Joint Modelling of Longitudinal and Survival Data

Rui Martins

CENTRO
DE INVESTIGAÇÃO
INTERDISCIPLINAR
EGAS MONIZ

Summary

Data in longitudinal studies

- Multiple biomarkers, *e.g.* blood pressure or CD4 counts, are often collected repeatedly over time (longitudinal data)
- time to an event of interest, *e.g.* death from any cause (survival data)
- Examples
  - PSA repeated measures and time to a recurrence of prostate cancer
  - CD4 repeated measures and time to AIDS

Questions of interest

- Separate Analysis
  - does treatment affect survival?
  - are the average longitudinal evolutions different between males and females?
- Joint Analysis
  - what is the effect of the longitudinal evolution of CD4 cell count in the hazard rate for death?
  - how the association between markers evolves over time (evolution of the association)
  - how marker-specific evolutions are related to each other (association of the evolutions)

Questions of interest

- Separate Analysis
  - does treatment affect survival?
  - are the average longitudinal evolutions different between males and females?

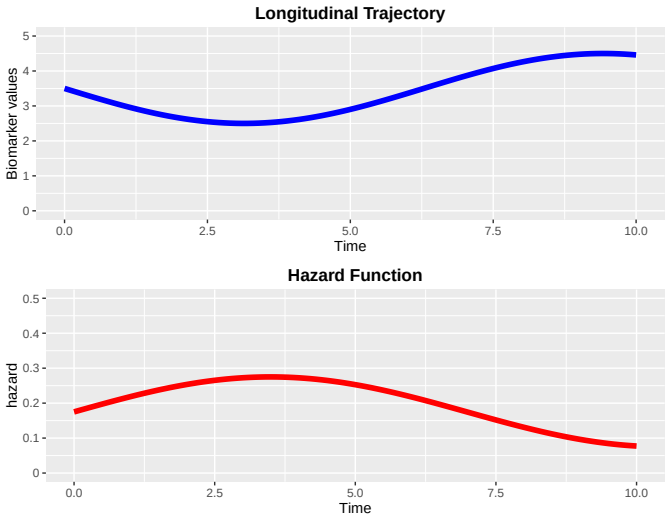
- Joint Analysis
  - what is the effect of the longitudinal evolution of CD4 cell count in the hazard rate for death?
  - how the association between markers evolves over time (evolution of the association)
  - how marker-specific evolutions are related to each other (association of the evolutions)

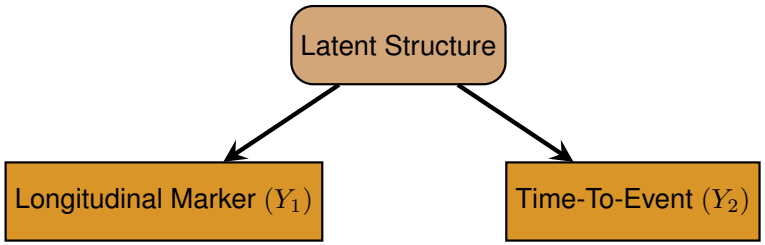
Issues

- Longitudinal studies are often affected by (informative) drop-out, e.g. due to death: MCAR, MAR, MNAR
- Biomarkers are often measured with error
- Survival analysis assumes that covariates are measured without error: Internal *vs* External covariates

Principle

Simultaneous modelling of correlated longitudinal and survival data





If the two processes are associated $\Rightarrow$ define a model for their joint probability distribution: $f(y_1, y_2)$

Objectives of a joint analysis

- explore the association between the two processes
- describe the longitudinal process stopped by the event
- **predict the risk of event adjusted for the longitudinal process**

Applications

- Arose primarily in the field of AIDS, relating CD4 trajectories to progression to AIDS in HIV$^+$ patients (Faucett and Thomas, 1996)
- Further developed in cancer, particularly modelling PSA levels and their association with prostate cancer recurrence (Proust-Lima and Taylor, 2009)

Joint Models to Analyse Longitudinal and Survival Data

Statistical Models

- Longitudinal model
  - mixed effects model
  - splines, etc.
  - multiple markers of progression and/or different nature
  - Gaussian, binary, Poisson
  - continuous but non Gaussian
- Survival model
  - Relative risk model (Proportional hazard model)
  - Accelerated failure time model
  - Competing risks; recurrent events; multiple events
- Linking structure
  - depends on the purposes
  - without consensus
  - *still evolving* ...

