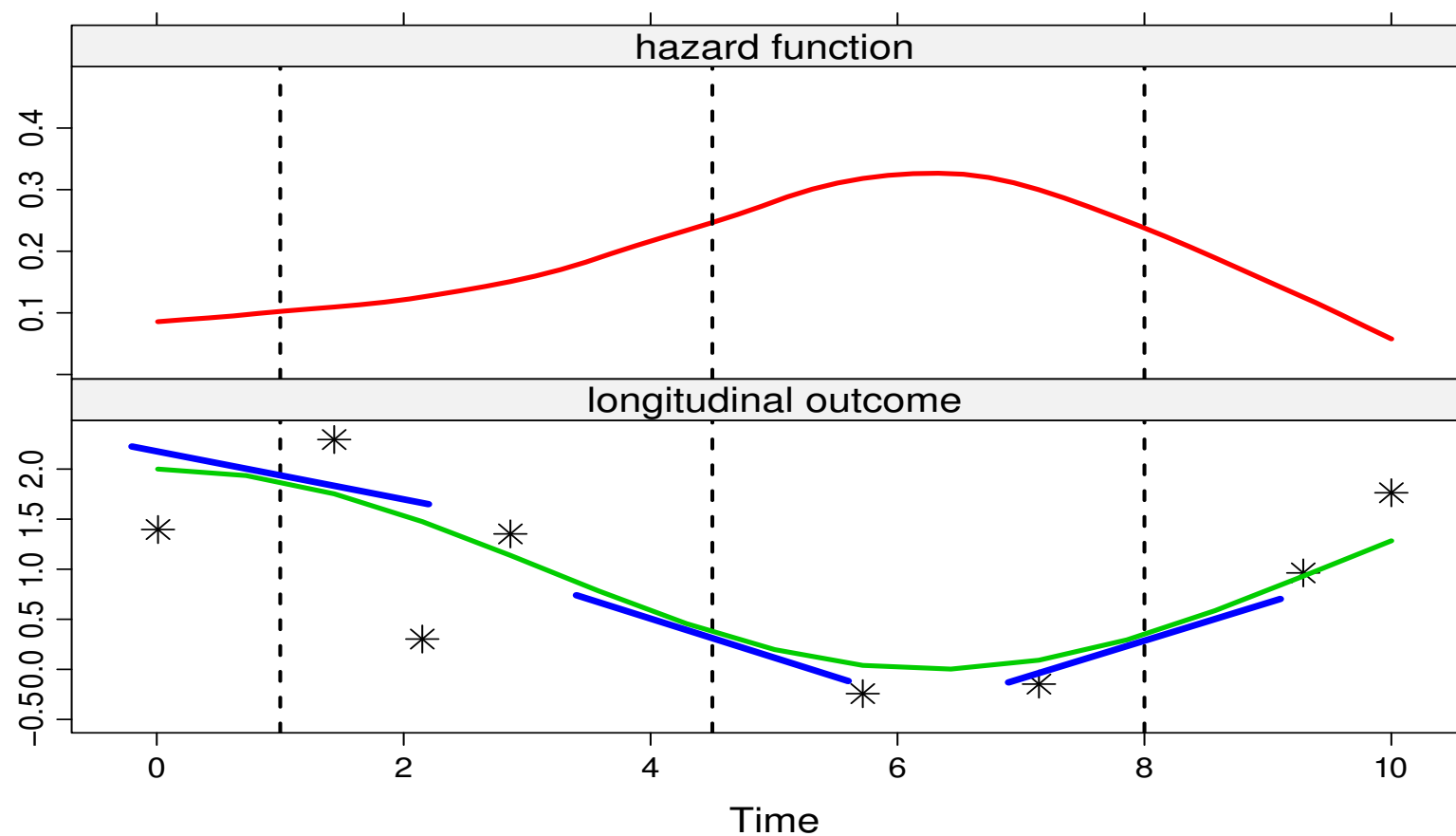
- *Cumulative Effects:* The hazard of an event at t is associated with the whole area under the trajectory up to t:

$$h_i\{t \mid \mathcal{M}_i(t)\} = h_0(t) \exp\left\{\gamma^\top w_i + \alpha \int_0^t m_i(s) \, ds\right\}$$

- Area under the longitudinal trajectory taken as a summary of $\mathcal{M}_i(t)$

- *Weighted Cumulative Effects (convolution):* The hazard of an event at t is associated with the area under the weighted trajectory up to t:

$$h_i\{t \mid \mathcal{M}_i(t)\} = h_0(t) \exp\left\{\gamma^\top w_i + \alpha \int_0^t \varpi(t-s) m_i(s) \, ds\right\},$$

where $\varpi(\cdot)$ an appropriately chosen weight function, e.g.,

▷ Gaussian density

▷ Student's-t density

▷ . . .

- *Random Effects:* The hazard of an event at t is associated <u>only</u> with the random effects of the longitudinal model:

$$h_i\{t \mid \mathcal{M}_i(t)\} = h_0(t)\exp(\gamma^\top w_i + \alpha^\top b_i)$$

- Features:

  ▷ avoids numerical integration for the survival function

  ▷ interpretation of α more difficult, especially in high-dimensional random-effects settings

- $\boxed{\text{Example:}}$ Sensitivity of inferences for the longitudinal process to the choice of the parameterization for the AIDS data

- We use the same mixed model as before, i.e.,

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

$$= \beta_0 + \beta_1 t + \beta_2 \{t \times \mathtt{ddI}_i\} + b_{i0} + b_{i1} t + \varepsilon_i(t)$$

and the following four survival submodels

5.1 Parameterizations (cont'd)

- Model I (current value)

$$h_i(t) = h_0(t) \exp\{\gamma \mathtt{ddI}_i + \alpha_1 m_i(t)\}$$

- Model II (current value + current slope)

$$h_i(t) = h_0(t) \exp\{\gamma \mathtt{ddI}_i + \alpha_1 m_i(t) + \alpha_2 m_i'(t)\},$$

where

  $\triangleright m_i'(t) = \beta_1 + \beta_2 \mathtt{ddI}_i + b_{i1}$

- Model III (random slope)

