Joint models for longitudinal and survival data
What they are and when to use them

Dimitris Rizopoulos
Department of Biostatistics, Erasmus Medical Center, the Netherlands

d.rizopoulos@erasmusmc.nl

Annual J&J Quantitative Sciences Statistics Conference

November 9th, 2016

1.1 Introduction

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

- Explicit outcomes
  - ▷ multiple longitudinal responses (e.g., markers, blood values)
  - ▷ time-to-event(s) of particular interest (e.g., death, relapse)

- Implicit outcomes
  - ▷ missing data (e.g., dropout, intermittent missingness)
  - ▷ random visit times

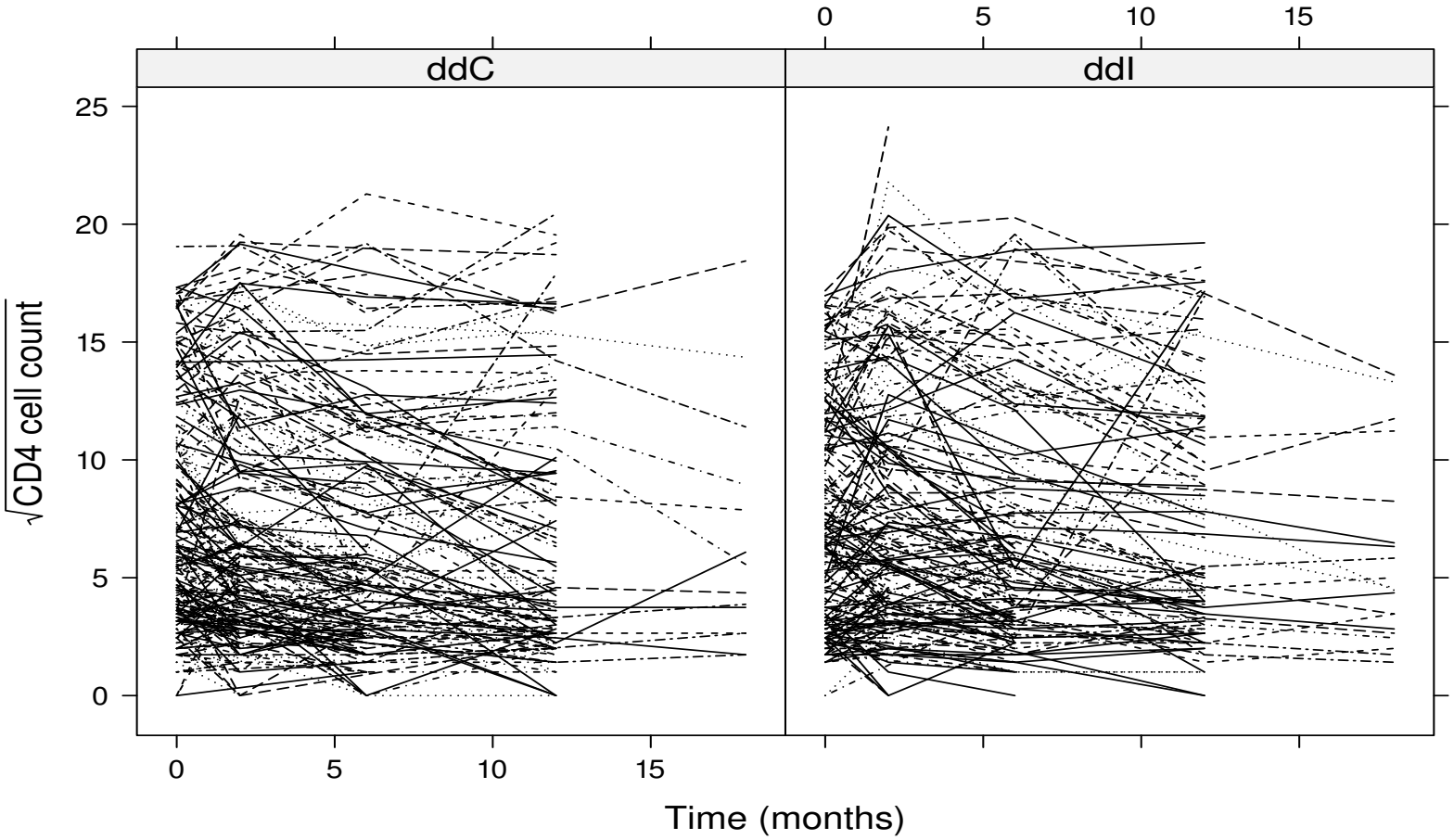
1.2 Illustrative Case Study

- 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)

- Outcomes of interest:
  - ▷ time to death
  - ▷ randomized treatment: 230 patients ddI and 237 ddC
  - ▷ CD4 cell count measurements at baseline, 2, 6, 12 and 18 months
  - ▷ prevOI: previous opportunistic infections

1.2 Illustrative Case Study (cont'd)



1.3 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 males and females?
  - ▷ ...