Statistical Models

Main families of joint models

- Latent classes (Proust-Lima et al., 2012)
- Shared parameters (Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Gould et al., 2014)
- random-effects models
- Pattern-mixture models
- Selection models

General idea

How to specify the joint distribution, $f(y_1, y_2)$?

- directly
- factorize

$$f(y_1, y_2) = f(y_1|y_2)f(y_2) = f(y_2|y_1)f(y_1)$$

- use latent variables

$$f(y_1, y_2) = \int f(y_1, y_2|\boldsymbol{b})f(\boldsymbol{b})d\boldsymbol{b} = \int f(y_1|\boldsymbol{b})f(y_2|\boldsymbol{b})f(\boldsymbol{b})d\boldsymbol{b}$$

$$f(y_1, y_2) = \int f(y_1, y_2|\boldsymbol{b})f(\boldsymbol{b}_1, \boldsymbol{b}_2)d\boldsymbol{b} = \int f(y_1|\boldsymbol{b}_1)f(y_2|\boldsymbol{b}_2)f(\boldsymbol{b}_1, \boldsymbol{b}_2)d\boldsymbol{b}_1 d\boldsymbol{b}_2$$

Longitudinal submodel

Assumes observations of a normally distributed longitudinal marker for each time,

$$y_i(t)|\boldsymbol{\theta} \sim \mathcal{N}\left(m_i(t), \sigma_e^2\right)$$

- the mean level trajectory (linear mixed-effects model)

$$m_i(t) = \mathbf{x}_i^\top(t)\boldsymbol{\beta}_1 + \mathbf{z}_i^\top(t)\boldsymbol{b}_i$$

- random-effects

$$\boldsymbol{b}_i \sim \mathcal{N}\left(0, \Sigma_b\right)$$

- more flexibility through polynomials or splines in $\mathbf{x}_i$ and $\mathbf{z}_i$.

Survival submodel

Estimates T_i^*, but using only $T_i = \min(T_i^*, C_i)$ and δ_i.
Assumes the following hazard model,

$$h_i(t) = h_0(t) \exp\{\mathbf{v}_i^\top \boldsymbol{\beta}_2 + \gamma \times m_i(t)\}$$

- where $h_0(t)$ is the baseline hazard function (Weibull, piecewise exponential, splines...)
- $\mathbf{v}_i$ is a vector of time-independent baseline covariates with an associated vector of log hazard ratios, $\boldsymbol{\beta}_2$
- $\gamma \times m_i(t)$ represents the linking structure.

How can we link longitudinal and survival data?

- Use the observed baseline biomarker values
  - We're ignoring all the repeated measures and measurement error
- Use the repeated measures as a time-varying covariate
  - We're still ignoring the measurement error
- Model the longitudinal outcome, and use predictions as a time-varying covariate
  - Uncertainty in the longitudinal outcome is not carried through
- Model both processes simultaneously in a joint model, defining a joint probability distribution
  - Reduce bias and maximize efficiency

Linking structure

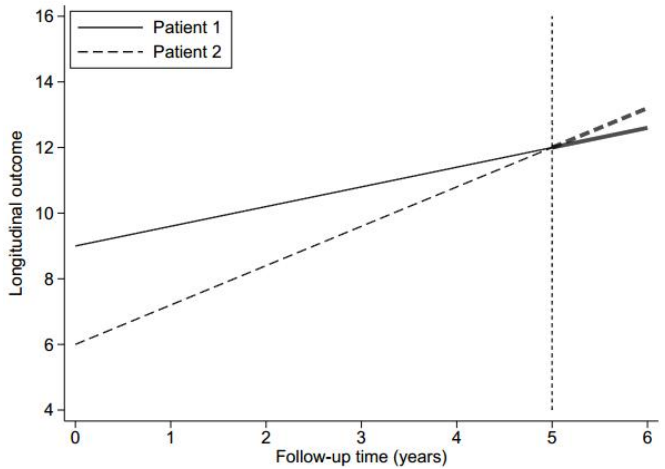
What are the main characteristics of the longitudinal trajectory associated with the Survival?