$$h_i(t) = h_0(t) \exp\{\gamma \mathtt{ddI}_i + \alpha_3 b_{i1}\}$$
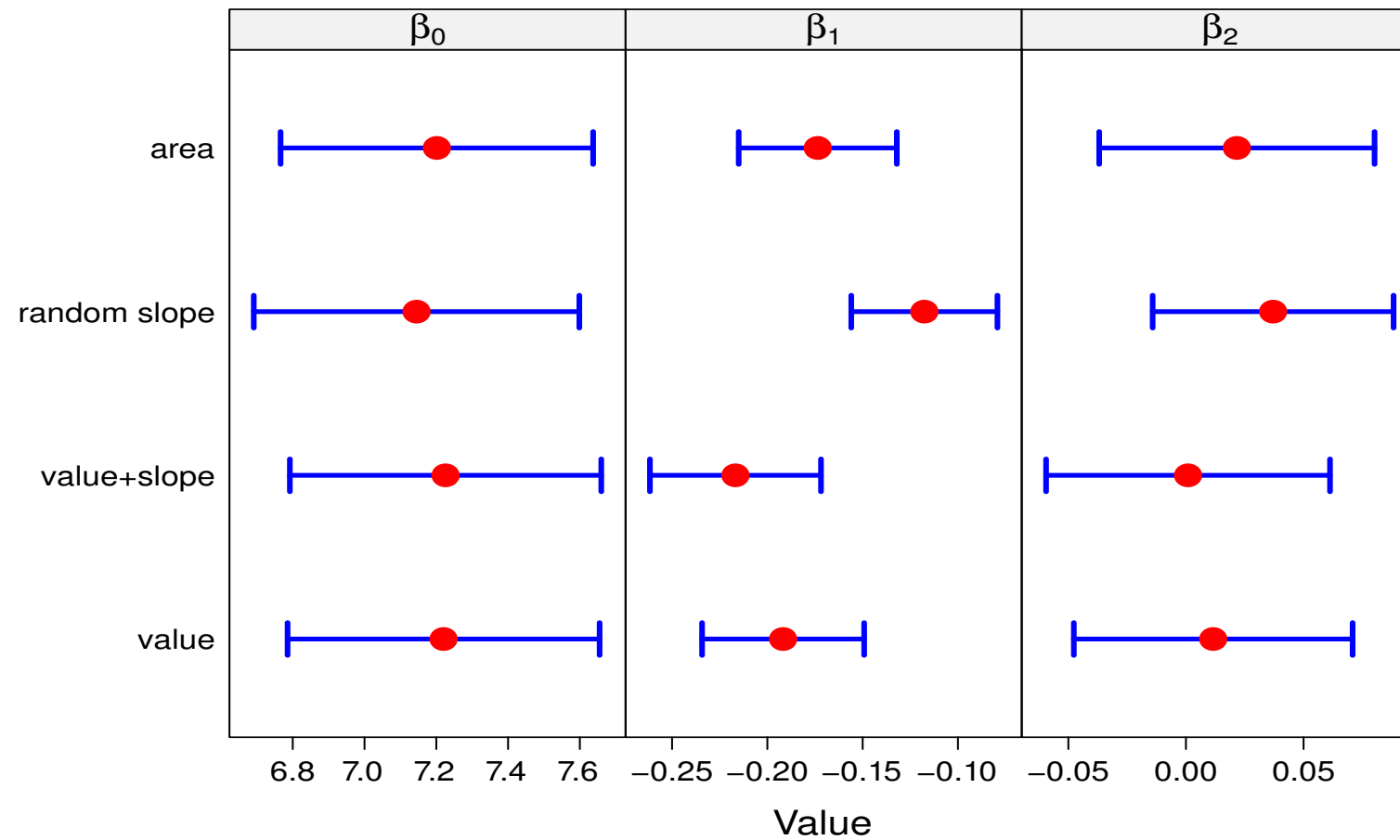
- Model IV (area)

$$h_i(t) = h_0(t) \exp\left\{\gamma \mathtt{ddI}_i + \alpha_4 \int_0^t m_i(s) \, ds\right\},$$

where

▷ $\int_0^t m_i(s) \, ds = \beta_0 t + \frac{\beta_1}{2} t^2 + \frac{\beta_2}{2} \{t^2 \times \mathtt{ddI}_i\} + b_{i0} t + \frac{b_{i1}}{2} t^2$

5.1 Parameterizations (cont'd)

- There are noticeable differences between the parameterizations

  ▷ especially in the slope parameters

- Therefore, a sensitivity analysis should not stop at the standard joint model parameterization but also consider alternative association structures

5.2 Parameterizations in R

R> Lagged effects can be fitted using the `lag` argument of `jointModelBayes()`. For example, the following code fits a joint model for the PBC dataset with

▷ random intercepts and random slopes for log serum bilirubin, and

▷ a relative risk model with P-splines baseline hazard and the *true* effect at the previous year

```
lmeFit <- lme(log(serBilir) ~ year, random = ~ year | id, data = pbc2)

coxFit <- coxph(Surv(years, status2) ~ 1, data = pbc2.id, x = TRUE)

jointFit <- jointModelBayes(lmeFit, coxFit, timeVar = "year",
    baseHaz = "P-splines", lag = 1)

summary(jointFit)
```

5.2 Parameterizations in R (cont'd)

R> For the time-dependent slopes and cumulative effects parameterizations, arguments `param` and `extraForm` of `jointModelBayes()` should be used

▷ Options are

* `param = "td-value"`
* `param = "td-extra"`
* `param = "td-both"`
* `param = "shared-betasRE"`
* `param = "shared-RE"`

5.2 Parameterizations in R (cont'd)

R> Include also the slope of the longitudinal biomarker

```
dform = list(fixed = ~ 1, random = ~ 1,
    indFixed = 2, indRandom = 2)

jointFit2 <- jointModelBayes(lmeFit, coxFit, timeVar = "year",
    param = "td-both", extraForm = dform, baseHaz = "P-splines")

summary(jointFit2)
```

5.3 Multiple Biomarkers

- So far we have concentrated on a single continuous marker

- But very often we may have several markers we wish to study, some of which could be categorical

- Example: In the PBC dataset we have used serum bilirubin as the most important marker, but during follow-up several other markers have been recorded

  ▷ serum cholesterol (continuous)

  ▷ edema (3 categories)

  ▷ ascites (2 categories)

  ▷ ...

5.3 Multiple Biomarkers (cont'd)

> **We need to extend the basic joint model!**

- To handle multiple longitudinal markers of different types we use Generalized Linear Mixed Models

  ▷ We assume $Y_{i1}, \ldots, Y_{iJ}$ for each subject, each one having a distribution in the exponential family, with expected value

$$m_{ij}(t) = E(y_{ij}(t) \mid b_{ij}) = g_j^{-1}\{x_{ij}^\top(t)\beta_j + z_{ij}^\top(t)b_{ij}\},$$

  with $g(\cdot)$ denoting a link function

▷ Correlation between the outcomes is built by assuming a multivariate normal distribution for the random effects

$$b_i = (b_{i1}^\top, \ldots, b_{iJ}^\top)^\top \sim \mathcal{N}(0, D)$$

• The expected value of each longitudinal marker is incorporated in the linear predictor of the survival submodel

$$h_i\{t \mid \mathcal{M}_i(t)\} = h_0(t) \exp\Big\{\gamma^\top w_i + \sum_{j=1}^{J} \alpha_j m_{ij}(t)\Big\},$$

where $\mathcal{M}_i(t) = \{m_{ij}(s), 0 \le s < t, 1 \le j \le J\}$

5.4 Multiple Biomarkers in R

R>   Joint models for multiple longitudinal outcomes can be fitted with function `mvJointModelBayes()` from package **JMbayes**

- The use of this function mimics the one of `jointModelBayes()` but with some small differences, namely

