Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Bayesian joint ordinal and survival modeling for breast cancer risk assessment

C. Armero[1], C. Forné[2,3], M. Rué[2,4], A. Forte[1], H. Perpiñán[1,5], G. Gómez[6], and M. Baré[4,7]

We propose a joint model to analyze the structure and intensity of the association between longitudinal measurements of an ordinal marker and time to a relevant event. The longitudinal process is defined in terms of a proportional-odds cumulative logit model. Time-to-event is modelled through a left-truncated proportional-hazards model which incorporates information of the longitudinal marker as well as baseline covariates. Both longitudinal and survival processes are connected by means of a common vector of random effects.

General inferences are discussed under the Bayesian approach and include the posterior distribution of the probabilities associated to each longitudinal category and the assessment of the impact of the baseline covariates and the longitudinal marker on the hazard function. The flexibility provided by the joint model makes possible to dynamically estimate individual event-free probabilities and predict future longitudinal marker values.

The model is applied to the assessment of breast cancer risk in women attending a population-based screening program. The longitudinal ordinal marker is mammographic breast density measured with the BI-RADS scale in biennial screening exams.

Keywords: BI-RADS scale; Latent process; Left-truncated proportional-hazards model; Proportional-odds cumulative logit model

[1] *Department of Statistics and Operational Research, Universitat de València, Doctor Moliner, 50, 46100-Burjassot, Spain.*
[2] *Department of Basic Medical Sciences, Universitat de Lleida-IRBLleida, Avda. Rovira Roure, 80, 25198-Lleida, Spain.*
[3] *Oblikue Consulting, Barcelona, Spain.*
[4] *Health Services Research Network in Chronic Diseases (REDISSEC).*
[5] *Fundación para el Fomento de la Investigación Sanitaria y Biomédica (FISABIO), Generalitat Valenciana, Spain.*
[6] *Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain.*
[7] *Clinical Epidemiology and Cancer Screening. Corporació Sanitària Parc Taulí-UAB, Sabadell, Parc Taulí s/n, 08208, Sabadell, Spain.*

1. Introduction

The current evidence on benefits and harms supports the personalization of screening as a crucial step to improve early detection of breast cancer [1,2]. A number of risk models were designed to measure the individual probability of developing breast cancer [3–5]. In the context of individualized breast cancer screening, the utility of these risk models has been questioned due to their low discrimination power [6]. The inclusion of a baseline measure of breast density - a characteristic of the breast tissue - in the risk models improved the accuracy of the breast cancer risk estimate [7–10].

Several studies have shown that high breast density is associated with increased breast cancer risk [7,11–14], with risk estimates in the range four - to six - fold for women with very high breast density compared to women with low breast density [11,12]. Other studies have examined whether changes in breast density over time are associated to changes in breast cancer risk [15–21], and have suggested that monitoring changes in breast density could help to identify women at greater risk of disease. In most of the cases, the statistical methods used did not account for relevant characteristics of prospective studies like non-ignorable dropout mechanisms or internal time-dependent covariates [22].

Joint modeling of longitudinal and time-to-event data is an increasingly productive area of statistical research that assesses the association between longitudinal and survival processes. It enhances longitudinal modeling by allowing for the inclusion of non-ignorable dropout mechanisms, and survival modeling by the inclusion of internal time-dependent covariates [22]. Joint models were introduced during the 90s [23–25] and since then, have been applied to a great variety of studies in epidemiological and biomedical areas. Shared-parameter models are a type of joint models where the longitudinal and time-to-event processes are connected by means of a common set of subject-specific random effects. These models make possible to quantify both the population and individual effects of the underlying longitudinal outcome on the risk of an event, and obtain individualized time-dynamic predictions. Recently, Rizopoulos proposed an overview of the theory and applications of joint modeling [26] and developed the `JM` [27] and `JMbayes` [28] `R` packages for the frequentist and Bayesian shared-effects' approaches, respectively. Serrat *et al.* illustrate the application of both statistical approaches to joint modeling longitudinal measures of prostate specific antigen (PSA) and prostate cancer detection in men participating in a screening trial [29].

When longitudinal outcomes are ordinal, joint models become more complex. Different approaches, that use constraints in the probabilities of the categorical outcomes or the discretization of a continuous latent variable, have been proposed [30–33]. The non-linear and longitudinal nature of the data produce a complex likelihood function, difficult to maximize under the frequentist paradigm. This could be a reason why the standard software for joint models does not include longitudinal ordinal variables yet. Some relevant works on the subject use the frequentist [34–36] and Bayesian [33,37,38] approaches, respectively.