- **Current value parameterisation**

$$h_i(t) = h_0(t) \exp\{\mathbf{v}_i^\top \boldsymbol{\beta}_2 + \gamma \times m_i(t)\}$$

- **Time-dependent slope**

$$h_i(t) = h_0(t) \exp\{\mathbf{v}_i^\top \boldsymbol{\beta}_2 + \gamma_1 \times m_i(t) + \gamma_2 \times m_i'(t)\}$$



Linking structure II – shared parameters

A time-independent association

Longitudinal submodel

$$m_i(t) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t$$

Survival submodel

$$h_i(t) = h_0(t) \exp\{\mathbf{v}_i^\top \boldsymbol{\beta}_2 + \gamma(b_{0i} + b_{1i})\}$$

$$h_i(t) = h_0(t) \exp\{\mathbf{v}_i^\top \boldsymbol{\beta}_2 + \gamma_1 b_{0i} + \gamma_2 b_{1i}\}$$

random-effects: $\boldsymbol{b}_i = (b_{0i}, b_{1i})$ with density $f(\boldsymbol{b}_i)$; usually a Gaussian

Linking structure III – random-effects

Random-effects parameterisation

$$m_i(t) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t$$

$$h_i(t) = h_0(t) \exp\{\mathbf{v}_i^\top \boldsymbol{\beta}_2 + \gamma(b_{2i})\}$$

random-effects: $\boldsymbol{b}_i = (b_{0i}, b_{1i}, b_{2i})$ with density $f(\boldsymbol{b}_i)$; usually a Gaussian

Joint likelihood

$$L(\boldsymbol{\theta}, \mathbf{b} \mid \mathcal{D}) = \prod_{i=1}^{N} \left(\prod_{j=1}^{n_i} p(y_i(t_{ij}) | \boldsymbol{\theta}, \boldsymbol{b}_i) \right) p(T_i, \delta_i | \boldsymbol{\theta}, \boldsymbol{b}_i)$$

Joint likelihood

$$L(\boldsymbol{\theta}, \mathbf{b} \mid \mathcal{D}) = \prod_{i=1}^{N} \left(\prod_{j=1}^{n_i} p(y_i(t_{ij}) | \boldsymbol{\theta}, \boldsymbol{b}_i) \right) p(T_i, \delta_i | \boldsymbol{\theta}, \boldsymbol{b}_i)$$

where

$$p(y_i(t_{ij}) \mid \boldsymbol{\theta}, \boldsymbol{b}_i) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left\{ - \frac{[y_i(t_{ij}) - m_i(t_{ij})]^2}{2\sigma_e^2} \right\}$$

Joint likelihood

$$L(\boldsymbol{\theta}, \mathbf{b} \mid \mathcal{D}) = \prod_{i=1}^{N} \left(\prod_{j=1}^{n_i} p(y_i(t_{ij}) | \boldsymbol{\theta}, \boldsymbol{b}_i) \right) p(T_i, \delta_i | \boldsymbol{\theta}, \boldsymbol{b}_i)$$

where

$$p(T_i, \delta_i \mid \boldsymbol{\theta}, \boldsymbol{b}_i) = [h_0(T_i) \exp\{\mathbf{v}_i^\top \boldsymbol{\beta}_2 + \gamma \times m_i(T_i)\}]^{\delta_i} \times \exp\left\{ - \int_0^{T_i} h_0(u) \exp\{\mathbf{v}_i^\top \boldsymbol{\beta}_2 + \gamma \times m_i(u)\} du \right\}$$

Prediction

Probably the most "marketable" feature of Joint Models

- A growing interest in a tailored made medical decision
  - Personalized Medicine
  - Shared Decision Making
- This is of high relevance in various diseases
  - cancer research, cardiovascular diseases, HIV research ...
- Bayesian approach

Physicians are interested in accurate prognostic tools that will inform them about the future prospect of a patient to adjust medical care

Dynamic Predictions

Example: HIV/AIDS patients receiving HAART therapy

- Interest in predicting survival probabilities for a new patient i that has provided a set of CD4 measurements up to a specific time point t
- What do we know for the patient?
  - a series of CD4 and/or viral load
  - is event-free up to the last measurement
- General Questions:
  - Can we utilize CD4 or viral load measurements to predict survival?
  - When to plan the next visit for a patient?
- Survival probabilities and the visiting plan can be dynamically updated as additional information is recorded - *Dynamic Predictions* (Rizopoulos 2011)

Predictions (Bayesian)

Suppose a new individual data, $\tilde{\mathcal{D}} = \{\tilde{\mathbf{y}}, \tilde{T} = t, \tilde{\delta} = 0\}$