  ▷ we fit a *multivariate* mixed model using `mvglmer()`,

  ▷ following we fit a Cox model using `coxph()`, and

  ▷ and we give the resulting objects as input in `mvJointModelBayes()`

5.4 Multiple Biomarkers in R (cont'd)

R> An example for the PBC dataset using serum bilirubin (continuous) and spiders (binary)

```
multMixedFit <- mvglmer(list(log(serBilir) ~ year + (year | id),
                             spiders ~ year + (1 | id)), data = pbc2,
                        families = list(gaussian, binomial))

coxFit <- coxph(Surv(years, status2) ~ drug + age, data = pbc2.id,
                model = TRUE)

multJMFit <- mvJointModelBayes(multMixedFit, coxFit, timeVar = "year")
summary(multJMFit)
```

R> Function `mvJointModelBayes()` also allows for

    ▷ right, left, interval censored data

    ▷ left truncated data

    ▷ exogenous time-varying covariates

5.5 Variable Selection

- <u>Note:</u> In the previous extension of joint models,

  ▷ multiple longitudinal markers

  we used the default parameterization that includes the current value term $m_i(t)$ in the linear predictor of the survival submodel(s)

Nonetheless, all the other parameterizations we have seen earlier are also applicable

- For example in the case of multiple longitudinal outcomes

$$g_j\big[E\{y_{ij}(t) \mid b_{ij}\}\big] = m_{ij}(t) = x_{ij}^\top(t)\beta_j + z_{ij}^\top(t)b_{ij}$$

$$h_i\{t \mid \mathcal{M}_i(t)\} = h_0(t)\exp\Big\{\gamma^\top w_i + \sum_{j=1}^{J}\sum_{l=1}^{L} f_{jl}(m_{ij}(t), \alpha_{jl})\Big\}$$

where $\mathcal{M}_i(t) = \{m_{ij}(s), 0 \le s < t, 1 \le j \le J\}$

- In this case we face a challenging model selection problem

- Different possible solutions

  ▷ lasso

  ▷ ridge

  ▷ horseshoe

  ▷ . . .

5.6 Variable Selection in R

R>   Function `mvJointModelBayes()` also allows to consider multiple parameterization per outcome in a similar manned as `jointModelBayes()` does

R>   It also implements a global-local ridge-type prior for the association parameters

$$\alpha_{jl} \sim \mathcal{N}(0, \tau \psi_{jl})$$

$$\tau^{-1} \sim Gamma(0.1, 0.1)$$

$$\psi_{jl}^{-1} \sim Gamma(1, 0.01)$$

R> An example for the PBC dataset using serum bilirubin (continuous) and spiders (binary)

```
Forms <- list("log(serBilir)" = "value",
              "log(serBilir)" = list(fixed = ~ 1, random = ~ 1,
                          indFixed = 2, indRandom = 2, name = "slope"),
              "spiders" = list(fixed = ~ 0 + year + I(year^2/2), random = ~ 0 + year,
                          indFixed = 1:2, indRandom = 1, name = "area"))

multJMFit2 <- update(multJMFit, Formulas = Forms)
summary(multJMFit2)

Ints <- list("log(serBilir)" = ~ drug, "log(serBilir)_slope" = ~ drug,
             "spiders_area" = ~ drug)
multJMFit3 <- update(multJMFit2, Interactions = Ints,
             priors = list(shrink_alphas = TRUE))
summary(multJMFit3)
```
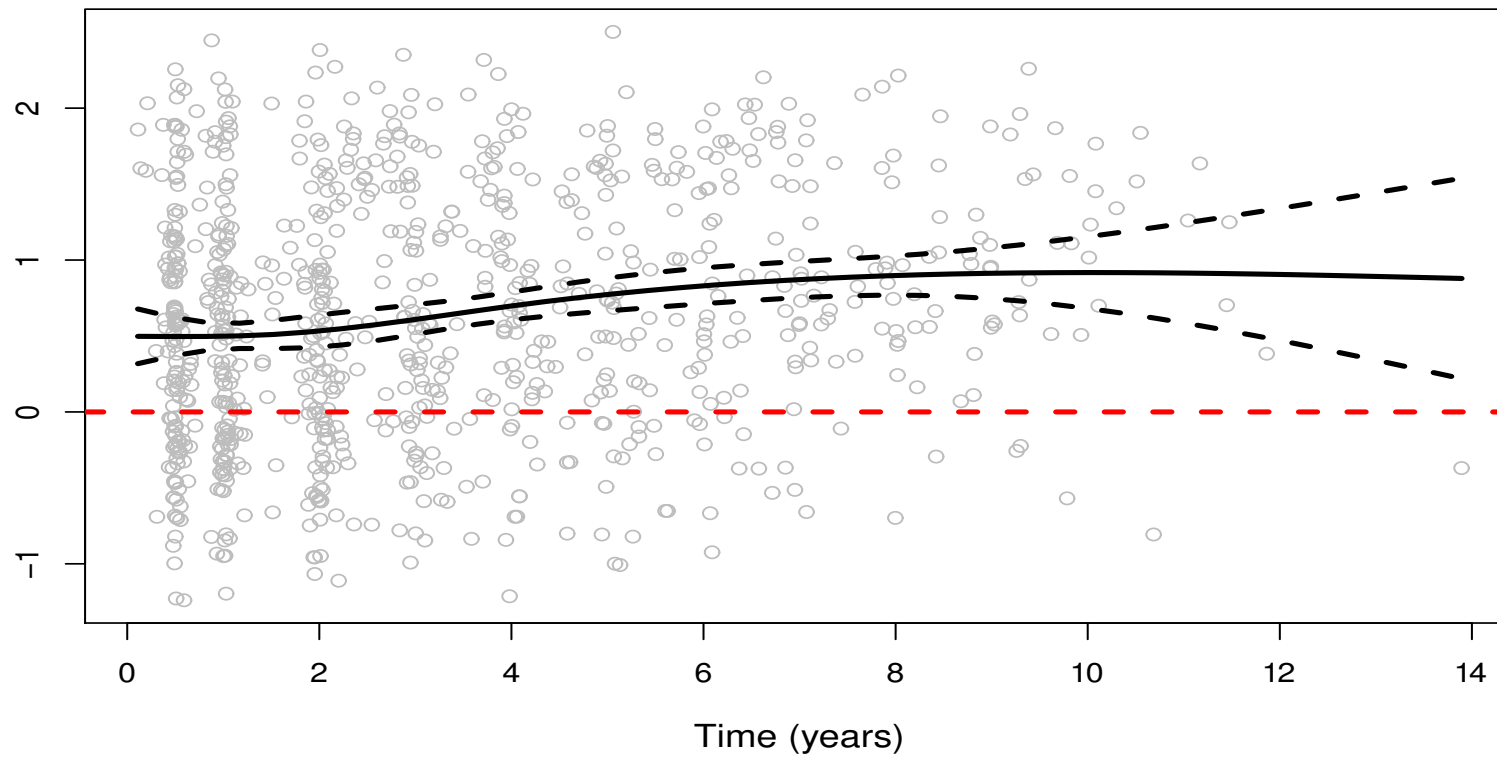
- Standard joint models assume a constant regression coefficient for the effect of the covariates.