1.3 Research Questions (cont'd)

- Focus on multiple outcomes
  - ▷ Complex effect estimation: how strong is the association between the longitudinal **outcome** and the hazard rate of death?
  - ▷ Handling implicit outcomes: focus on the longitudinal outcome but with **dropout**

1.3 Research Questions (cont'd)

In the AIDS dataset:

- Research Question:
  - ▷ Investigate the longitudinal evolutions of CD4 cell count correcting for dropout
  - ▷ Can we utilize CD4 cell counts to predict survival

1.4 Goals

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

- Survival data:
  - ▷ Cox model, accelerated failure time models, ...

- Longitudinal data
  - ▷ mixed effects models, GEE, marginal models, ...

1.4 Goals (cont'd)

- **Goals** of this talk:
  - ▷ introduce joint models
  - ▷ link with missing data
  - ▷ sensitivity analysis

2.1 Missing Data in Longitudinal Studies

- A major challenge for the analysis of longitudinal data is the problem of missing data
  - ▷ studies are designed to collect data on every subject at a set of pre-specified follow-up times
  - ▷ often subjects miss some of their planned measurements for a variety of reasons
- We can have different patterns of missing data

Missing Data in Longitudinal Studies

Subject	Visits				
	1	2	3	4	5
1	x	x	x	x	x
2	x	x	x	?	?
3	?	x	x	x	x
4	?	x	?	x	?

- ▷ Subject 1: Completer
- ▷ Subject 2: dropout
- ▷ Subject 3: late entry
- ▷ Subject 4: intermittent

2.1 Missing Data in Longitudinal Studies (cont'd)

- Implications of missingness:
  - ▷ we collect less data than originally planned ⇒ *loss of efficiency*
  - ▷ not all subjects have the same number of measurements ⇒ *unbalanced datasets*
  - ▷ missingness may depend on outcome ⇒ *potential bias*

- For the handling of missing data, we introduce the missing data indicator

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

2.1 Missing Data in Longitudinal Studies (cont'd)

- We obtain a partition of the complete response vector y_i
  - ▷ observed data y_i^o, containing those y_{ij} for which $r_{ij} = 1$
  - ▷ missing data y_i^m, containing those y_{ij} for which $r_{ij} = 0$

- **For the remaining we will focus on dropout** $\Rightarrow$ notation can be simplified
  - ▷ Discrete dropout time: $r_i^d = 1 + \sum_{j=1}^{n_i} r_{ij}$ (ordinal variable)
  - ▷ **Continuous time**: T_i^* denotes the time to dropout

2.2 Missing Data Mechanisms

- To describe the probabilistic relation between the measurement and missingness processes Rubin (1976, Biometrika) has introduced three mechanisms

- *Missing Completely At Random (MCAR)*: The probability that responses are missing is unrelated to both y_i^o and y_i^m

$$p(r_i \mid y_i^o, y_i^m) = p(r_i)$$

- Examples
  - ▷ subjects go out of the study after providing a pre-determined number of measurements
  - ▷ laboratory measurements are lost due to equipment malfunction

2.2 Missing Data Mechanisms (cont'd)

- Features of MCAR:
  - ▷ The observed data y_i^o can be considered a random sample of the complete data y_i
  - ▷ We can use any statistical procedure that is valid for complete data
    - * sample averages per time point
    - * linear regression, ignoring the correlation (consistent, but not efficient)
    - * t-test at the last time point
    - * ...

2.2 Missing Data Mechanisms (cont'd)

- *Missing At Random (MAR)*: The probability that responses are missing is related to y_i^o, but is unrelated to y_i^m

$$p(r_i \mid y_i^o, y_i^m) = p(r_i \mid y_i^o)$$

- Examples
  - ▷ study protocol requires patients whose response value exceeds a threshold to be removed from the study
  - ▷ physicians give rescue medication to patients who do not respond to treatment

2.2 Missing Data Mechanisms (cont'd)

- Features of MAR:
  - ▷ The observed data cannot be considered a random sample from the target population
  - ▷ Not all statistical procedures provide valid results

Not valid under MAR	Valid under MAR
sample marginal evolutions	sample subject-specific evolutions
methods based on moments, such as GEE	likelihood based inference
mixed models with misspecified correlation structure	mixed models with correctly specified correlation structure
marginal residuals	subject-specific residuals

2.2 Missing Data Mechanisms (cont'd)

- *Missing Not At Random (MNAR)*: The probability that responses are missing is related to y_i^m, and possibly also to y_i^o

$$p(r_i \mid y_i^m) \quad \text{or} \quad p(r_i \mid y_i^o, y_i^m)$$

- Examples
  - ▷ in studies on drug addicts, people who return to drugs are less likely than others to report their status
  - ▷ in longitudinal studies for quality-of-life, patients may fail to complete the questionnaire at occasions when their quality-of-life is compromised