- **Future longitudinal values** at time $s > t$

$$p(\tilde{y}(s) \mid \mathcal{D}, \tilde{\mathcal{D}}) = \int\int p(\tilde{y}(s)|\tilde{\mathcal{D}}, \tilde{\boldsymbol{b}}, \boldsymbol{\theta})\, p(\tilde{\boldsymbol{b}}|\tilde{\mathcal{D}}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}\, d\tilde{\boldsymbol{b}}$$

- **Future survival probabilities** at time $s > t$

$$p(\tilde{T}^* > s \mid \mathcal{D}, \tilde{T}^* > t, \tilde{\mathbf{y}}) = \int\int \frac{\tilde{S}(s \mid \tilde{\mathbf{y}}, \tilde{\boldsymbol{b}})}{\tilde{S}(t \mid \tilde{\mathbf{y}}, \tilde{\boldsymbol{b}})}\, p(\tilde{\boldsymbol{b}}|\tilde{\mathcal{D}}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}\, d\tilde{\boldsymbol{b}}$$

Software

- R - JM, joineR, frailtypack, INLA
- WinBUGS, Stan, JMBayes
- STATA - stjm command

Computationally intensive

Software

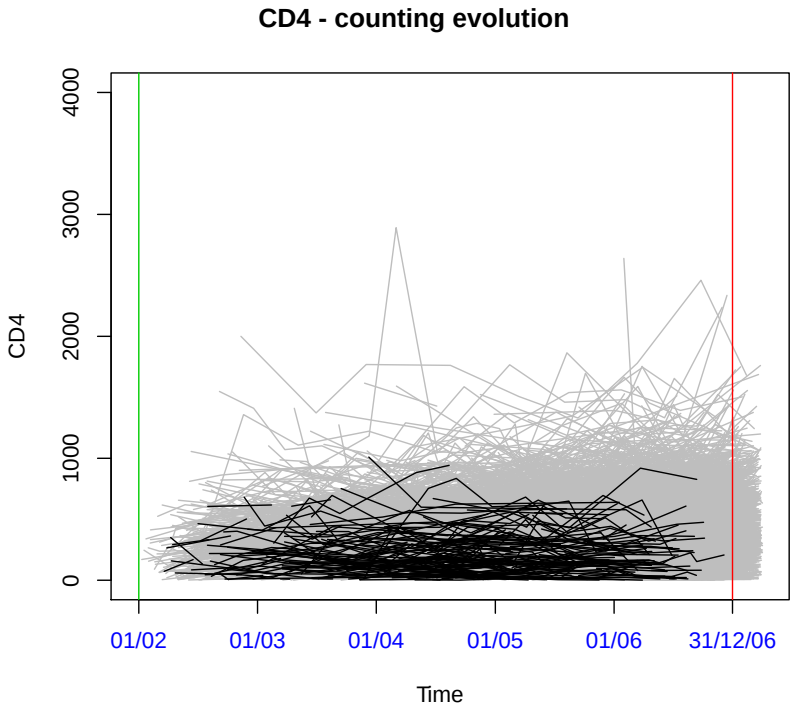
```
jointModel(lmeObject, survObject, timeVar,
  parameterization = c("value", "slope", "both"),
  method = c("weibull-PH-aGH", "weibull-PH-GH", "weibull-AFT-aGH",
    "weibull-AFT-GH", "piecewise-PH-aGH", "piecewise-PH-GH",
    "Cox-PH-aGH", "Cox-PH-GH", "spline-PH-aGH", "spline-PH-GH",
    "ch-Laplace"),
  interFact = NULL, derivForm = NULL, lag = 0, scaleWB = NULL,
  CompRisk = FALSE, init = NULL, control = list(), ...)
```