  ▷ when treatment is initiated, the strength of the association between the longitudinal and survival outcomes may also change

Time−varying coefficient for log Serum Bilirubin

A time-varying coefficient joint model (VCJM)

Specifically,

$$h_i\{t \mid \mathcal{M}_i(t)\} = h_0(t) \exp[\gamma^\top w_i + \sum_{j=1}^{J} \sum_{l=1}^{L} f_{j\ell}\{m_{ij}(t), \lambda_{j\ell}(t)\}],$$

where

- w_i is a vector of baseline covariates with a corresponding vector of regression coefficients γ

- $f_{j\ell}\{m_{ij}(t), \lambda_{j\ell}(t)\}$ is the outcome (j = 1, ..., J) and the form of association (ℓ = 1, ..., L) between the longitudinal and the survival outcomes

- We consider estimation of the function $\lambda_{j\ell}(t)$ using the regression P-spline method, where

$$\lambda_{j\ell}(t) = \sum_{u=1}^{U} \alpha_{j\ell u} B_{j\ell u}(t, \nu),$$

where

  - $\alpha_{j\ell u}$ is a set of parameters that capture the strength of association between the longitudinal and survival outcomes

  - $B_{j\ell u}(t, \nu)$ denotes the q-th basis function of a B-spline with knots $\nu_1, \ldots, \nu_Q$

- The idea behind the P-spline method is to assume a high number of knots and penalize the coefficients to tackle the problem of the large number or parameters.

R>   An example for the PBC dataset using serum bilirubin (continuous) and spiders (binary)

```
Ints_tveffect <- list("log(serBilir)_value" = ~ 0 + tve(years, df = 8))
JMFit_tveffect <- mvJointModelBayes(multMixedFit, coxFit, timeVar = "year",
                       Interactions = Ints_tveffect)

plot(JMFit_tveffect, which = "tv_effect")
```

Chapter 6

Dynamic Predictions

6.1 Survival Probabilities

- Nowadays there is great interest for prognostic models and their application to personalized medicine

- Examples are numerous

  ▷ cancer research, cardiovascular diseases, HIV research, . . .

<div style="border:1px solid black;padding:1em;text-align:center;color:red;font-weight:bold">
Physicians are interested in accurate prognostic tools that will inform them about the future prospect of a patient in order to adjust medical care
</div>

6.1 Survival Probabilities (cont'd)

- We are interested in predicting survival probabilities for a new patient j that has provided a set of serum bilirubin measurements up to a specific time point t

- Example: We consider Patients 2 and 25 from the PBC dataset that have provided us with 9 and 12 serum bilirubin measurements, respectively

  ▷ **Dynamic Prediction** survival probabilities are dynamically updated as additional longitudinal information is recorded

- We need to account for the endogenous nature of the marker

  ▷ providing measurements up to time point $t \Rightarrow$ the patient was still alive at time t
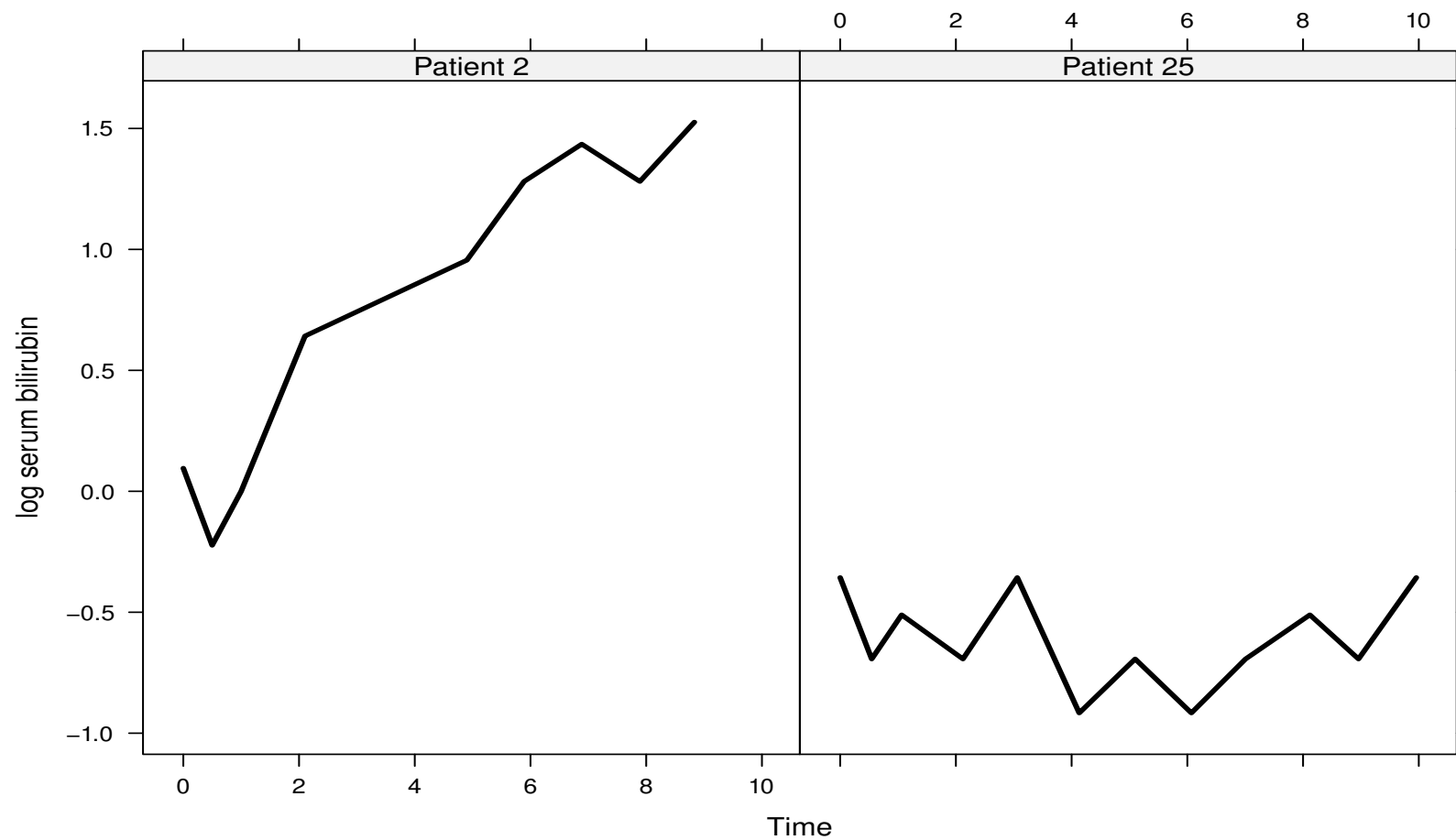
- More formally, for a new subject j we have available measurements up to time point t

$$\mathcal{Y}_j(t) = \{y_j(s), 0 \leq s \leq t\}$$

and we are interested in

$$\pi_j(u \mid t) = \Pr\{T_j^* \geq u \mid T_j^* > t, \mathcal{Y}_j(t), \mathcal{D}_n\},$$

