Bayesian joint models for longitudinal and survival data

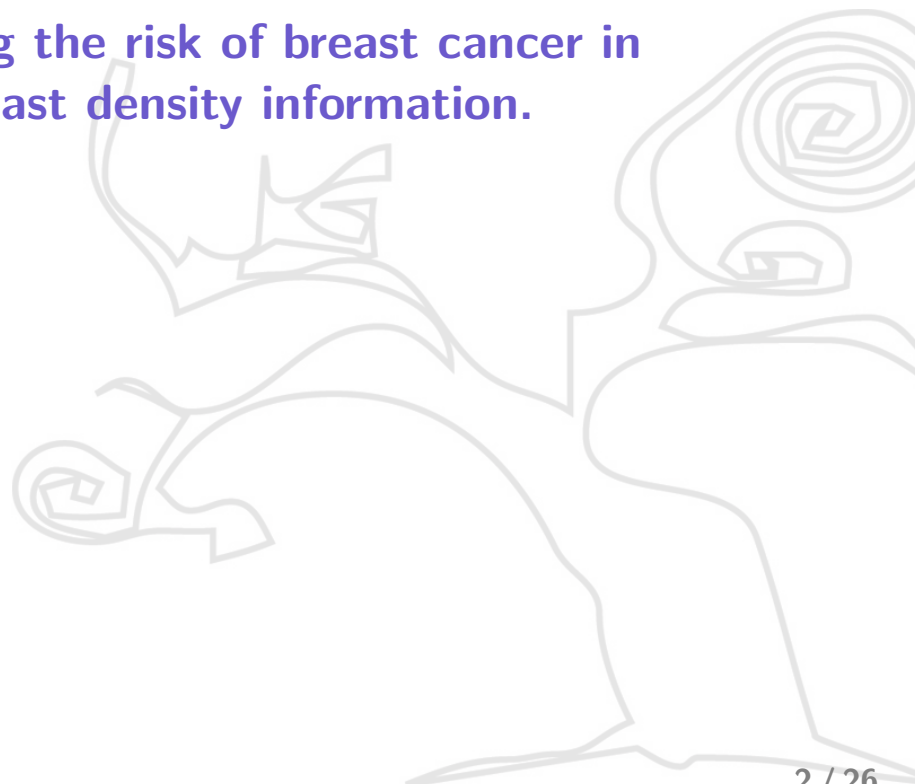
C. Armero[(1)], M. Rué[(2)], A. Forte[(1)], C. Forné[(2)],
H. Perpiñán[(1)], M. Baré[(3)], and G. Gómez-Melis[(4)]

[(1)]VaBaR and Universitat de València,
[(2)] Universitat de Lleida,
[(3)]Epidemiology and Assessment Unit UDIAT-Diagnostic Centre, Corporació Sanitària Parc Taulí, Sabadell,
[(4)]Universitat Politècnica de Catalunya

ScoVa16
València, January 26-27, 2016

Outline

- Survival, longitudinal, and joint models in VaBaR
- A Bayesian joint model for assessing the risk of breast cancer in women with longitudinal ordinal breast density information.



Sección 1 | Survival, longitudinal and joint models in VaBaR

Survival, longitudinal and joint models in VaBaR, I

- Survival models
- Longitudinal models
- A **joint model** is a joint probabilistic model for a longitudinal, $\boldsymbol{y}$, and a survival process, $\boldsymbol{s}$.
- A **Bayesian joint model** with parameters and hyperparameters $\boldsymbol{\theta}$, random effects $\boldsymbol{b}$, and covariates $\boldsymbol{x}$

$$f(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{b} \mid \boldsymbol{x}) = f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{x})\, \pi(\boldsymbol{\theta}, \boldsymbol{b})$$

- **Joints models**
  - Longitudinal studies with missing data generated by **non-ignorable mechanisms**. Survival models for modeling the missing process.
  - Survival studies. **Endogeneous temporal covariates** are modeled in terms of longitudinal models.
  - Survival and longitudinal studies.

Survival, longitudinal and joint models in VaBaR, II

- **Analysis of the risk of prostate cancer with longitudinal prostate-specific antigen data.** Joint model with survival objectives. Anabel, Hèctor, and Carmen from VaBaR and M. Rué (Universitat de Lleida), and G. Gómez, X. Piulachs, and C. Serrat (Universitat Politècnica de Catalunya).
- **Analysis of the risk of breast cancer with mammographic breast density data.** Joint model with survival objectives. Anabel, Hèctor, and Carmen from VaBaR and M. Rué and C. Forné (Universitat de Lleida), G. Gómez (Universitat Politècnica de Catalunya), and M. Baré (Hospital de Sabadell).
- **Progression of chronic kidney disease in Valencian children.** Joint model with longitudinal objective: The missing data due to the exit of children because of recovery or dialysis are modeled in terms of a competing risk survival model with left truncation. Anabel, Hèctor and Carmen from VaBaR.
- **Dynamic estimation and prediction of individual longitudinal and survival information.** Sequential Monte Carlos methods to update posterior distributions for which we only have approximate random samples. Danilo, Anabel and Carmen from VaBaR.
- **Progression of viral infections in plants.** Survival. Multistate models. Joint models. Elena and Carmen from VaBaR, and Luis Rubio from IVIA.
- **Survival probabilities in seabirds using capture and recapture methods.** Multistate. Blanca, David and Carmen from Vabar and Jonas Hentati-Sundberg, Olof Olsson, and Henrik Österblom from Stockholm Resilience Centre (Stockholm University).