2.2 Missing Data Mechanisms (cont'd)

- Features of MNAR
  - ▷ The observed data cannot be considered a random sample from the target population
  - ▷ Only procedures that explicitly model the joint distribution $\{y_i^o, y_i^m, r_i\}$ provide valid inferences $\Rightarrow$ **analyses which are valid under MAR will not be valid under MNAR**

2.2 Missing Data Mechanisms (cont'd)

We cannot tell from the data at hand whether the missing data mechanism is MAR or MNAR

Note: We can distinguish between MCAR and MAR

3.1 Joint Modeling Framework

- To account for possible MNAR dropout, we need to postulate a model that relates
  - ▷ the CD4 cell count, with
  - ▷ the time to dropout

Joint Models for Longitudinal and Time-to-Event Data

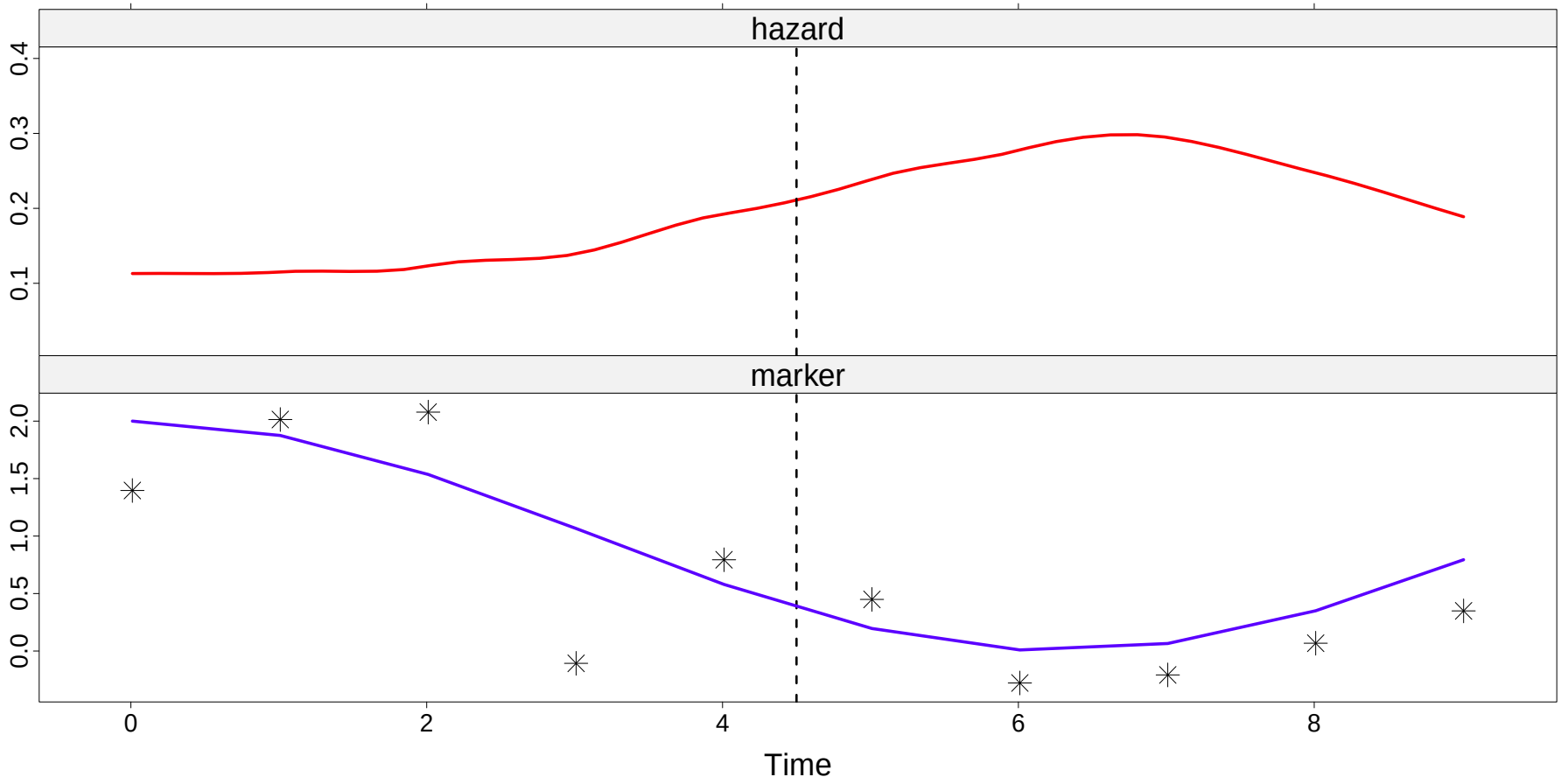
- Intuitive idea behind these models
  1. use an appropriate model to describe the evolution of the marker in time for each patient
  2. the estimated evolutions are then used in a Cox model