The objective of this paper is to propose a Bayesian joint model for assessing the structure and intensity of the association between longitudinal measures of an ordinal marker and a time-to-event outcome. In particular, we use a proportional-odds cumulative logit model [30] for the ordinal measurements and a proportional hazard model with left-truncation for the time to an event of interest. We have applied the model to analyze the risk of breast cancer in women attending a population-based screening program with regard to repeated measurements of mammographic breast density.

Section 2 presents a description of the motivating dataset. Section 3 formulates the joint model and discusses general inferences for 1) dynamic probabilities associated to the different ordinal categories, 2) the impact of baseline covariates and the longitudinal marker on the hazard function, 3) dynamic estimation of survival probabilities, and 4) prediction of future longitudinal outcomes. Section 4 applies the model developed in Section 3 to study age at diagnosis of breast cancer in women who participate in a population-based screening program. Finally, Section 5 contains a discussion and some conclusions.

2. Motivating data

2.1. Study design and study population

This is an observational prospective study including 13 760 women that participated for the first time in the breast cancer early-detection program in the Vallès Occidental Est (BCEDP-VOE) area in Catalonia (Spain), between October 1995 and June 1998. The BCEDP-VOE invites women aged 50-69 years for biennial mammographic exams. At study entry, the participants were 50-70 years old and did not have a personal history of breast cancer. They were followed for vital status or possible diagnosis of breast cancer until December 2013 [39–41].

Of the initial 13 760 women, we excluded seven without follow-up data, as well as 38 women who were diagnosed with breast cancer and nine who died within six months of baseline. Twenty-one women were also excluded for not having breast density measurements within the 50-70 age interval. We analyzed invasive breast cancer and ductal carcinoma in situ (DCIS) diagnosed during follow-up. The final sample included 13 685 women, with 431 diagnosed with breast cancer.

2.2. Variables and data description

At the first mammographic exam, the study participants answered a questionnaire that included information on family history of breast cancer, prior breast procedures, age at menarche, age at first birth, and menopausal status. Family history refers to absence/presence of first-degree relatives with breast cancer. Prior breast procedures included breast biopsy, fine needle aspiration, cyst aspiration, breast reconstruction, lumpectomy and surgical treatment.

Breast density is a characteristic of the breast tissue that is reflected in mammograms. Breasts are considered dense if the connective and epithelial tissues predominate over the fatty tissue. At all mammographic exams breast density was rated and recorded according to the BI-RADS system [40] which categorizes breast density in four groups: a, almost entirely fatty (low density); b, scattered fibroglandular densities (medium density); c, heterogeneously dense (high density); and d, extremely dense (very high density). This longitudinal breast density data is a remarkable and unique characteristic of the BCEDP-VOE among the breast cancer screening programs in Spain. Breast density measures within 6 months before breast cancer diagnosis could be affected by the presence of preclinical breast cancer, therefore they were excluded. The mammographic exams performed before age 50 or after age 70 were excluded in order to avoid sample biases. A total of 81 621 measures of breast density were included in the longitudinal analysis, with median [range] 4 [1 to 9] and 6 [1 to 15] measures for women with or without breast cancer diagnosis, respectively.

We considered that the time origin for the event of interest (diagnosis of breast cancer) was age 50 years, the lower limit of the screening age interval. We defined the time-to-event as the time elapsed from the origin to diagnosis of breast cancer. For women without a breast cancer diagnosis at the study end, the censoring time was obtained as the minimum of time to death and time to the last screening exam plus 2.5 years that correspond to the active follow-up for cancer identification. It is important to remark that women who entered the program over 50 years had delayed entry times that may induce length biased sampling or left truncation.

TABLE 1 AROUND HERE

Among 13 685 women aged 50 years and older, 431 developed breast cancer –336 invasive cancers and 95 DCIS–, and 513 died within 2.5 years of the last mammogram. Median follow-up was 12.7 years for women without breast cancer and 8.2 years for women with breast cancer. Table 1 shows the baseline characteristics of women and the breast density measurements at first and last examination according to their breast cancer diagnosis status at the end of follow-up. High breast density categories were more prevalent among women with breast cancer, in both the first and the last mammogram. Furthermore, between the first and last exams, the prevalence of low density

categories increased, as described in the literature.