where

▷ where $u > t$, and

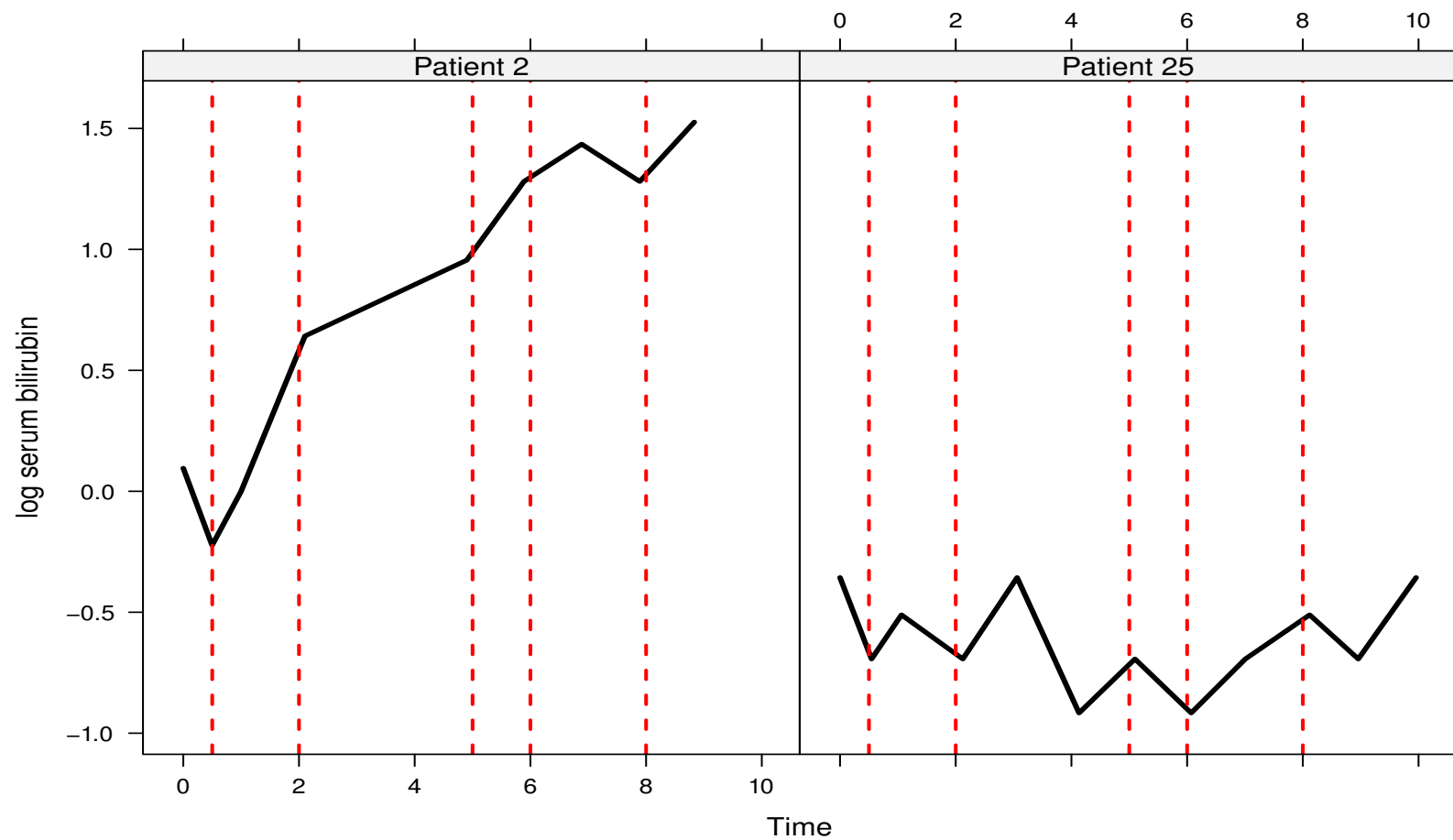▷ $\mathcal{D}_n$ denotes the sample on which the joint model was fitted

6.1 Survival Probabilities (cont'd)

- We assume that the joint model has been fitted to the data at hand

- Based on the fitted model we can estimate the conditional survival probabilities
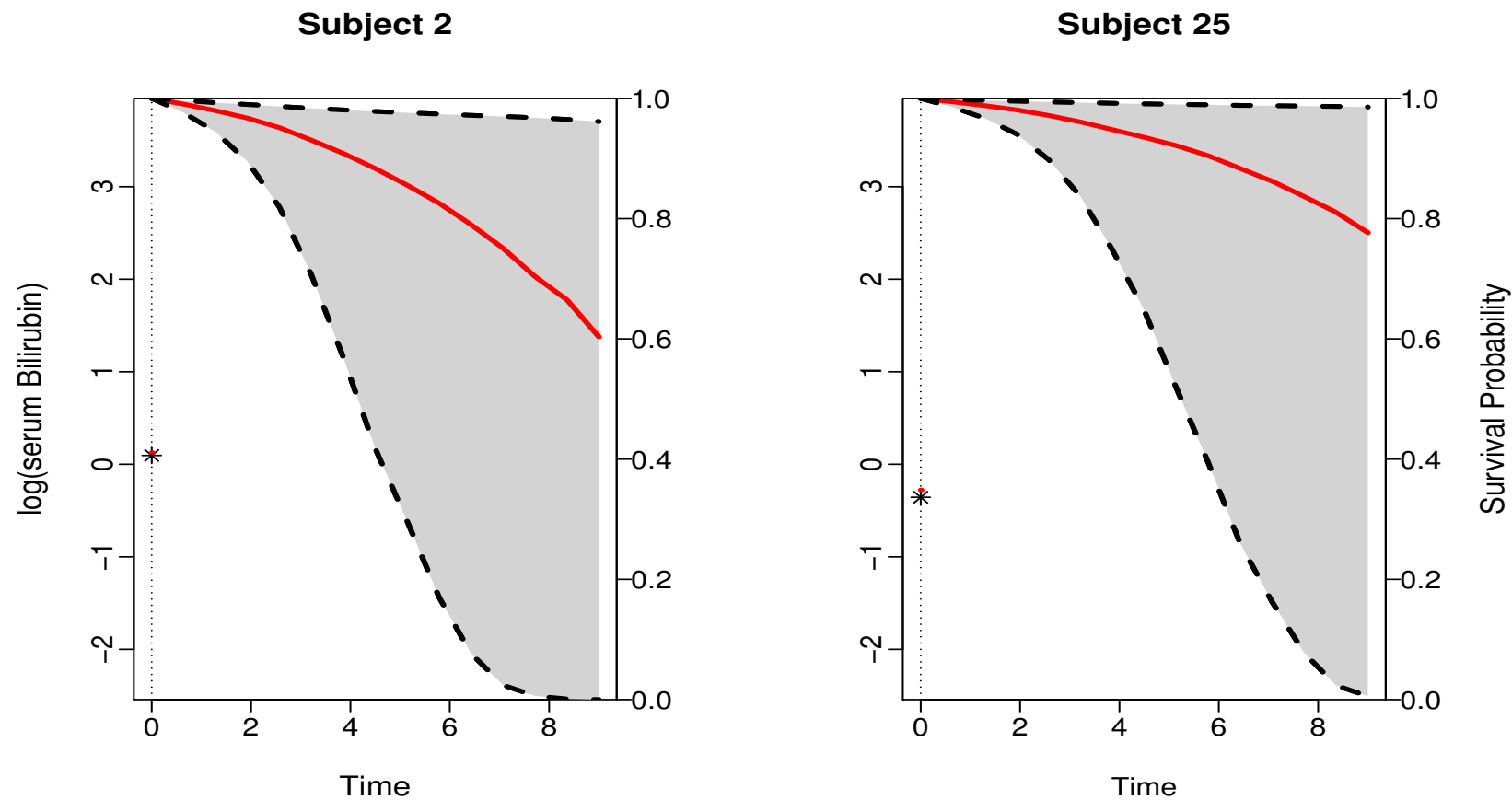 (Rizopoulos, 2011, Biometrics)