Survival, longitudinal and joint models in VaBaR, III

- **Analysis of the risk of prostate cancer with longitudinal prostate-specific antigen data.** Joint model with survival objectives. Anabel, Hèctor, and Carmen from VaBaR and M. Rué (Universitat de Lleida), and G. Gómez, X. Piulachs, and C. Serrat (Universitat Politècnica de Catalunya).
- **Analysis of the risk of breast cancer with mammographic breast density data. Joint model with survival objectives. Anabel, Hèctor, and Carmen from VaBaR and M. Rué and C. Forné (Universitat de Lleida), G. Gómez (Universitat Politècnica de Catalunya), and M. Baré (Hospital de Sabadell).**
- **Progression of chronic kidney disease in Valencian children.** Joint model with longitudinal objective: The missing data due to the exit of children because of recovery or dialysis are modeled in terms of a competing risk survival model with left truncation. Anabel, Hèctor and Carmen from VaBaR.
- **Dynamic estimation and prediction of individual longitudinal and survival information.** Sequential Monte Carlos methods to update posterior distributions for which we only have approximate random samples. Danilo, Anabel and Carmen from VaBaR.
- **Progression of viral infections in plants.** Survival. Multistate models. Joint models. Elena and Carmen from VaBaR, and Luis Rubio from IVIA.
- **Survival probabilities in seabirds using capture and recapture methods.** Multistate. Blanca, David and Carmen from Vabar and Jonas Hentati-Sundberg, Olof Olsson, and Henrik Österblom from Stockholm Resilience Centre (Stockholm University).

Sección 2 | A Bayesian joint model for assessing the risk of breast cancer in women with longitudinal breast density information

A Bayesian joint model for assessing the risk of breast cancer in women with longitudinal breast density information

- Background
- Scientific objective and study population
- Data
- Statistical methods
- Results
  - Estimation
  - Prediction
- Conclusions

Background

- **Breast density (BD)** is a characteristic of the breast tissue that is reflected in mammograms. For most women, breast density decreases with age.
- The ordinal **BI-RADS scale** (American College of Radiology, 2013) is a commonly tool used to measure breast density. It categorizes breast density in four groups
  - *a* (1): almost entirely fatty **(low density)**,
  - *b* (2): scattered fibroglandular densities **(medium density)**,
  - *c* (3): heterogeneously dense **(high density)**, and
  - *d* (4): extremely dense **(very high density)**
- Several studies have shown that high breast density is associated with an increased **breast cancer (BC)** risk.
- Some research (Huo et al., 2014) concluded that longitudinal measurement of mammographic density might be used to personalize breast cancer screening **(Biomarker)**.

Objective and material

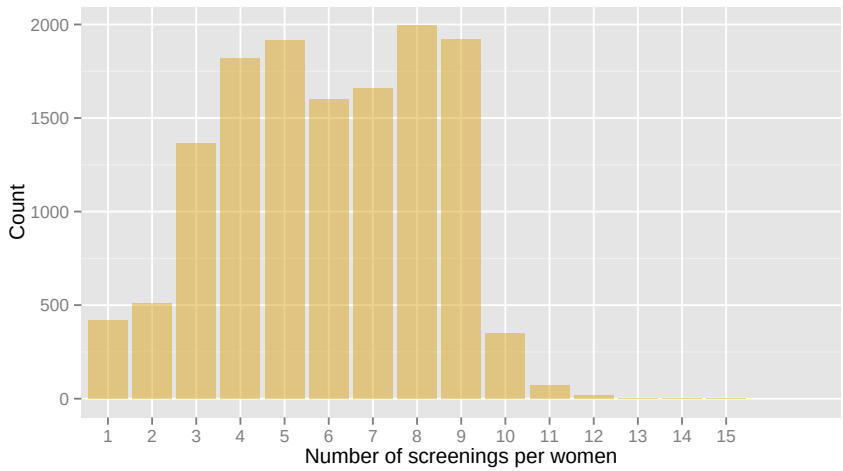
- **Objective:** To study the **association structure** and **intensity** between mammographic longitudinal measures of breast density and breast cancer risk.
- **Study population:**
  - 13 685 women attending a population-based screening program in the BC early-detection program in the Vallès Occidental East area in Catalonia (Spain), between October 1995 and June 1998.
  - Women were followed for vital status or possible diagnosis of BC until December, 2013 (Baré *et al.*, 2003, 2006, 2008).
  - At study entry
    - 50-69 years old.
    - No history of BC.
    - Invited for a biennial mammographic exam.