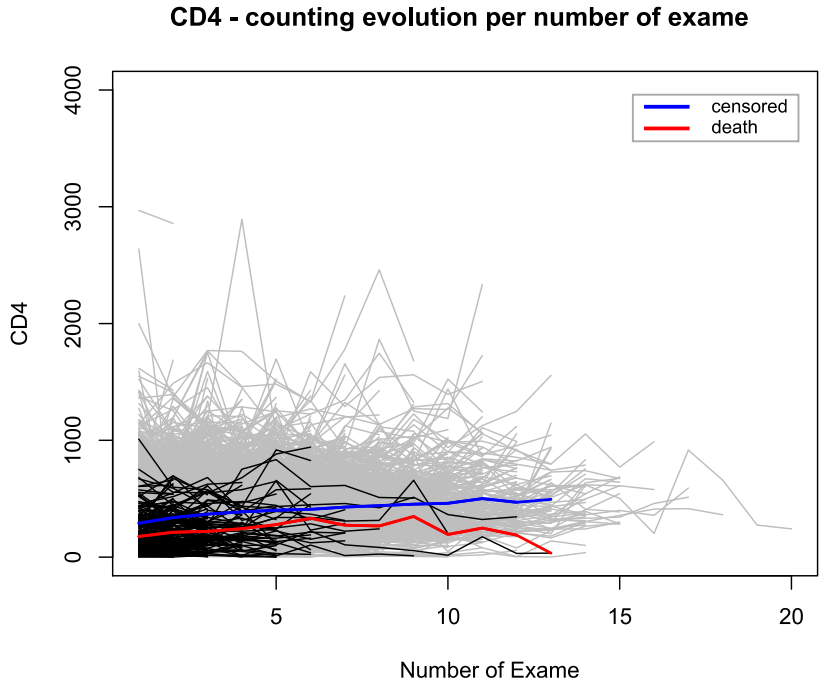
```
jointModelBayes(lmeObject, survObject, timeVar,
    param = c("td-value", "td-extra", "td-both", "shared-betasRE", "shared-RE"),
    extraForm = NULL, baseHaz = c("P-splines", "regression-splines"),
    transFun = NULL, densLong = NULL, lag = 0, df.RE = NULL,
    estimateWeightFun = FALSE, weightFun = NULL, init = NULL,
    priors = NULL, scales = NULL, control = list(), ...)
```

Database - Martins, Silva, Andreozzi (2016) Stat. Med.

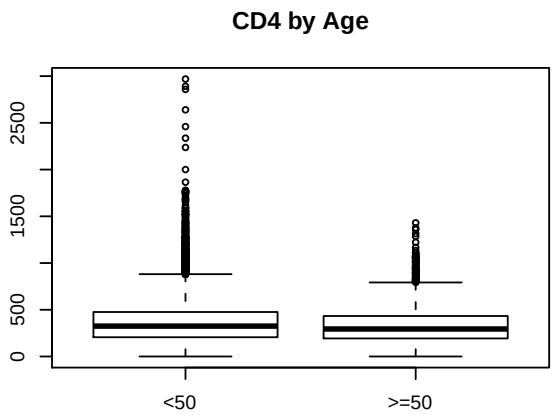
- network of 88 laboratories located in every state in Brazil during 2002–2006;
- **Sample**: $n = 4654$ individuals;
- **Outcomes**: CD4$^+$T lymphocyte counts and survival time;
- **Covariates**: age (<50=0, $\geq$50=1); gender (Female=0, Male=1); prevoi (previous opportunistic infection at study entry=1, no previous infection=0); region (the 27 states of Brazil); time;
- **Patients**: 320 deaths. 88% between 15 and 49 years old; 60% males. 61% no previous infection. Initial CD4 median: 245 cells/mm^3 (men - 226 cells/mm^3; women - 263 cells/mm^3).