FIGURE 1 AROUND HERE

To illustrate the longitudinal breast density measurements, we randomly selected eight women without and eight women with cancer (Figure 1). A high variability of the breast density trajectories can be observed: some women experience an increase of breast density, while others remain stable, fluctuate, or experience a decrease. The plots show the biennial periodicity of the screening exams, as well as the unbalanced number of measures between women, due to different reasons. Not all women entered the study at the same age, not all the scheduled screening exams were taken, or not always breast density was rated.

3. The joint model

We propose a model with two submodels: (1) a proportional-odds cumulative logit model for the longitudinal ordinal measurements based on the idea of a continuous latent variable [30,33], and (2) a left-truncated proportional hazard model for the time-to-event, which incorporates information from the longitudinal process. Both processes are connected through a shared vector of random effects, which, in the presence of covariates and parameters, endows them with conditional independence [26].

Let $\{D_1, D_2, \ldots, D_K\}$ denote the set of ordinal categories and y_{ij} the longitudinal category of individual i, $i = 1, \ldots, n$, at time t_{ij}, $j = 1, \ldots, n_i$. We assumed an underlying continuous latent variable y^*_{ij} that determines the ordinal category of individual i at time t_{ij}. This latent variable has no interest *per se* but it is useful for motivating and interpreting the cumulative logit model. The relationship between y_{ij} and y^*_{ij} is stated as

$$y_{ij} = D_k \;\Leftrightarrow y^*_{ij} \in (\gamma_{k-1},\, \gamma_k],\;\; k = 1, \ldots, K,$$

where $-\infty = \gamma_0 < \gamma_1 < \ldots < \gamma_{K-1} < \gamma_K = \infty$ are unknown cut-points. We assumed a logistic distribution for y^*_{ij}, $\text{Lo}(m_{ij}, s = 1)$, with location parameter m_{ij} (mean) and a common scale parameter $s = 1$ for achieving identifiability. The choice of that distribution implies a logit link for the cumulative probabilities

$$q_{ijk} = P(y_{ij} > D_k) = P(y^*_{ij} > \gamma_k) = \frac{1}{1 + \exp(\gamma_k - m_{ij})}, \tag{1}$$

and therefore,

$$\text{logit}\, q_{ijk} = \log\left(\frac{q_{ijk}}{1 - q_{ijk}}\right) = m_{ij} - \gamma_k.$$

Despite the fact that $s = 1$ in the logistic distribution of the latent variable, the model is overparameterized (any set of k probabilities can be obtained increasing the cut-points in the same quantity). To obtain an identifiable model, we arbitrarily introduced a reference point on the latent scale, in particular $\gamma_{K/2} = 0$ if K is even and $\gamma_{(K-1)/2} = 0$ or $\gamma_{(K+1)/2} = 0$ if K is odd.

We considered a mixed-effects model to describe the subject-specific time trajectories of the latent variable

$$y^*_{ij} = m_{ij} + \epsilon_{ij} = \boldsymbol{x}^{(l)\prime}_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i + \epsilon_{ij}, \tag{2}$$

where $\boldsymbol{x}^{(l)}_{ij}$ is a P dimensional vector of covariates relevant to the longitudinal process, as indicated by superscript (l), for individual i at time t_{ij} with regression coefficient (populational) $\boldsymbol{\beta}$; $\boldsymbol{z}_{ij}$ the vector of explanatory variables

attached to the vector of random effects $\boldsymbol{b}_i$ for the i-th individual at time t_{ij}; and ϵ_{ij} an error term for the i-th individual at time t_{ij}, modeled in terms of a logistic distribution, Lo(0, 1). The random effects $\boldsymbol{b} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)^T$ are conditionally i.i.d. (given the hyperparameter vector $\boldsymbol{\phi}$) with $(\boldsymbol{b}_i \mid \boldsymbol{\phi}) \sim f(b_i \mid \boldsymbol{\phi})$.