3.1 Joint Modeling Framework (cont'd)

- Some notation
  - ▷ y_i: Longitudinal responses
  - ▷ T_i: Dropout time for patient i
  - ▷ δ_i: Dropout indicator, i.e., equals 1 for MNAR events

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

3.1 Joint Modeling Framework (cont'd)



3.1 Joint Modeling Framework (cont'd)

- We define a standard joint model

  ▷ Survival Part: Relative risk model

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

  where
  * $m_i(t)$ = underlying CD4 cell count at time t
  * α quantifies how strongly associated CD4 cell count with the risk of dropping out
  * w_i baseline covariates

3.1 Joint Modeling Framework (cont'd)

▷ Longitudinal Part: Reconstruct $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ using $y_i(t)$ and a mixed effects model (we focus on continuous markers)

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

where

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

3.1 Joint Modeling Framework (cont'd)

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

- Joint Models 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

  ▷ b_i a vector of random effects that explains the interdependencies

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

3.2 Link with Missing Data Mechanisms

- To show this connection more clearly
  - ▷ T_i^*: true time-to-event
  - ▷ y_i^o: longitudinal measurements before T_i^*
  - ▷ y_i^m: longitudinal measurements after T_i^*

- **Important to realize** that the model we postulate for the longitudinal responses is for the complete vector $\{y_i^o, y_i^m\}$
  - ▷ implicit assumptions about missingness

3.2 Link with Missing Data Mechanisms (cont'd)

- Missing data mechanism:

$$p(T_i^* \mid y_i^o, y_i^m) = \int p(T_i^* \mid b_i)\, p(b_i \mid y_i^o, y_i^m)\, db_i$$

still depends on y_i^m, which corresponds to nonrandom dropout

Intuitive interpretation: Patients who dropout show different longitudinal evolutions than patients who do not

3.3 Link with Missing Data Mechanisms (cont'd)

- What about censoring?
  - ▷ censoring also corresponds to a discontinuation of the data collection process for the longitudinal outcome

- Likelihood-based inferences for joint models provide valid inferences when censoring is MAR
  - ▷ a patient relocates to another country (MCAR)
  - ▷ a patient is excluded from the study when her longitudinal response exceeds a pre-specified threshold (MAR)
  - ▷ censoring depends on random effects (MNAR)

3.3 Link with Missing Data Mechanisms (cont'd)

- Joint models belong to the class of *Shared Parameter Models*

$$p(y_i^o, y_i^m, T_i^*) = \int p(y_i^o, y_i^m \mid b_i)\, p(T_i^* \mid b_i)\, p(b_i) db_i$$

the association between the longitudinal and missingness processes is explained by the *shared* random effects b_i

3.3 Link with Missing Data Mechanisms (cont'd)

- The other two well-known frameworks for MNAR data are
  - ▷ Selection models

$$p(y_i^o, y_i^m, T_i^*) = p(y_i^o, y_i^m)\, p(T_i^* \mid y_i^o, y_i^m)$$

  - ▷ Pattern mixture models:

$$p(y_i^o, y_i^m, T_i^*) = p(y_i^o, y_i^m \mid T_i^*)\, p(T_i^*)$$

- These two model families are primarily applied with discrete dropout times and cannot be easily extended to continuous time

3.4 MNAR Analysis of the AIDS data

- Example: In the AIDS dataset
  - ▷ 58 (5%) completers
  - ▷ 184 (39%) died before completing the study
  - ▷ 225 (48%) dropped out before completing the study

- A comparison between
  - ▷ linear mixed-effects model $\Rightarrow$ all dropout MAR
  - ▷ joint model $\Rightarrow$ death is set MNAR, and dropout MAR

  is warranted

3.4 MNAR Analysis of the AIDS data (cont'd)

- We fitted the following joint model

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

  where

  ▷ $h_0(t)$ is assumed piecewise-constant

- The MAR analysis entails only the linear mixed model

3.4 MNAR Analysis of the AIDS data

	LMM (MAR)	JM (MNAR)
	value (s.e.)	value (s.e)
Intercept	7.19 (0.22)	7.20 (0.22)
Time	−0.16 (0.02)	−0.23 (0.04)
Treat:Time	0.03 (0.03)	0.01 (0.06)

- ▷ We observe some sensitivity for the time effect
- ▷ The interaction with treatment remains non significant under both analyses

3.5 CD4 and the risk of of death

- Turn focus on the link between CD4 cell count and risk of death

	JM	Cox
	log HR (std.err)	log HR (std.err)
Treat	0.33 (0.16)	0.31 (0.15)
$CD4^{1/2}$	−0.29 (0.04)	−0.19 (0.02)

3.5 CD4 and the risk of of death (cont'd)

- A unit decrease in $CD4^{1/2}$, results in a
  - ▷ **Joint Model**: 1.3-fold increase in risk (95% CI: 1.24; 1.43)
  - ▷ **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