- Example: Dynamic predictions of survival probabilities for Patients 2 and 25 from the PBC dataset: We fit the joint model

- Longitudinal submodel

  ▷ fixed effects: Linear & quadratic time, treatment and their interaction

  ▷ random effects: Intercept, linear & quadratic time effects

- Survival submodel

  ▷ treatment effect $+$ *underlying* serum bilirubin level

  ▷ P-splines
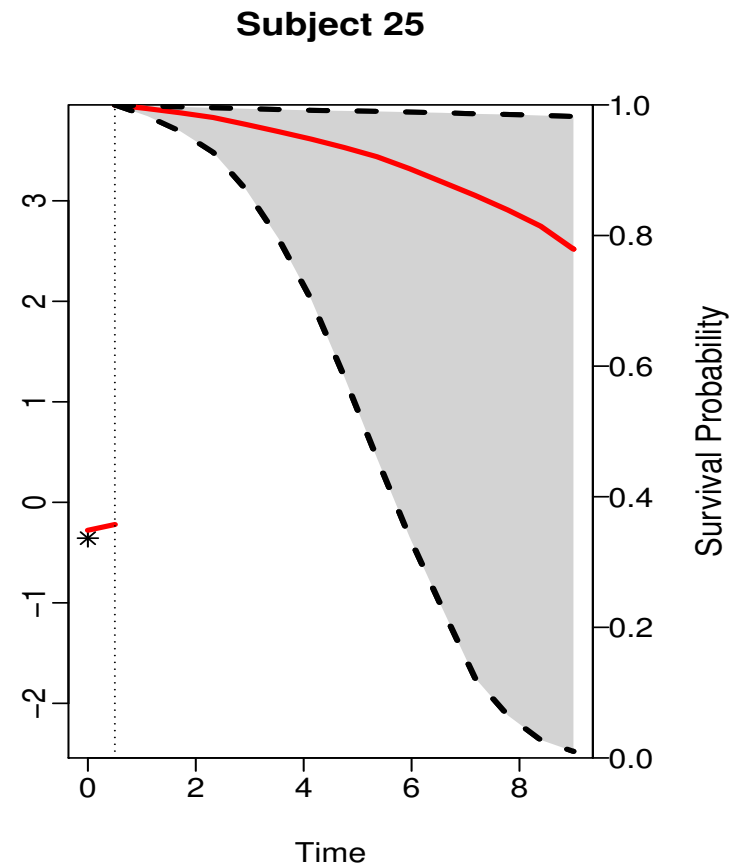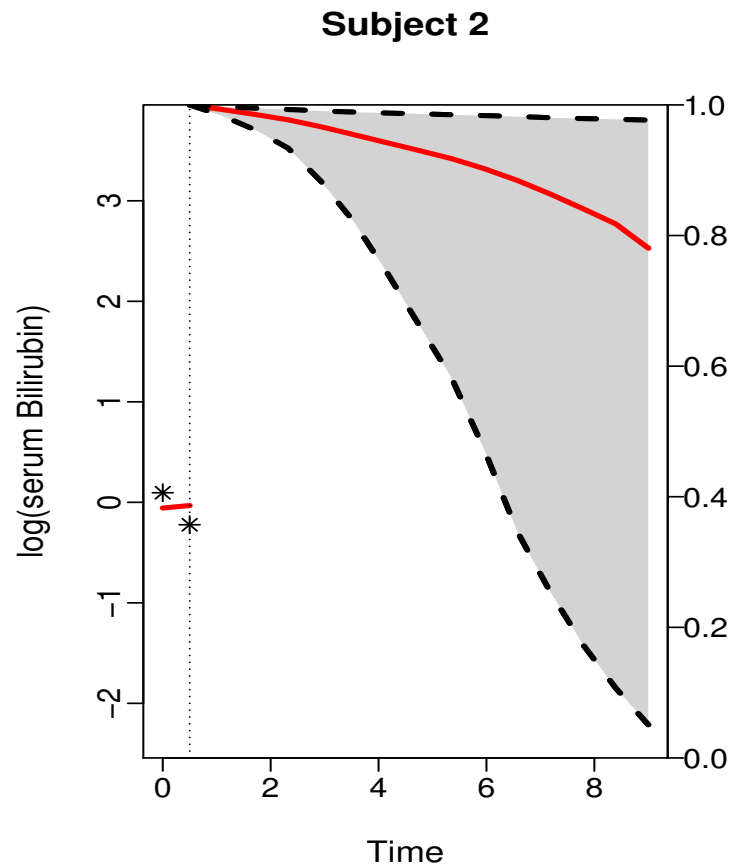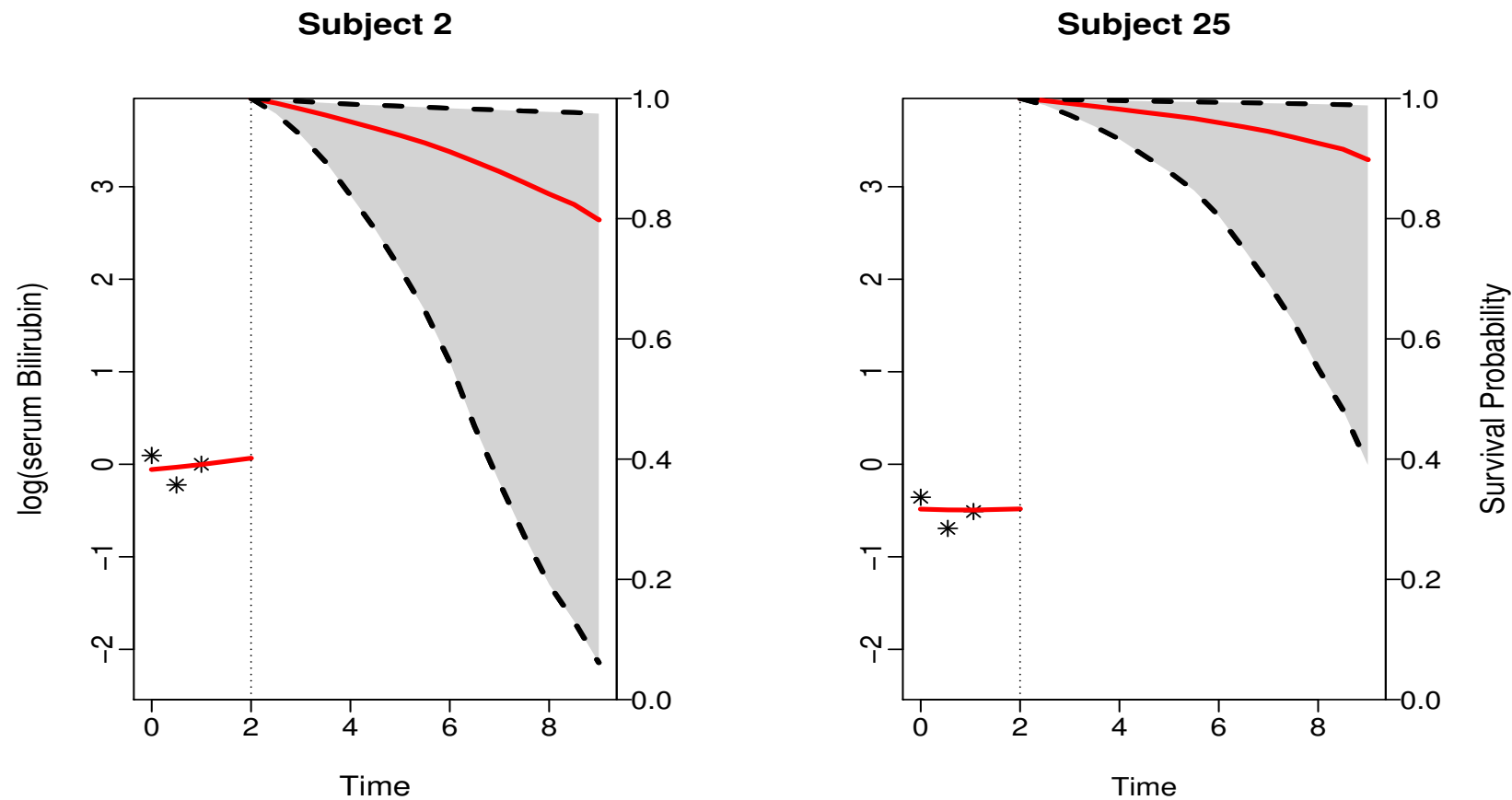
6.2 Dynamic Predictions in R (cont'd)