Let T_i, $i = 1, \ldots, n$, be the observed event time for the i-th subject, obtained as the minimum between the true failure time, T_i^*, and the right-censoring time, C_i, $T_i = \min(T_i^*, C_i)$. The event indicator $\delta_i = I(T_i^* \leq C_i)$ takes the value 1 if the observed time corresponds to a true event time, and 0 otherwise. In addition, event times corresponding to individuals who enter the study at delayed entry times introduce left-truncation, thus defining the subsequent hazard function as zero in the period before the entrance of the individual to the system [43]. In particular, we consider the hazard function of T_i^* in terms of a left-truncated relative risk regression model

$$h_i(t) = h_0(t)\, \exp\{\boldsymbol{x}_i^{(s)\prime}\boldsymbol{\eta} + \alpha\, m_{it}\},\ t > a_i, \tag{3}$$

and zero otherwise, where $h_0(t)$ is the baseline risk function; $\boldsymbol{x}_i^{(s)}$ represents the vector of baseline covariates relevant to the survival process, as indicated by superscript (s), with associated coefficients $\boldsymbol{\eta}$; α assesses the effect of the longitudinal marker of subject i on the event of interest in terms of the latent variable mean; a_i is the delayed entry time of individual i. It is important to comment that left-truncated data will add computational complexity to the modeling because the likelihood function corresponding to this type of data will incorporate conditional probabilities which contain the information that the individual is alive in the period between their theoretical entrance at time zero and their real entrance to the system.

To complete the Bayesian model, it is necessary to elicit a prior distribution, $\pi(\boldsymbol{\theta})$ for all the unknown parameters and hyperparameters of the joint model $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}, \alpha, \boldsymbol{\gamma}, \boldsymbol{\phi})^T$. Our joint model contains parameters and hyperparameters, $\boldsymbol{\theta}$, and random effects $\boldsymbol{b}$. From a Bayesian perspective, $\pi(\boldsymbol{\theta}, \boldsymbol{b} \mid \mathcal{D})$, where $\mathcal{D}$ represents all the data collected from the longitudinal and the survival processes, is the joint posterior distribution of the parameters, hyperparameters, and random effects, which can be obtained by the hierarchical modeling

$$\pi(\boldsymbol{\theta}, \boldsymbol{b} \mid \mathcal{D}) = L(\boldsymbol{\theta}, \boldsymbol{b} \mid \mathcal{D})\, f(\boldsymbol{b} \mid \boldsymbol{\phi})\, \pi(\boldsymbol{\theta}). \tag{4}$$

$L(\boldsymbol{\theta}, \boldsymbol{b} \mid \mathcal{D})$ is the likelihood function of $\boldsymbol{\theta}$ and $\boldsymbol{b}$ for data $\mathcal{D}$, $f(\boldsymbol{b} \mid \boldsymbol{\phi})$ the distribution of the random effects $\boldsymbol{b}$ given $\boldsymbol{\phi}$ introduced before, and $\pi(\boldsymbol{\theta})$ the prior distribution for $\boldsymbol{\theta}$. Markov Chain Monte Carlo (MCMC) simulation methods allow to obtain an approximated random sample from the posterior $\pi(\boldsymbol{\theta}, \boldsymbol{b} \mid \mathcal{D})$, which is the key element and the starting point of all relevant inferences.

Finally, it is worth noting that when inference is carried out under the Bayesian formulation, the shared joint model will induce conditional independence between the longitudinal and the survival processes given not only the random effects and covariates but also given all the parameters and hyperparameters in the model, as a result of its stochastic role in Bayesian inference.

3.1. Dynamic probabilities associated to ordinal categories

From expression (1), the probability distribution of the ordinal marker y_{it} for individual i at time t can be computed in terms of the logistic distribution of their latent variable y_{it}^* as

$$P(y_{it} = D_k \mid \boldsymbol{x}_{it}^{(l)}, \boldsymbol{\theta}, \boldsymbol{b}_i) = P(y_{it}^* \in (\gamma_{k-1}, \gamma_k) \mid \boldsymbol{x}_{it}^{(l)}, \boldsymbol{\theta}, \boldsymbol{b}_i),\ k = 1, 2, \ldots, K. \tag{5}$$

These probabilities depend on $\boldsymbol{\theta}$, $\boldsymbol{b}_i$, and the relevant covariates associated to that individual. Consequently, we could use the posterior marginal distribution $\pi(\boldsymbol{\theta}, \boldsymbol{b}_i \mid \mathcal{D})$ for computing the posterior distribution, $\pi(P(y_{it} = D_k \mid \boldsymbol{x}_{it}^{(l)}, \boldsymbol{\theta}, \boldsymbol{b}_i) \mid \mathcal{D})$, of all the relevant dynamic probabilities for each individual in the study.