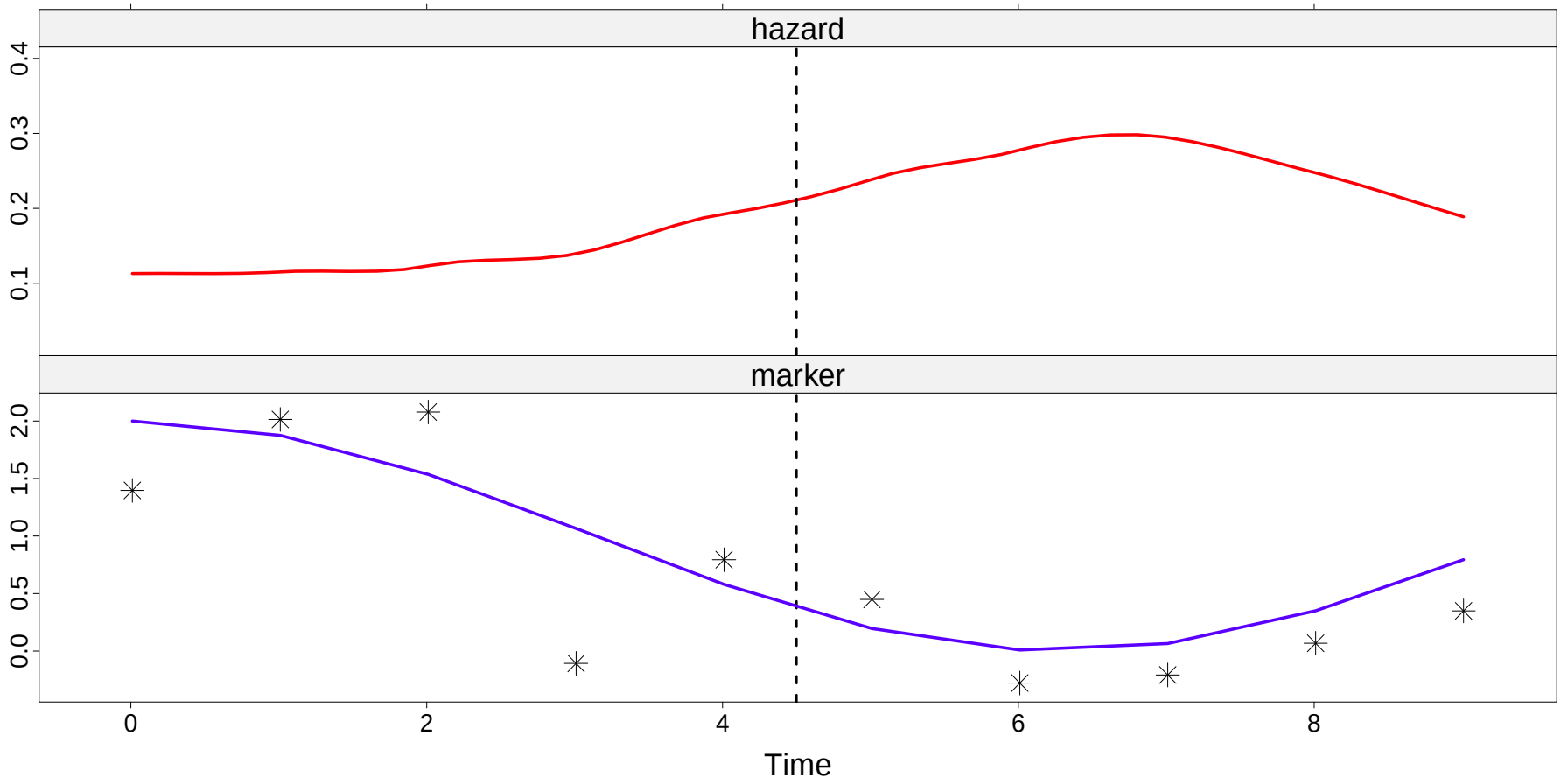
4.1 Association Structures

- The standard assumption is

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

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

4.1 Association structures (cont'd)



4.1 Association Structures (cont'd)

- The standard assumption is

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

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

Is this the only option? Is this the most optimal for prediction?