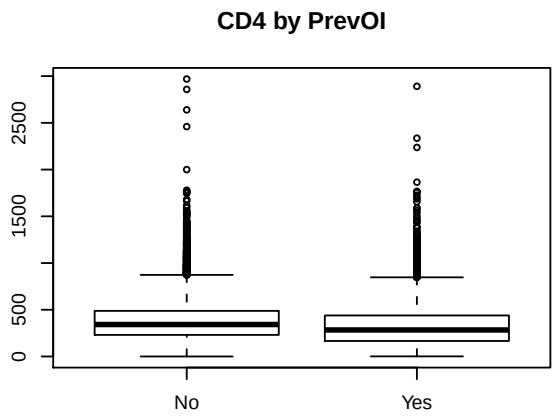
Exploratory trajectory plots

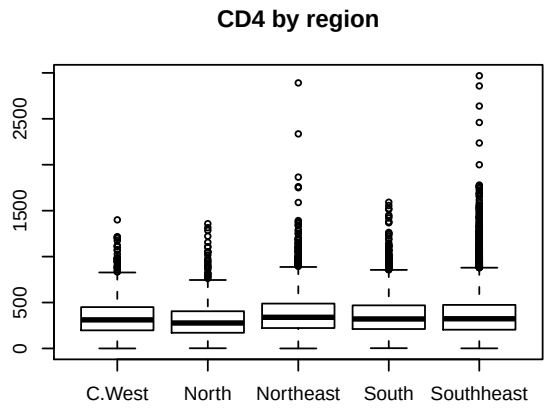


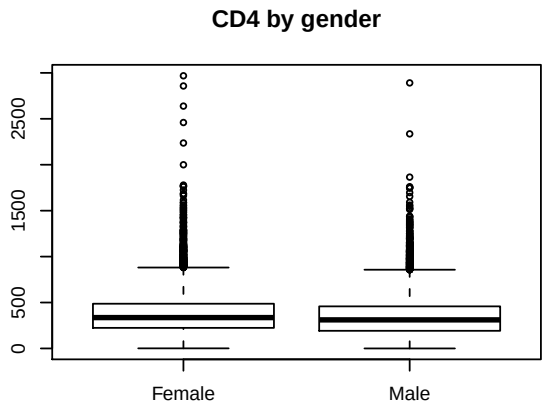


CD4 transformation









Adjusted joint model

Longitudinal specification

$$\sqrt{\text{CD4}} \mid \boldsymbol{b}_{ik}, \boldsymbol{\beta}_1, \sigma_e^2 \sim \mathcal{N}(m_{ikj}, \sigma^2)$$

$$\begin{aligned} m_{ikj} = {} & \beta_{11} + \beta_{12} t_{ikj} + \beta_{13} t_{ikj}^2 + \beta_{14} t_{ikj}^3 + \\ & b_{1ik} + b_{2ik} t_{ikj} + b_{3ik} t_{ikj}^2 + b_{4ik} t_{ikj}^3 + \\ & \beta_{15} \text{gender}_{ik} + \beta_{16} \text{age}_{ik} + \beta_{17} \text{PrevOI}_{ik} \end{aligned}$$

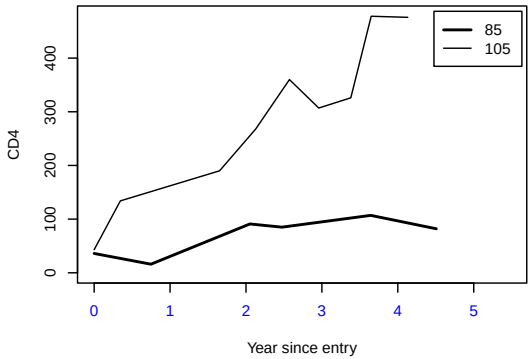
Survival specification

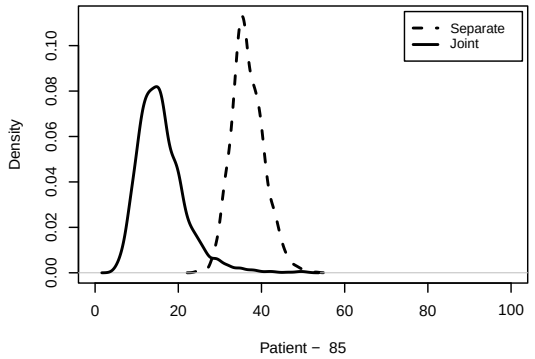
$$T_{ik} \mid \boldsymbol{b}_{ik}, \boldsymbol{\beta}_2, Q_k \sim \mathcal{W}(1, \lambda_{ik}(t)) \equiv \mathcal{E}(\lambda_{ik}(t))$$

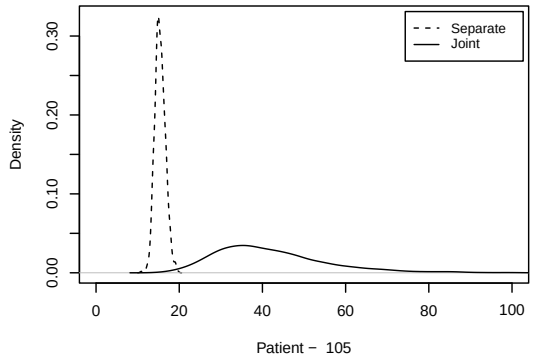
$$\lambda_{ik}(t) = \exp\{\beta_{21} + \beta_{22}\text{gender}_{ik} + \beta_{23}\text{age}_{ik} + \beta_{24}\text{PrevOI}_{ik} + \sum_{s=1}^{4} \gamma_s b_{sik} + Q_k\}$$