R> Individualized predictions of survival probabilities are computed by function survfitJM() – for example, for Patient 2 from the PBC dataset we have

```
lmeFit <- lme(log(serBilir) ~ year*drug + I(year^2)*drug,
        random = ~ year + year^2 | id, data = pbc2)

coxFit <- coxph(Surv(years, status2) ~ drug, data = pbc2.id, x = TRUE)

jointFit <- jointModelBayes(lmeFit, coxFit, timeVar = "year",
                            baseHaz = "P-splines")

sfit <- survfitJM(jointFit, newdata = pbc2[pbc2$id %in% c(2, 25), ])

plot(sfit)
plot(sfit, include.y = TRUE)
```

Chapter 7

Closing

7.1 Concluding Remarks

- **When we need joint models for longitudinal and survival outcomes?**

  ▷ to investigate the association between the longitudinal and the survival outcome

  ▷ to handle endogenous time-varying covariates in a survival analysis context

- **How joint models work?**

  ▷ a mixed model for the longitudinal outcome

  ▷ a relative risk model for the event process

  ▷ explain interrelationships with shared random effects

7.1 Concluding Remarks (cont'd)

- **Extensions**

    ▷ under the full conditional independence assumption we can easily extend the basic joint model

    ▷ multiple longitudinal outcomes / different association parameters / time-varying effects

    ▷ though more computationally intensive

- **Where to pay attention when defining joint models?**

    ▷ model flexibly the subject-specific evolutions for the longitudinal outcome

    ▷ use parametric but flexible models for the baseline hazard function

    ▷ consider how to model the association structure between the two processes
    $\Rightarrow$ Parameterization

- **Individualized predictions**

  ▷ joint models can provide subject-specific predictions for the longitudinal and survival outcomes

  ▷ these are dynamically updated as extra information is recorded for the subjects

  ▷ $\Rightarrow$ joint models constitute an excellent tool for personalized medicine

- **What we did not cover**

  ▷ diagnostics for joint models using residuals

  ▷ assessment of predictive performance

  ▷ nonrandom dropout

  ▷ . . .

The End!

7.2 Additional References

- Andrinopoulou, E.R., Eilers, P.H.C., Takkenberg, J.J.M. and Rizopoulos, D. (2017). Improved dynamic predictions from joint models of longitudinal and survival data with time-varying effects using P-splines. *Biometrics*, to appear.

- Andrinopoulou, E.R., Rizopoulos, D., Takkenberg, J.J.M. and Lesaffre, E. (2017). Combined dynamic predictions using joint models of two longitudinal outcomes and competing risk data. *Statistical Methods in Medical Research*, to appear.

- Andrinopoulou, E.R., Rizopoulos, D., Takkenberg, J.J.M. and Lesaffre, E. (2014). Joint modeling of two longitudinal outcomes and competing risk data. *Statistics in Medicine* **33**, 3167–3178.

- Brown, E. and Ibrahim, J. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* **59**, 221–228.

- Brown, E. Ibrahim, J. and DeGruttola, V. (2005). A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics* **61**, 64–73.

- Chi, Y.-Y. and Ibrahim, J. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* **62**, 432–445.

- DeGruttola, V. and Tu, X. (1994). Modeling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics* **50**, 1003–1014.

- Elashoff, R., Li, G. and Li, N. (2008). A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics* **64**, 762–771.

- Faucett, C. and Thomas, D. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statistics in Medicine* **15**, 1663–1685.

- Garcia-Hernandez, A. and Rizopoulos, D. (2017). %JM: A SAS macro to fit jointly generalized mixed models for longitudinal data and time-to-event responses. *Journal of Statistical Software*, to appear.

- Gerds, T. and Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal* **48**, 1029–1040.

- Heagerty, P. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61**, 92–105.

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

7.2 Additional References (cont'd)

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

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

- Lin, H., Turnbull, B., McCulloch, C. and Slate, E. (2002). Latent class models for joint analysis of longitudinal biomarker and event process: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association* **97**, 53–65.

- Liu, L. and 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* **58**, 65–81.

- Mauff, K., Steyerberg, E.W., Nijpels, G., van der Heijden, A. and Rizopoulos, D. (2017). Extension of the association structure in joint models to include weighted cumulative effects. *Statistics in Medicine*, to appear.