4.1 Association Structures (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., patient 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

4.2 Time-dependent Slopes

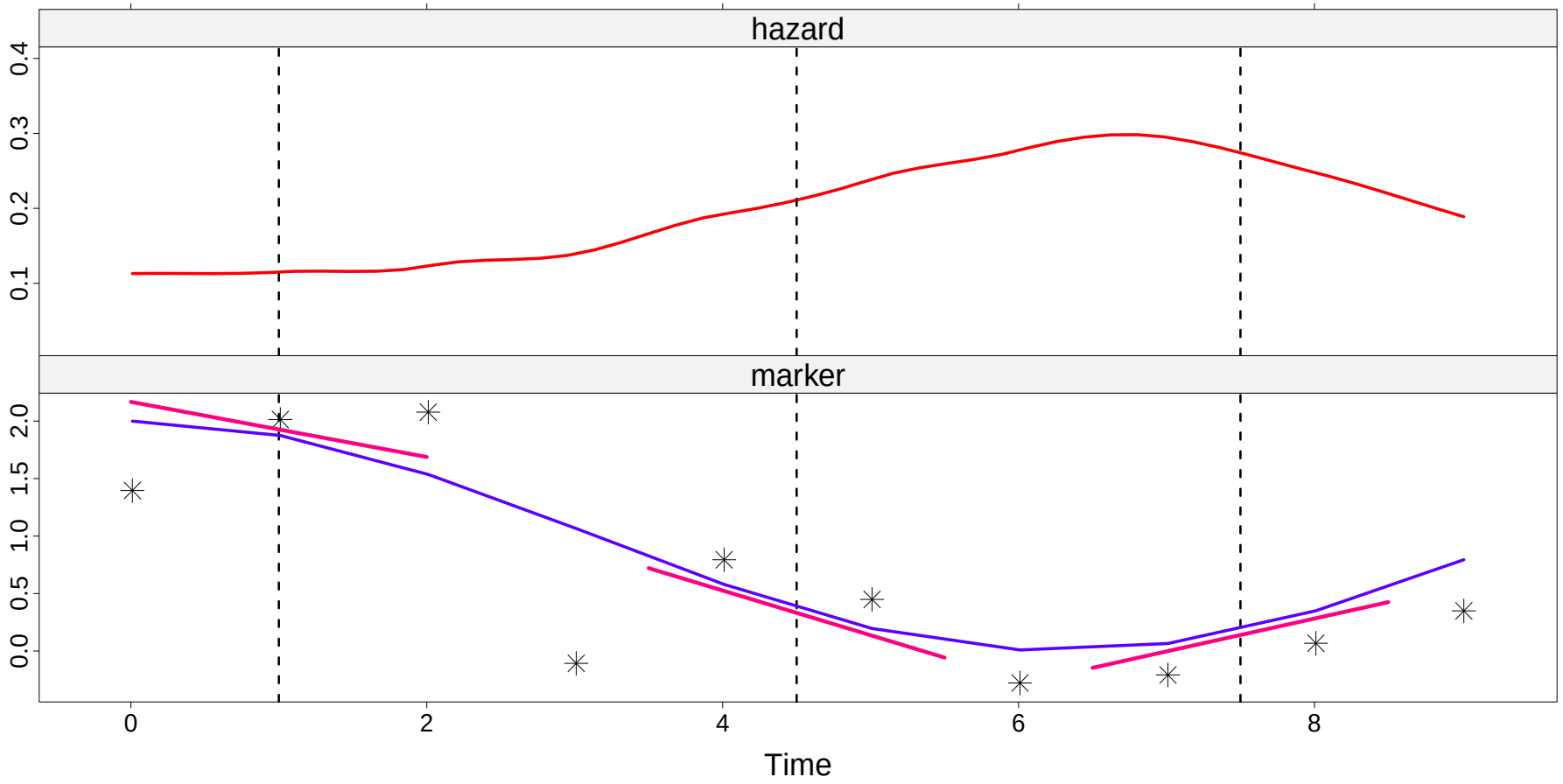
- The hazard for an event at t is associated with both the current value and the slope of the trajectory at t (Ye et al., 2008, Biometrics):