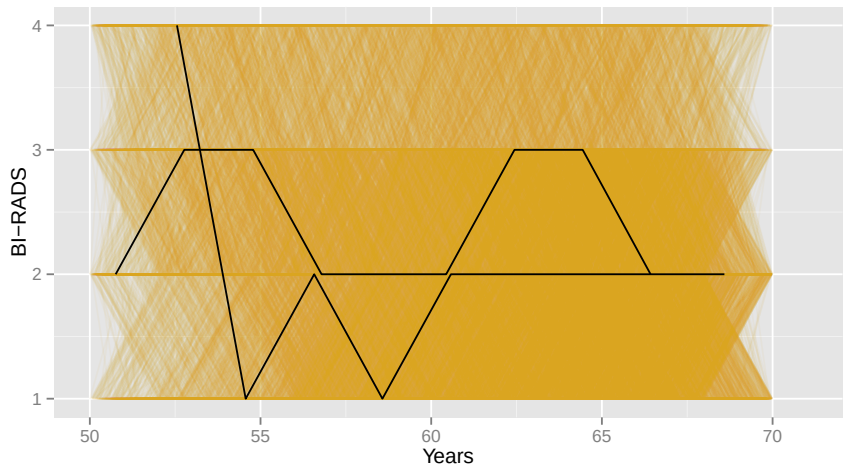
Study variables and scales: Breast density and breast cancer data

- Data from the **13 685** women attending our population-based screening program.
  - **81 621** mammographic measures of breast density with the subsequent identification and age of the woman.
  - **Time scale**: Age (50-69 years).
  - **Baseline covariates:**
    - First-degree relative with BC: Yes-No.
    - Prior breast procedures (breast biopsy, fine needle aspiration, cyst aspiration, breast reconstruction, lumpectomy and surgical treatment): Yes-No.
  - **Survival time:** Time from the initial age at entrance to the screening program until BC-diagnosis (Administrative censoring).
  - **Time zero**: 50 years. Women who entered the programme after this age produced left truncated data.

Data description

- **13 685** women and **81 621** mammograms





- **Prior breast procedures:** 7.1 % (no cancer) and 13.2 % (cancer)
- **First-degree relative with BC:** 5.2 % (no cancer) and 9.8 % (cancer)
- **431** women diagnosed with BC (3.15 %)

Statistical methodology

- **Elements:**
  - **Longitudinal data:** Breast density data
  - **Survival data:** Time from the initial age of the screening program to BC detection.
- **Joint models for longitudinal and survival data.**
  1. **Longitudinal submodel.**
  2. **Survival submodel.**
  3. **Connexion between both submodels.**

Objective and material and methods

Joint models

- **Different approaches:** conditional models, shared-parameter models, random-effects models, and joint latent class models.
- **Shared-parameter models:** Connection between the longitudinal ($\boldsymbol{y}$) and survival ($\boldsymbol{s}$) models is a common vector of random-effects $\boldsymbol{b}$ which endows them with conditional independence.

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{b}, \boldsymbol{\theta} \mid \boldsymbol{x}) &= f(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}, \boldsymbol{x})\, \pi(\boldsymbol{b}, \boldsymbol{\theta}) \\ &= f(\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{\theta}_y, \boldsymbol{x}_y)\, f(\boldsymbol{s} \mid \boldsymbol{b}, \boldsymbol{\theta}_s, \boldsymbol{x}_s)\, \pi(\boldsymbol{b}, \boldsymbol{\theta}) \end{aligned}$$

where $\boldsymbol{\theta}$ are the vector of parameters and hyperparameters of the model.

The joint model

Longitudinal submodel: A cumulative logit model for the longitudinal ordinal measures of breast density based on the idea of a continuous latent variable (Lunn et al., 2001; Luo et al., 2014).

- y_{ij} breast density in the BI-RADS scale, $y_{ij} \in \{1, 2, 3, 4\}$, of woman i, $i = 1, ..., n$, at time t_{ij} (age $50 + t_{ij}$), $j = 1, ..., n_i$.
- We assume an underlying continuous latent variable y_{ij}^* that determines the breast density BI-RADS category of woman i at time t_{ij}.
  - Relationship between y_{ij} and y_{ij}^*

$$
\begin{aligned}
y_{ij}^* &\sim \text{Logistic}(m_{ij}, s = 1) \\
y_{ij} = k &\leftrightarrow y_{ij}^* \in (\gamma_{k-1}, \gamma_k], \\
\text{logit}\, P(y_{ij} > k) &= m_{ij} - \gamma_k, \\
m_{it} &= \beta_0 + b_{i0} + (\beta_1 + b_{i1})t,
\end{aligned}
$$

with $-\infty = \gamma_0 < \gamma_1 < \gamma_2 = 0 < \gamma_3 < \gamma_4 = \infty$ unknown cutpoints, β_0, β_1 unknown parameters and $(b_{0i}, b_{1i}$ subject specific random effects).

The joint model

Survival submodel: A left-truncated Cox proportional hazard model for the time-to-BC detection that incorporates information from the longitudinal process.