A complementary and overall perspective of the temporal evolution of the different categories of the ordinal marker is based on the marginal distribution

$$P(y_{it} = D_k \mid \boldsymbol{x}_{it}^{(l)}, \boldsymbol{\theta}) = \int P(y_{it} = D_k \mid \boldsymbol{x}_{it}^{(l)}, \boldsymbol{\theta}, \boldsymbol{b}_i)\, f(\boldsymbol{b}_i \mid \boldsymbol{\phi})\, \mathrm{d}\boldsymbol{b}_i, \tag{6}$$

which is computed by integrating out the random effects of the conditional distribution (5). This distribution only depends on $\boldsymbol{\theta}$. It can be interpreted as the time-specific population distribution of the longitudinal marker for a generic individual of the population with covariate values $\boldsymbol{x}_{it}^{(l)}$. Consequently, we can use our current information about $\boldsymbol{\theta}$ expressed through $\pi(\boldsymbol{\theta} \mid \mathcal{D})$ and compute the posterior distribution $\pi(P(y_{it} = D_k \mid \boldsymbol{x}_{it}^{(l)}, \boldsymbol{\theta}) \mid \mathcal{D})$ for each ordinal category D_k. This posterior distribution provides point estimates of these relevant probabilities such as posterior expectations

$$\mathrm{E}(P(y_{it} = D_k \mid \boldsymbol{x}_{it}^{(l)}, \boldsymbol{\theta}) \mid \mathcal{D}) = \int P(y_{it} = D_k \mid \boldsymbol{x}_{it}^{(l)}, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\theta} = P(y_{it} = D_k \mid \boldsymbol{x}_{it}^{(l)}, \mathcal{D}), \tag{7}$$

as well as credible intervals for measuring the uncertainty of the estimation.

Our model also allows to explore the estimated relationship between the ordinal and latent variables. We could address the posterior distribution of the latent variable y_{it}^* for each time with regard to each individual in the study or a generic one. The logistic distribution, $\mathrm{Lo}(m_{it}, s = 1)$, for the latent variable y_{it}^* is a conditional distribution with an unknown mean that depends on $\boldsymbol{\theta}$ and $\boldsymbol{b}_i$. The subsequent marginal distribution $f(y_{it}^* \mid \boldsymbol{x}_{it}^{(l)}, \boldsymbol{\theta})$ can be obtained as in (6) integrating out the random effects, and can be interpreted as a time-specific population distribution of the latent variable for a generic individual. Again, this marginal distribution is also a conditional distribution that depends on the population parameters $\boldsymbol{\theta}$ and the posterior distribution of the latent variable can be estimated as

$$f(y_{it}^* \mid \boldsymbol{x}_{it}^{(l)}, \mathcal{D}) = \int f(y_{it}^* \mid \boldsymbol{x}_{it}^{(l)}, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \mathcal{D})\, \mathrm{d}\boldsymbol{\theta}. \tag{8}$$

3.2. Impact of the covariates on the risk of the event

The hazard ratio (HR) of an individual with covariates $\boldsymbol{x}$ having the event as compared to an individual with covariates $\boldsymbol{x}^*$ is $\exp\{\sum_{p=1}^{P} \eta_p(x_p - x_p^*)\}$, where P is the number of covariates. This hazard ratio only depends on $\boldsymbol{\eta}$, the vector of regression coefficients in 3. Consequently, its posterior distribution,

$$\pi(\exp\{\textstyle\sum_{p=1}^{P} \eta_p(x_p - x_p^*)\} \mid \mathcal{D}), \tag{9}$$

computed from the approximate MCMC sample from the posterior marginal $\pi(\boldsymbol{\eta} \mid \mathcal{D})$, provides all the relevant information about that HR.

The association parameter α allows to assess the relationship of the mean of the latent density with the hazard function, but does not provide a direct link with the ordinal longitudinal variable. To facilitate an interpretation of the association parameter in terms of the ordinal measurements, we propose the following *ad-hoc* procedure:

1. Compute the posterior mean, $\mathrm{E}(\gamma_k \mid \mathcal{D})$, of the cut-points γ_k and construct the posterior intervals $(\mathrm{E}(\gamma_{(k-1)} \mid \mathcal{D}),\ \mathrm{E}(\gamma_k \mid \mathcal{D}))$, $k = 1, \ldots, K$.
2. Define for a given time t a representative value $\widetilde{m}_{kt}$, $k = 1, \ldots, K$ of the mean of each ordinal category in the latent scale as follows
   (a) Compute the median of the posterior distribution (8) in each interval $(\mathrm{E}(\gamma_{(k-1)} \mid \mathcal{D}),\ \mathrm{E}(\gamma_k \mid \mathcal{D}))$, and consider them, $\widetilde{y}_{kt}^*$, as the representative value of the latent variable y^* in each ordinal category.
   (b) For each $\widetilde{y}_{kt}^*$, generate a value $\widetilde{m}_{kt}$ of the latent mean according to the general formulation (2) $y^* = m + \epsilon$, or equivalently $m = y^* - \epsilon$, where ϵ is a random error with logistic distribution $\mathrm{Lo}(0, 1)$.