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

where

$$m_i'(t) = \frac{d}{dt}\{x_i^\top(t)\beta + z_i^\top(t) b_i\}$$

4.2 Time-dependent Slopes (cont'd)



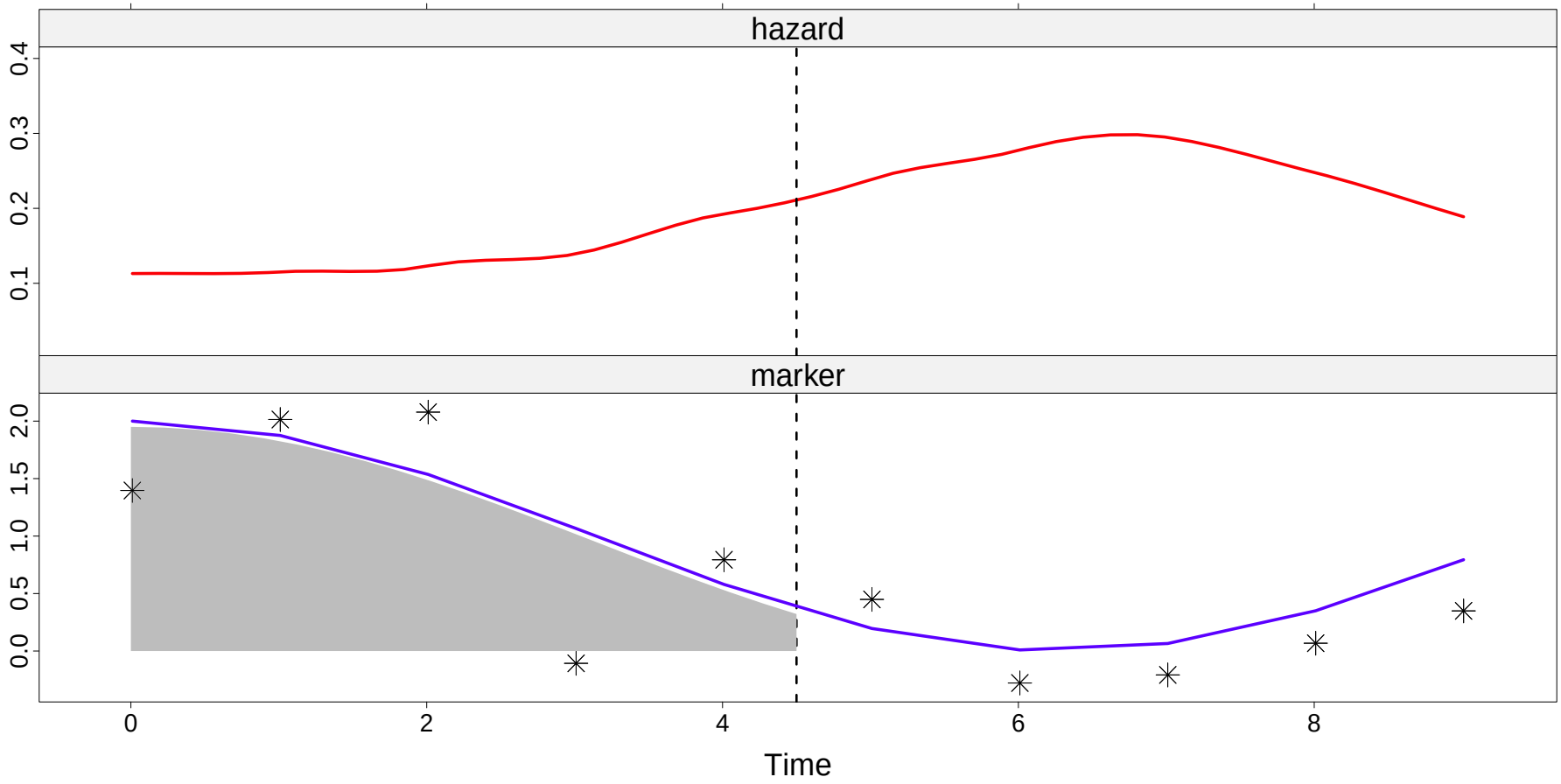
4.3 Cumulative Effects

- The hazard for an event at t is associated with 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 $\mathcal{M}_i(t)$

4.3 Cumulative Effects (cont'd)



4.4 Weighted Cumulative Effects

- The hazard for an event at t is associated with the area under the weighted trajectory up to t:

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

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

- ▷ Gaussian density
- ▷ Student's-t density
- ▷ ...

4.5 Parameterizations & Sensitivity Analysis

- 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.,

$$\begin{aligned} y_i(t) &= m_i(t) + \varepsilon_i(t) \\ &= \beta_0 + \beta_1 t + \beta_2 \{t \times \texttt{ddI}_i\} + b_{i0} + b_{i1} t + \varepsilon_i(t) \end{aligned}$$

  and the following four survival submodels

4.5 Parameterizations & Sens. Analysis (cont'd)

- Model I (current value)

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

- Model II (current value + current slope)

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

where

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

4.5 Parameterizations & Sens. Analysis (cont'd)

- Model III (random slope)