- T_i, $i = 1, \ldots, n$ observed BC detection time for the i-th woman, $T_i = \min(T_i^*, C_i)$, T_i^* is the true failure time and C_i the right-censoring time. The event indicator, $\delta_i = I(T_i^* \leq C_i)$.
- Hazard function of T_i^* in terms of the left truncated Cox proportional hazard model

$$h_i(t|\boldsymbol{x}_i, \boldsymbol{\theta}_{is}, t_i^* > a_i) = h_0(t \mid \lambda, \eta_0) \exp\{\eta_1 Famhist_i + \eta_2 Brstproc_i + \alpha\, m_{it}\}, \quad t > a_i$$

  - $h_0(t \mid \lambda, \eta_0) = \lambda t^{\lambda-1} e^{\eta_0}$: baseline risk function of a Weibull distribution;
  - $\boldsymbol{x}_i$: baseline covariates, *First-degree relative with BC (Famhist)* and *prior breast procedures (Brstproc)* with regression coefficients η_1 and η_2;
  - α: effect of the individual trajectory of breast density of a woman over their BC risk in terms of the latent BD mean;
  - a_i: age over 50 at which woman i enters the screening program thus providing a left truncated time (Uzunogullari et al., 1992);
  - $\boldsymbol{\theta}_{is} = (\lambda, \eta_0, \eta_1, \eta_2, \alpha, \boldsymbol{\theta}_{il})^T$, with $\boldsymbol{\theta}_{il} = (\beta_0, \beta_1, b_{0i}, b_{1i})^T$.

The joint model

1. **Longitudinal submodel:** A cumulative logit model for the longitudinal ordinal measures of breast density based on the idea of a continuous latent variable (Lunn et al., 2001; Luo et al., 2014), and

$$
\begin{aligned}
y_{ij} = k &\leftrightarrow y_{ij}^* \in (\gamma_{k-1}, \gamma_k], \\
\text{logit}\, P(y_{ij} > k) &= m_{ij} - \gamma_k, \\
m_{it} &= \beta_0 + \boldsymbol{b}_{i0} + (\beta_1 + \boldsymbol{b}_{i1})t,
\end{aligned}
$$

2. **Survival submodel:** A left-truncated Weibull proportional hazard model for the time-to-BC detection, that incorporates information from the longitudinal process.

$$
\begin{aligned}
h_i(t|\boldsymbol{x}_i, \boldsymbol{\theta}_{is}, t_i^* > a_i) &= h_0(t \mid \lambda, \eta_0) \exp\{\eta_1 Famhist_i + \eta_2 Brstproc_i + \alpha\, m_i(t)\} \\
&= h_0(t \mid \lambda, \eta_0) \exp\{\eta_1 Famhist_i + \eta_2 Brstproc_i + \alpha(\beta_0 + \boldsymbol{b}_{i0} + (\beta_1 + \boldsymbol{b}_{i1})t)\},
\end{aligned}
$$

3. Both processes are connected through a **shared vector of random effects**, which, in the presence of covariates and parameters, endows both processes with conditional independence (Rizopoulos, 2012).

Bayes Inference

- **Bayesian Inference:** Complete specification of the joint model needs the elicitation of a prior distribution for the subsequent parameters and hyperparameters.
  - **Prior independence** as a default specification and wide proper prior distributions with the aim of giving all inferential prominence to the data.
  - **Normal** distributions for the regression coefficients of the longitudinal and survival submodels and the association coefficient.
  - **Truncated Normal** distributions for the cutpoints of the latent scale, $-\infty = \gamma_0 < \gamma_1 < \gamma_2 = 0 < \gamma_3 < \gamma_4 = \infty$ selected to provide the same prior probability to each response category.
  - **Gamma** distribution, Ga(1, 1), for the parameter of the baseline hazard function because it mimics a constant baseline hazard function (Guo et al.,)
  - **Uniform** hyperdistribution distribution Un(0, 10) for the standard deviations σ_0 and σ_1 of the normal random effects $b_{i0} \sim \mathrm{N}(0, \sigma_0)$ and $b_{i1} \sim \mathrm{N}(0, \sigma_1)$.

Posterior distribution

Longitudinal submodel: A cumulative logit model for the longitudinal ordinal measures of breast density based on the idea of a continuous latent variable (Lunn et al., 2001; Luo et al., 2014).

$$y_{ij} = k \leftrightarrow y_{ij}^* \in (\gamma_{k-1}, \gamma_k],$$
$$\text{logit}\, P(y_{ij} > D_k) = m_{ij} - \gamma_k,$$
$$m_{it} = \beta_0 + \boldsymbol{b}_{i0} + (\beta_1 + \boldsymbol{b}_{i1})t,$$

Survival submodel: A left-truncated Weibull proportional hazard model for the time-to-BC detection, that incorporates information from the longitudinal process.