$$Q_k | \sigma_Q^2 \sim ICAR(\sigma_Q^2), \quad k = 1, \ldots, 27$$

Separate *vs* Joint



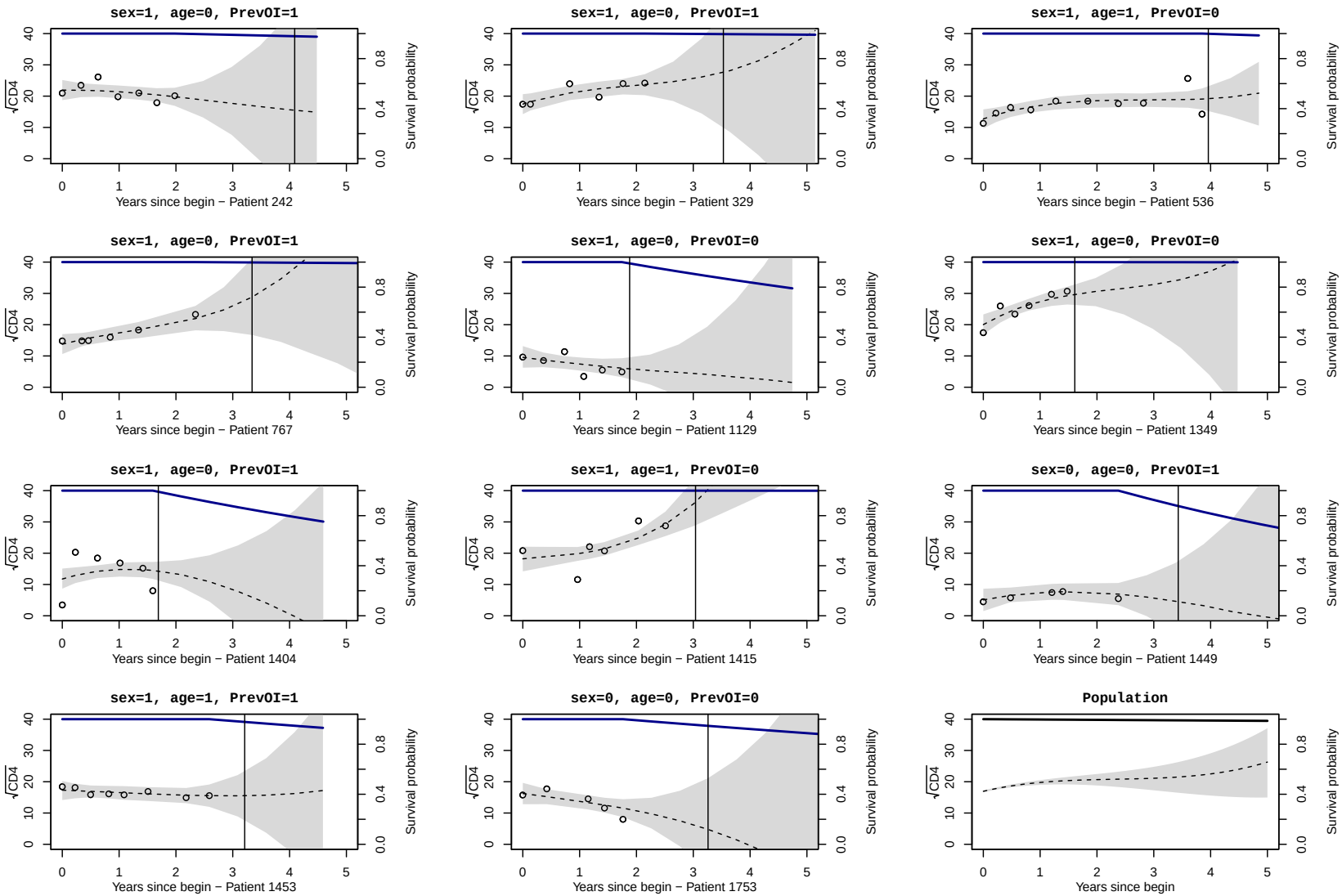




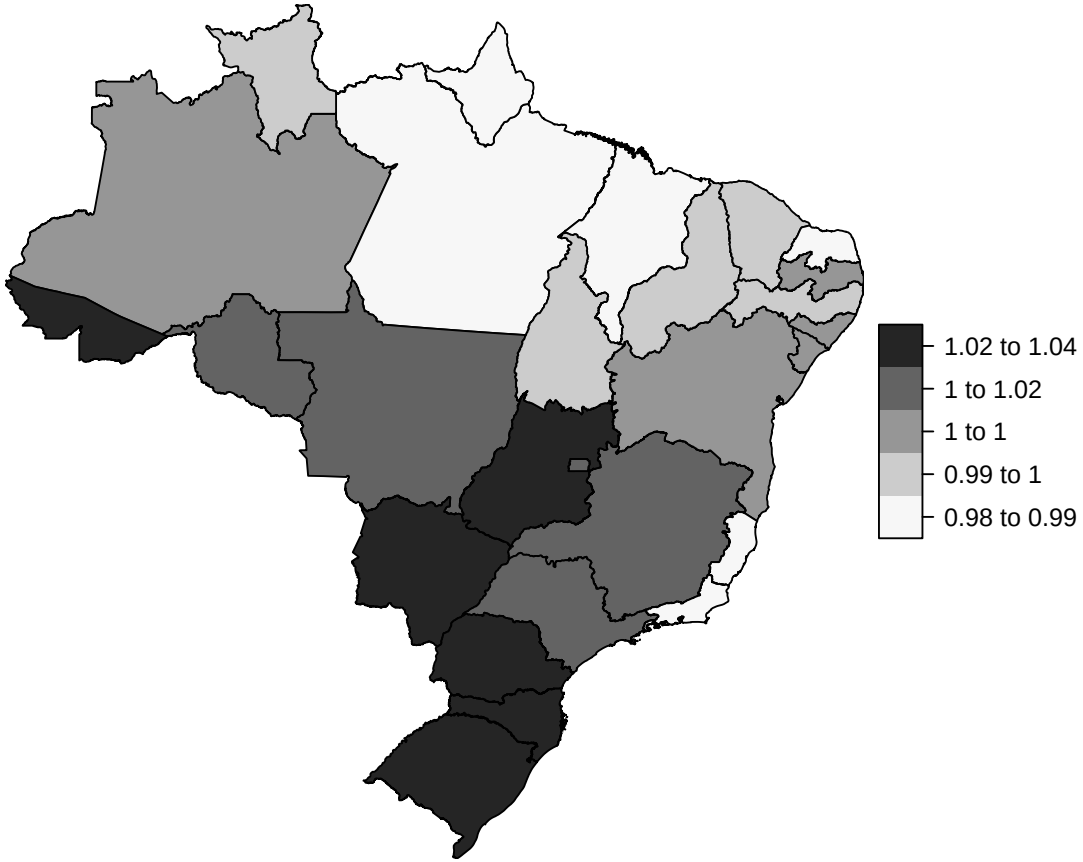
- two patients:
  - male, 31 years old, without previous opportunistic infection and censored time 1645 days
  - male, 29 years old, with previous opportunistic infection and censored time 1508 days

- **Posterior median survival time**: joint model improves the survival estimates

Predictions



Spatial relative risk



FAUCETT, C. L., AND THOMAS, D. C.
Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach.
Statistics in Medicine 15, 15 (Aug 1996), 1663–1685.

GOULD, A., BOYE, M., CROWTHER, M., IBRAHIM, J., QUARTEY, G., MICALLEF, S., AND BOIS, F.
Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the dia bayesian joint modeling working group.
Statistics in Medicine 34, 14 (Jun 30 2015), 2181–2195.

IBRAHIM, J. G., CHEN, M. H., AND SINHA, D.
Bayesian Survival Analysis.
Springer-Verlag, 2001.

MARTINS, R., SILVA, G. L., AND ANDREOZZI, V.
Bayesian joint modeling of longitudinal and spatial survival aids data.
Statistics in Medicine (2016), n/a–n/a.
sim.6937.

PROUST-LIMA, C., AND TAYLOR, J.
Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach.
Biostatistics 10, 3 (2009), 535–549.

RIZOPOULOS, D.
Jm: An r package for the joint modelling of longitudinal and time-to-event data.
Journal of Statistical software 35 (9) (2010), 1–33.

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

RIZOPOULOS, D.
Joint Models for Longitudinal and Time-to-Event Data With Applications in R.
Chapman and Hall/CRC, 2012.

Obrigado!