$$h_i(t) \;=\; h_0(t)\exp\{\gamma \texttt{ddI}_i + \alpha_3 b_{i1}\}$$

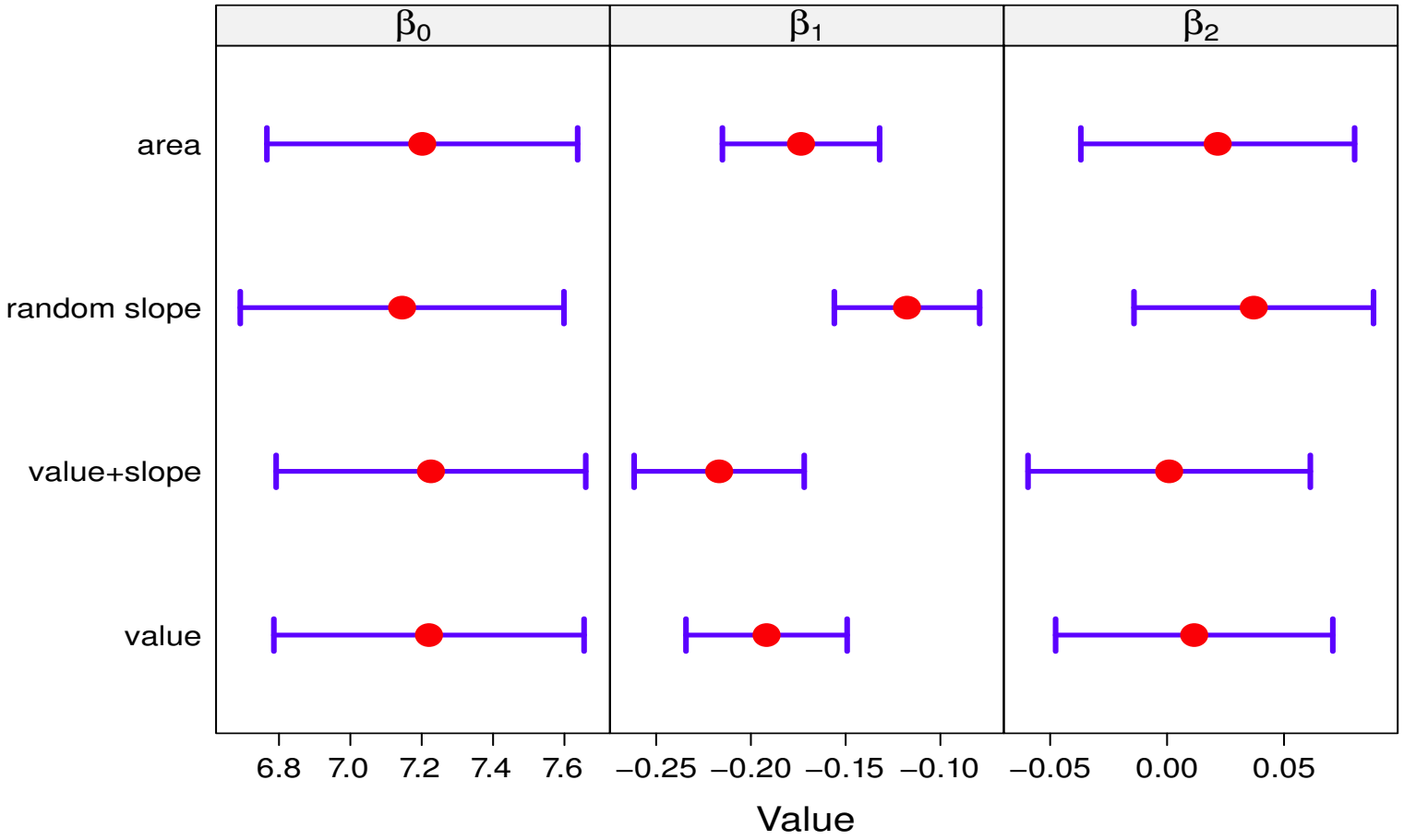
- Model IV (area)

$$h_i(t) \;=\; h_0(t)\exp\Big\{\gamma \texttt{ddI}_i + \alpha_4 \int_0^t m_i(s)\; ds\Big\},$$

where

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

4.5 Parameterizations & Sens. Analysis (cont'd)



5. Software

R> Joint models are fitted using function `jointModel()` from package **JM**. 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 <- jointModel(lmeFit, coxFit, timeVar = "obstime",
    method = "piecewise-PH-aGH")

summary(jointFit)
```

5. Software (cont'd)

R> 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

5. Software (cont'd)

R> Argument `method` specifies the type of relative risk model and the type of numerical integration algorithm – the syntax is as follows:

`<baseline hazard>-<parameterization>-<numerical integration>`

Available options are:

- `"piecewise-PH-GH"`: PH model with piecewise-constant baseline hazard
- `"spline-PH-GH"`: PH model with B-spline-approximated log baseline hazard
- `"weibull-PH-GH"`: PH model with Weibull baseline hazard
- `"weibull-AFT-GH"`: AFT model with Weibull baseline hazard
- `"Cox-PH-GH"`: PH model with unspecified baseline hazard

`GH` stands for standard Gauss-Hermite; using `aGH` invokes the pseudo-adaptive Gauss-Hermite rule

5. Software (cont'd)

- Software: R package **JM** freely available via
  http://cran.r-project.org/package=JM
  - ▷ it can fit a variety of joint models + many other features

- More info available at:

> Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data, with Applications in R*. Boca Raton: Chapman & Hall/CRC.
>
> Web site: http://jmr.r-forge.r-project.org/

5. Software (cont'd)

- Software: R package **JMbayes** freely available via
  http://cran.r-project.org/package=JMbayes
  - ▷ it can fit a variety of multivariate joint models + many other features

GUI interface for dynamic predictions using package **shiny**

5. Software (cont'd)

- SAS macro %JM by Alberto Garcia-Hernandez & D. Rizopoulos
  http://www.jm-macro.com/

Thank you for your attention!