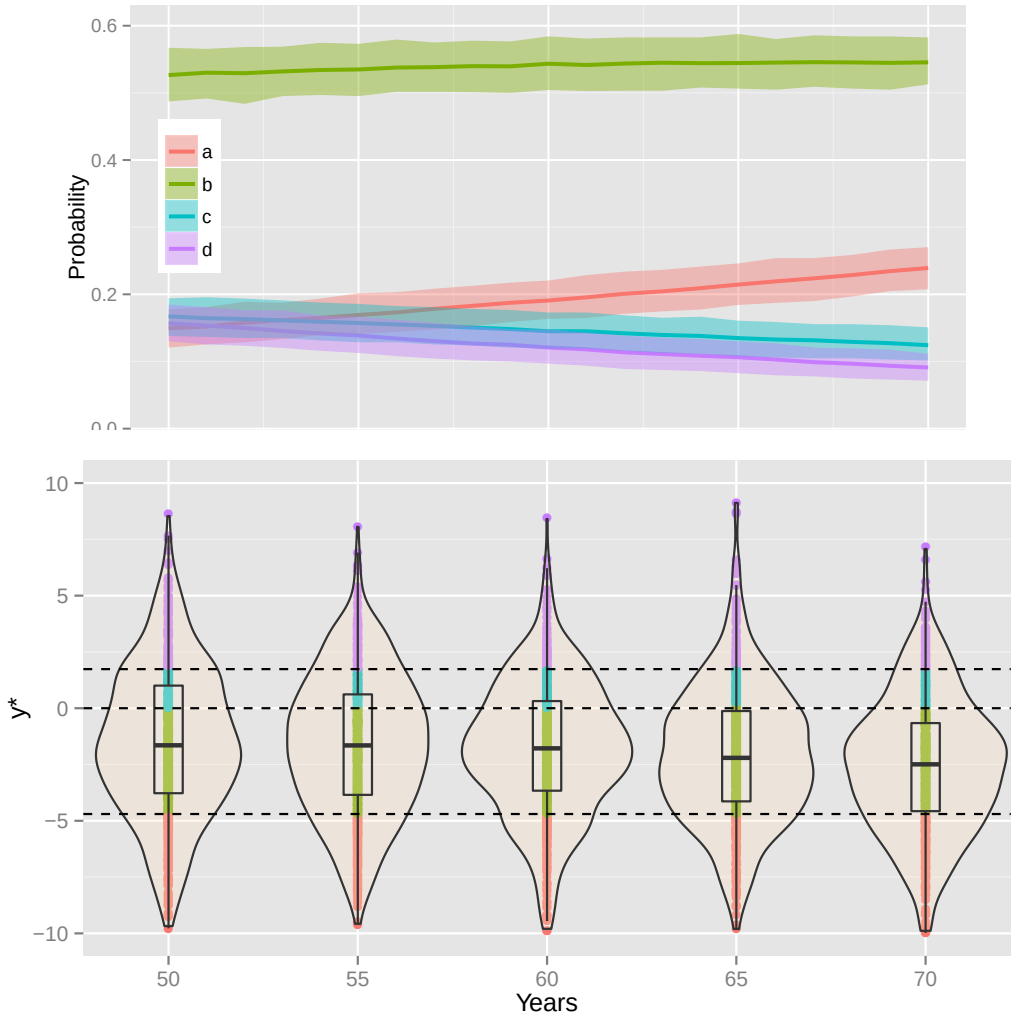
$$h_i(t|\boldsymbol{x}_i, \boldsymbol{\theta}_{is}, t_i^* > a_i) = h_0(t \mid \lambda, \eta_0) \exp\{\eta_1 Famhist_i + \eta_2 Brstproc_i + \alpha\, m_{it}\}$$
$$= h_0(t \mid \lambda, \eta_0) \exp\{\eta_1 Famhist_i + \eta_2 Brstproc_i + \alpha(\beta_0 + \boldsymbol{b}_{i0} + (\beta_1 + \boldsymbol{b}_{i1})t)\}, \quad t > a_i$$

- Posterior distribution: Approximated by Markov Chain Monte Carlo methods.

	Mean	Sd	2.5 %	Median	97.5 %	$P(\cdot > 0 \mid \mathcal{D})$
β_0	-1.4262	0.0346	-1.4964	-1.4251	-1.3608	0.0000
β_1	-0.0524	0.0018	-0.0560	-0.0524	-0.0489	0.0000
σ_0	2.6067	0.0227	2.5643	2.6059	2.6534	
σ_1	0.0053	0.0018	0.0015	0.0053	0.0087	
γ_1	-4.6994	0.0269	-4.7521	-4.6998	-4.6489	
γ_3	1.7362	0.0156	1.7060	1.7364	1.7675	
η_0	-7.6066	0.3369	-8.2476	-7.6011	-6.9337	0.0000
η_1	0.6227	0.1716	0.2747	0.6308	0.9517	0.9984
η_2	0.4535	0.1440	0.1644	0.4600	0.7210	1.0000
α	0.1490	0.0207	0.1089	0.1496	0.1887	1.0000
λ	1.5366	0.1044	1.3287	1.5387	1.7386	