- Proust-Lima, C., Joly, P., Dartigues, J. and Jacqmin-Gadda, H. (2009). Joint modelling of multivariate longitudinal outcomes and a time-to-event: A nonlinear latent class approach. *Computational Statistics and Data Analysis* **53**, 1142–1154.

- Proust-Lima, C. and Taylor, J. (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.

- Rizopoulos, D. (2016). The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *Journal of Statistical Software* **72**(7), 1–45.

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

- Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67**, 819–829.

- Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software* **35** (9), 1–33.

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

- Rizopoulos, D., Hatfield, L.A., Carlin, B.P. and Takkenberg, J.J.M. (2014). Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging. *Journal of the American Statistical Association* **109**, 1385–1397.

- Rizopoulos, D. and Lesaffre, E. (2014). Introduction to the special issue on joint modelling techniques. *Statistical Methods in Medical Research* **23**, 3–10.

- Rizopoulos, D., Molenberghs, G. and Lesaffre, E. (2017). Dynamic Predictions with Time-Dependent Covariates in Survival Analysis using Joint Modeling and Landmarking. *Biometrical Journal*, to appear.

- Rizopoulos, D., Taylor, J.M.G., van Rosmalen, J., Steyerberg, E.W. and Takkenberg, J.J.M. (2016). Personalized screening intervals for biomarkers using joint models for longitudinal and survival data. *Biostatistics* **17**, 149–164.

- Rizopoulos, D., Verbeke, G. and Lesaffre, E. (2009). Fully exponential Laplace approximation for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society, Series B* **71**, 637–654.

- Rizopoulos, D., Verbeke, G., Lesaffre, E. and Vanrenterghem, Y. (2008). A two-part joint model for the analysis of survival and longitudinal binary data with excess zeros. *Biometrics* **64**, 611–619.

- Rizopoulos, D., Verbeke, G. and Molenberghs, G. (2010). Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics* **66**, 20–29.

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

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

- Taylor, J., Park, Y., Ankerst, D., Proust-Lima, C., Williams, S., Kestin, L., Bae, K., Pickles, T., and Sandler, H. (2013). Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*, **69**, 206–213.

- Tseng, Y.-K., Hsieh, F. and Wang, J.-L. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika* **92**, 587–603.

- Tsiatis, A. and Davidian, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* **88**, 447–458.

7.2 Additional References (cont'd)

- Tsiatis, A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* **14**, 809–834.

- Tsiatis, A., DeGruttola, V., and Wulfsohn, M. (1995). Modeling the relationship of survival to longitudinal data measured with error: Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* **90**, 27–37.

- Viviani, S., Alfó, M. and Rizopoulos, D. (2014). Generalized linear mixed joint model for longitudinal and survival outcomes. *Statistics and Computing*, **24**, 417–427.

- Viviani, S., Rizopoulos, D. and Alfó, M. (2014). Local sensitivity of shared parameter models to nonignorability of dropout. *Statistical Modelling* **14**, 205–228.

- Wang, Y. and Taylor, J. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association* **96**, 895–905.

- Wu, M. and Bailey, K. (1988). Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine* **7**, 337–346.

- Wu, M. and Bailey, K. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics* **45**, 939–955.

- Wu, M. and Carroll, R. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **44**, 175–188.

- Wulfsohn, M. and Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–339.

- Xu, C., Baines, P. and Wang, J.-L. (2014). Standard error estimation using the EM algorithm for the joint modeling of survival and longitudinal data. *Biostatistics*, to appear.

- Xu, J. and Zeger, S. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics* **50**, 375–387.

- Ye, W., Lin, X., and Taylor, J. (2008). Semiparametric modeling of longitudinal measurements and time-to-event data a two stage regression calibration approach. *Biometrics* **64**, 1238–1246.

- Yu, M., Law, N., Taylor, J., and Sandler, H. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica* **14**, 835–862.

- Yu, M., Taylor, J. and Sandler, H. (2008). Individualized prediction in prostate cancer studies using a joint longitudinal-survival-cure model. *Journal of the American Statistical Association* **108**, 178–187.

- Zeng, D. and Cai, J. (2005). Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *The Annals of Statistics* **33**, 2132–2163.

- Zheng, Y. and Heagerty, P. (2007). Prospective accuracy for longitudinal markers. *Biometrics* **63**, 332–341.

7.3 Medical Papers

- Andrinopoulou, E.R., Rizopoulos, D., Jin, R., Bogers, A., Lesaffre, E. and Takkenberg, J. (2012). An introduction to mixed models and joint modeling: Analysis of valve function over time. *Annals of Thoracic Surgery* **93**, 1765–1772.

- Andrinopoulou, E.R., Rizopoulos, D., Geleijnse, M., Lesaffre, E., Bogers, A. and Takkenberg, J. (2015). Dynamic prediction of outcome for patients with severe aortic stenosis: Application of joint models for longitudinal and time-to-event data. *BMC Cardiovascular Disorders*, to appear.

- Daher Abdi, D.Z., Essig, M., Rizopoulos, D., Le Meur, Y., Premaud, A., Woillard, J.-B., Rerolle, J.-P., Marquet, P. and Rousseau, A. (2013). Impact of longitudinal exposure to mycophenolic acid on acute rejection in renal-transplant recipients using a joint modeling approach. *Pharmacological Research* **72**, 52–60.

- Ibrahim, J., Chu, H. and Chen, L.-M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology* **28**, 2796–2801.

- Nunez, J., Nunez, E., Rizopoulos, D., Minana, G., Bodi, V., Bondanza, L., Husser, O., Merlos, P., Santas, E., Pascual-Figal, D.,; Chorro, F. and Sanchis, J. (2014). Red blood cell distribution width is longitudinally associated with mortality and incident anemia in heart failure patients. *Circulation Journal* **78**, 410—418.

- Rizopoulos, D. and Takkenberg, J. (2014). Tools & Techniques: Dealing with time-varying covariates in survival analysis - joint models versus Cox models. *EuroIntervention* **10**, 285–288.

7.3 Medical Papers (cont'd)

- Thabut, G., Christie, J., Mal, H., Fournier, M., Brugiere, O., Leseche, G., Castier, Y. and Rizopoulos, D. (2013). Survival benefit of lung transplant for cystic fibrosis since lung-allocation-score implementation. *American Journal of Respiratory and Critical Care Medicine* **187**, 1335–1340.

- van der Linde, D., Roos-Hesselink, J., Rizopoulos, D., Heuvelman, H., Budts, W., van Dijk, A., Witsenburg, M., Yap, S., Bogers, A., Oxenius, A., Silversides, C., Oechslin, E. and Takkenberg, J. (2013). Surgical outcome of discrete subaortic stenosis in adults: A multicenter study. *Circulation* **127**, 1184–1191.

- van der Linde, D., Takkenberg, J., Rizopoulos, D., Heuvelman, H., Budts, W., van Dijk, A., Witsenburg, M., Yap, S., Bogers, A., Oxenius, A., Silversides, C., Oechslin, E. and Roos-Hesselink, J. (2013). Natural history of discrete subaortic stenosis in adults: A multicenter study. *European Heart Journal* **34**, 1548–1556.