3. Approximate the conditional HR, given α, of an individual in the ordinal category k having the event *versus* an individual in category k' at time t as $\mathrm{e}^{\alpha(\widetilde{m}_{kt}-\widetilde{m}_{k't})}$.
4. Compute the posterior distribution of the approximate HR in step 3 from the marginal posterior distribution of α as $\int \mathrm{e}^{\alpha(\widetilde{m}_{kt}-\widetilde{m}_{k't})}\,\pi(\alpha \mid \mathcal{D})\,\mathrm{d}\alpha$.

3.3. Prediction

Bayesian reasoning approaches the estimation of the conditional survival probability of an individual i with a given history provided by their baseline covariates and longitudinal follow-up $\mathcal{Y}_{in_i}$ (which guarantees that their survival time is greater than the time, t_{in_i}, of their last longitudinal measurement) through the posterior distribution $\pi(P(T_i \geq t \mid T_i > t_{in_i}, \mathcal{Y}_{in_i}, \boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{b}_i) \mid \mathcal{D})$, $t > t_{in_i}$. This posterior contains all relevant information about the location and variability of this conditional survival probability over time. In particular, its posterior mean can be more easily computed as

$$\begin{aligned} P(T_i \geq t \mid T_i > t_{in_i}, \mathcal{Y}_{in_i}, \boldsymbol{x}_i, \mathcal{D}) &= \int P(T_i \geq t \mid T_i > t_{in_i}, \mathcal{Y}_{in_i}, \boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{b}_i)\, \pi(\boldsymbol{\theta}, \boldsymbol{b}_i \mid \mathcal{D}, \mathcal{Y}_{in_i}, T_i > t_{in_i})\, \mathrm{d}(\boldsymbol{\theta}, \boldsymbol{b}_i) \\ &= \int P(T_i \geq t \mid T_i > t_{in_i}, \boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{b}_i)\, \pi(\boldsymbol{\theta}, \boldsymbol{b}_i \mid \mathcal{D}, \mathcal{Y}_{in_i}, T_i > t_{in_i})\, \mathrm{d}(\boldsymbol{\theta}, \boldsymbol{b}_i), \quad t > t_{in_i}, \end{aligned} \tag{10}$$

where the conditional probability $P(T_i \geq t \mid T_i > t_{in_i}, \mathcal{Y}_{in_i}, \boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{b}_i)$ does not depend on the particular longitudinal trajectory, $\mathcal{Y}_{in_i}$, as a result of the induced independence of the shared effects joint model, and $\pi(\boldsymbol{\theta}, \boldsymbol{b}_i \mid \mathcal{D}, \mathcal{Y}_{in_i}, T_i > t_{in_i})$ is the marginal posterior distribution of the common parameters' and random effects' vector for individual i, given $\mathcal{Y}_{in_i}$ and $T_i > t_{in_i}$.

We could also approach prediction of a future longitudinal measurement of an individual in the study [44]. The posterior predictive distribution of a new longitudinal measurement y_{i,n_i+1} at the time t_{i,n_i+1} of a future scheduled appointment for individual i with covariates $\boldsymbol{x}_i$ and longitudinal ordinal history $\mathcal{Y}_{in_i}$ is given by