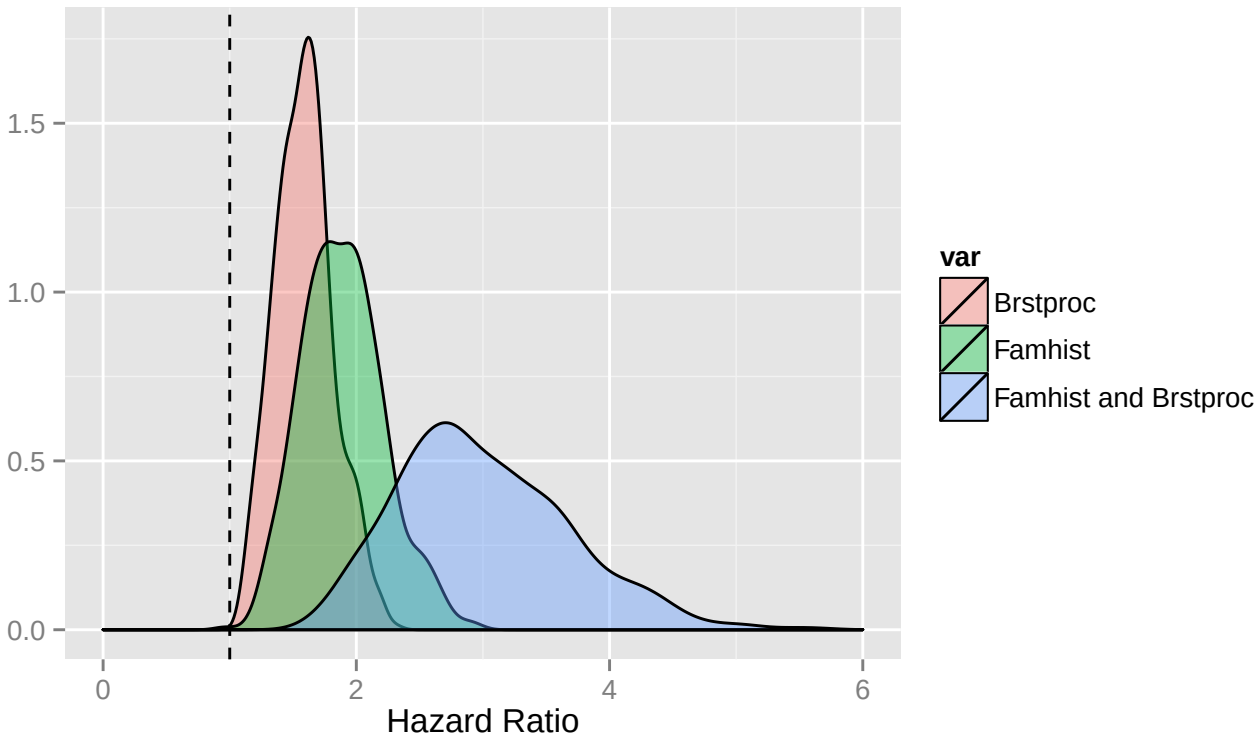
Posterior distribution for breast density

- Posterior distribution of breast density in the ordinal (posterior mean and 95 % credible interval) and latent (posterior distribution at 50, 55, 60, 65 and 70 years) scales.



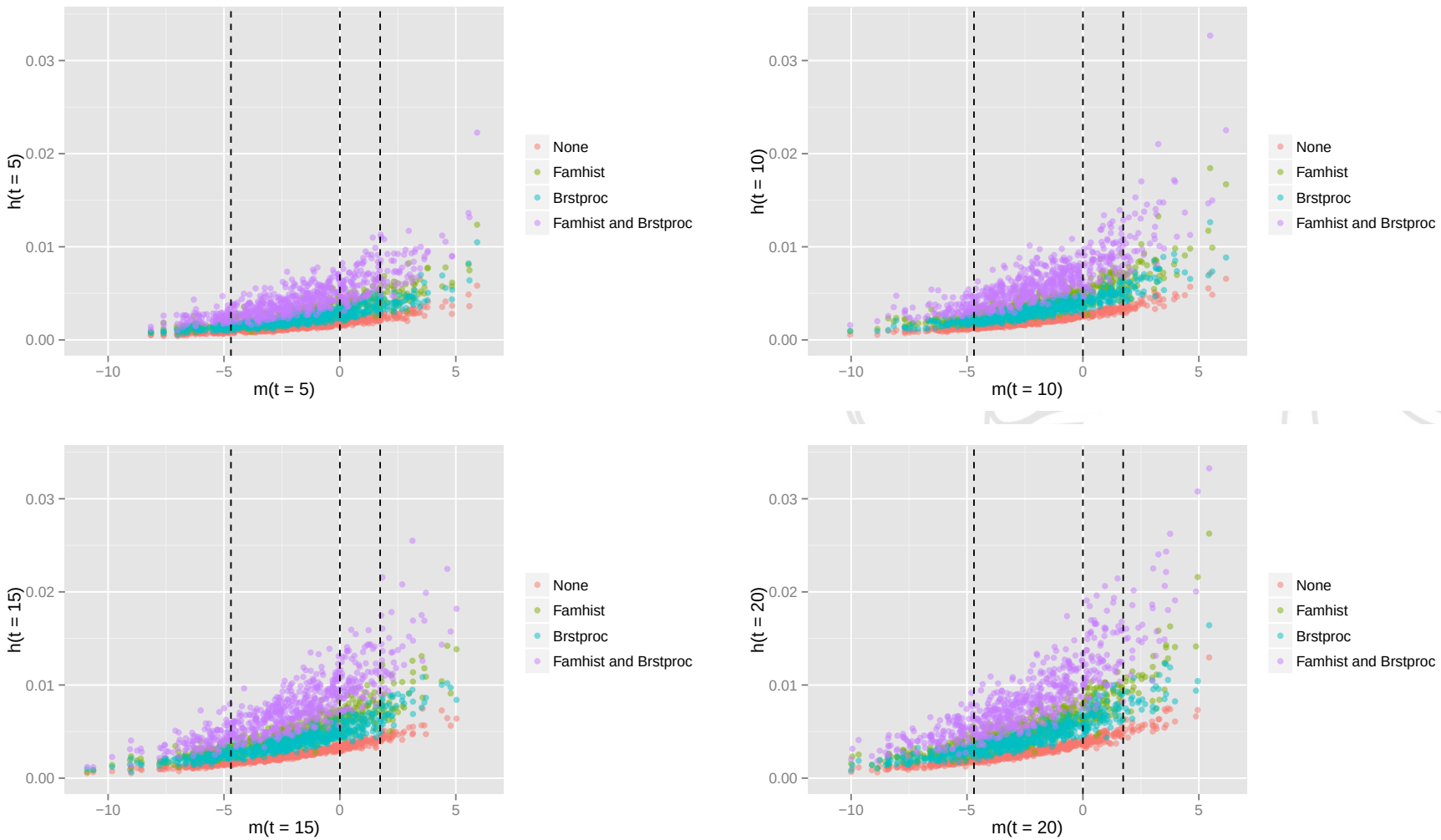
Hazard ratios

- **Posterior distribution for the hazard ratio (HR):**
  - HR of BC diagnosis in women with family history of BC *versus* women without family history of BC.
  - HR of BC diagnosis in women with prior breast procedures *versus* women without prior breast procedures.
  - HR of BC diagnosis in women with family history of BC and prior breast procedures *versus* women without family history of BC and without prior breast procedures.



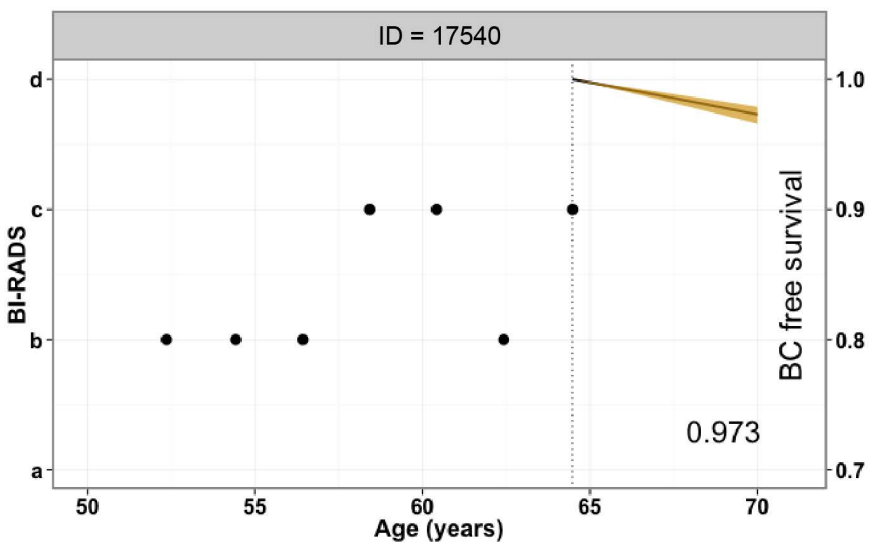
Hazard ratios

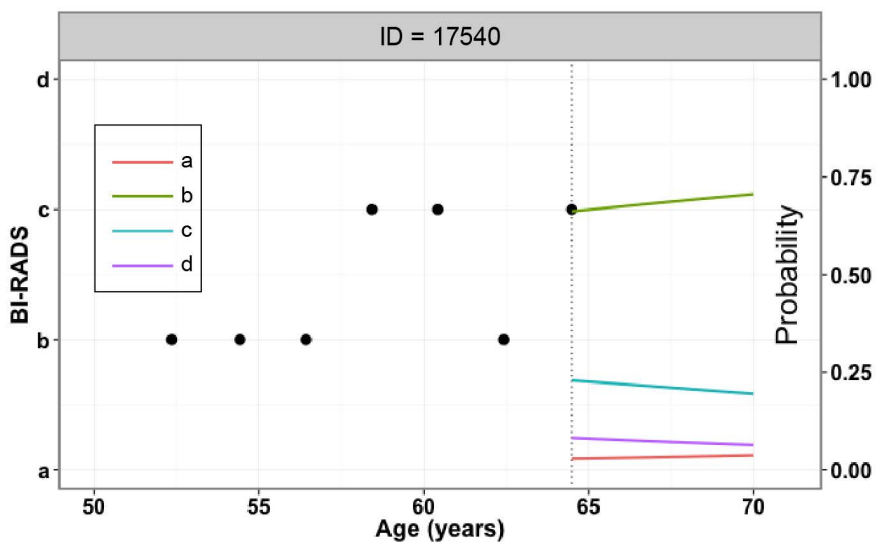
Posterior distribution for the hazard risk function with regard to the latent mean of BD:



Prediction

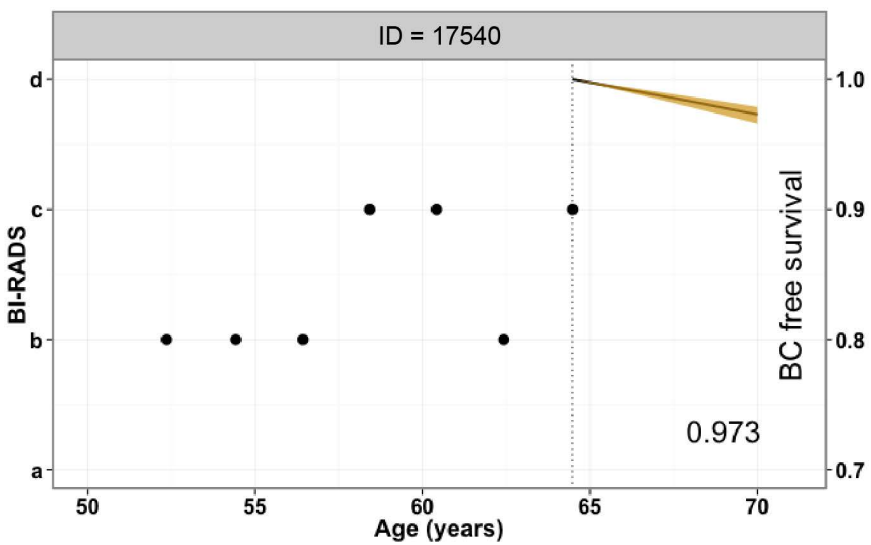
- Posterior distribution:
  - $\pi(P(T_i > t \mid \boldsymbol{\theta}, \boldsymbol{b}, \mathcal{Y}_i, \boldsymbol{x}_i, T_i > t_{i,n_i}) \mid \mathcal{D}),\ t > t_{i,n_i}$
  - $f(y_{i,t_{n_i+1}} \mid \mathcal{Y}_i, \boldsymbol{x}_i, \mathcal{D})$
- Posterior mean and 95 % credible interval of the survival probability distribution and posterior mean of the predictive mammogram breast density of woman id17540 (No family history of BC and no prior breast procedures) with a given follow-up which includes all her historical longitudinal mammogram and covariates.

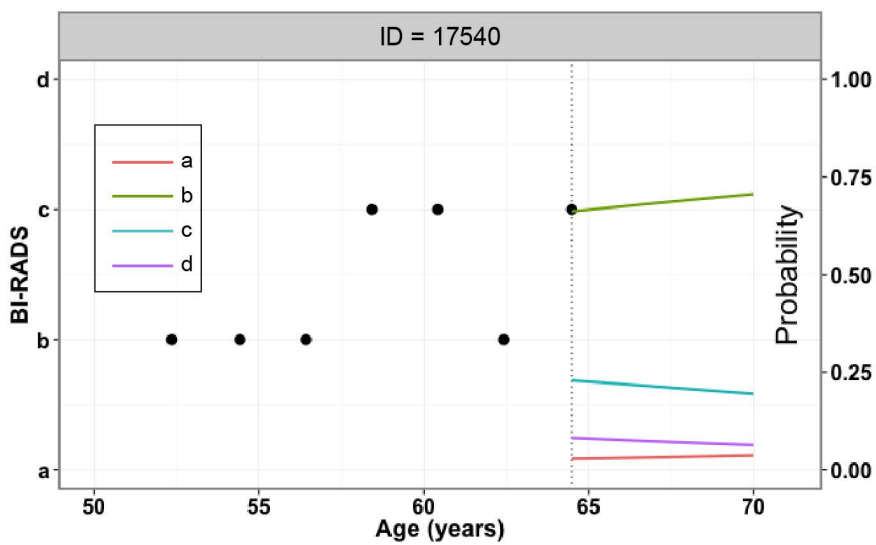




Prediction

- Posterior distribution:
  - $\pi(P(T_i > t \mid \boldsymbol{\theta}, \boldsymbol{b}, \mathcal{Y}_i, \boldsymbol{x}_i, T_i > t_{i,n_i}) \mid \mathcal{D}), \ t > t_{i,n_i}$
  - $f(y_{i,t_{n_i+1}} \mid \mathcal{Y}_i, \boldsymbol{x}_i, \mathcal{D})$
- Posterior mean and 95 % credible interval of the survival probability distribution and posterior mean of the predictive mammogram breast density of woman id17540 (No family history of BC and no prior breast procedures) with a given follow-up which includes all her historical longitudinal mammogram and covariates.





Sección 3 | Conclusions

Conclusions

- Joint models of longitudinal and survival data are a suitable tool for analyzing the relationship between mammographic breast density and breast cancer risk.
- Our joint model is a good starting modeling for learning about the problem but we need to introduce more flexibility and work more on the problem:
  - Non-linear trajectories in the longitudinal mammogram BD modeling.
  - Leave the Weibull baseline risk function and explore other proposals which allow for multimodal or heavy tailed patterns.
  - Assess the performance of the biomarker: Sensitivity, specificity and time-dependent receiver operating characteristic (ROC) curves.
  - Sensitivity of the prior distributions, in particular the hyperpriors for the random effects.
  - Improve our knowledge about latent variables in ordinal models.