$$\begin{aligned} P(y_{i,n_i+1} = D_k \mid T_i > t_{i,n_i+1}, \mathcal{Y}_{in_i}, \boldsymbol{x}_i, \mathcal{D}) &= P(y^*_{i,n_i+1} \in (\gamma_{k-1}, \gamma_k] \mid T_i > t_{i,n_i+1}, \mathcal{Y}_{in_i}, \boldsymbol{x}_i, \mathcal{D}) \\ &= \int P(y^*_{i,n_i+1} \in (\gamma_{k-1}, \gamma_k] \mid T_i > t_{i,n_i+1}, \mathcal{Y}_{in_i}, \boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{b}_i)\ \pi(\boldsymbol{\theta}, \boldsymbol{b}_i \mid \mathcal{D}, \mathcal{Y}_{in_i}, T_i > t_{i,n_i+1})\, \mathrm{d}(\boldsymbol{\theta}, \boldsymbol{b}_i) \\ &= \int P(y^*_{i,n_i+1} \in (\gamma_{k-1}, \gamma_k] \mid \boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{b}_i)\, \pi(\boldsymbol{\theta}, \boldsymbol{b}_i \mid \mathcal{D}, \mathcal{Y}_{in_i}, T_i > t_{i,n_i+1})\, \mathrm{d}(\boldsymbol{\theta}, \boldsymbol{b}_i) \end{aligned} \tag{11}$$

where

$$P(y^*_{i,n_i+1} \in (\gamma_{k-1}, \gamma_k] \mid \boldsymbol{x}_i, \theta, \boldsymbol{b}_i) = \frac{\mathrm{e}^{\gamma_k - m_{i,n_i+1}} - \mathrm{e}^{\gamma_{k-1} - m_{i,n_i+1}}}{(1+\mathrm{e}^{\gamma_k - m_{i,n_i+1}})(1+\mathrm{e}^{\gamma_{k-1} - m_{i,n_i+1}})}$$

is obtained from (1), with $m_{i,n_i+1} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t_{i,n_i+1}$. The conditional probability $P(y^*_{i,n_i+1} \in (\gamma_{k-1}, \gamma_k] \mid T_i > t_{i,n_i+1}, \mathcal{Y}_i, \boldsymbol{x}_i, \boldsymbol{\theta}, \boldsymbol{b}_i)$ is independent on the survival history, $T_i > t_{i,n_i+1}$, as a consequence of the shared random effects joint model. Note also that the different longitudinal measurements of the same individual are independent given $(\boldsymbol{\theta}, \boldsymbol{b}_i)$.

The previous two posteriors both apply to individuals in the study and to individuals of the population that could be involved in the study in the future. In this framework, some discussion about the posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{b} \mid \mathcal{D}, \mathcal{Y}_{in_i}, T_i > t_{in_i})$ becomes necessary. If the interest concentrates on a specific individual in the study, such as individual i, for whom we want to estimate the conditional probability (10) or the predictive distribution (11) from the current data $\mathcal{D}$, the information provided by $\mathcal{Y}_i$ and $T_i > t_{in_i}$ is already included in $\mathcal{D}$. Consequently, $\pi(\theta, \boldsymbol{b} \mid \mathcal{D}, \mathcal{Y}_{in_i}, T_i > t_{in_i}) = \pi(\theta, \boldsymbol{b} \mid \mathcal{D})$. If the interest focuses on sequentially estimate (10) or/and predict (11) as a result of their follow-up, we would need to sequentially update the current posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{b} \mid \mathcal{D})$ with

all that new relevant follow-up information, in particular new longitudinal measurements and the updated survival time.

Dynamic posterior estimation and prediction for individuals of the population who have not participated in the study are also possible. Let consider now a new subject i' who initially has not participated in the inferential process and enters the study at time $a_{i'}$ with given values $x_{i'}$ of the baseline covariates. The posterior distribution of their unconditional survival probability is $\pi(P(T_{i'} \geq t \mid T_{i'} > a_{i'}, \boldsymbol{x}_{i'}, \boldsymbol{\theta}, \boldsymbol{b}_{i'}) \mid \mathcal{D})$ with posterior mean

$$\begin{aligned} P(T_{i'} \geq t \mid T_{i'} > a_{i'}, \boldsymbol{x}_{i'}, \mathcal{D}) &= \int P(T_{i'} \geq t \mid T_{i'} > a_{i'}, \boldsymbol{x}_{i'}, \boldsymbol{\theta}, \boldsymbol{b}_{i'})\, \pi(\boldsymbol{\theta}, \boldsymbol{b}_{i'} \mid \mathcal{D}, T_{i'} > a_{i'})\, \mathrm{d}(\boldsymbol{\theta}, \boldsymbol{b}_{i'}) \\ &= \int \frac{P(T_{i'} \geq t \mid \boldsymbol{x}_{i'}, \boldsymbol{\theta}, \boldsymbol{b}_{i'})}{P(T_{i'} > a_{i'} \mid \boldsymbol{x}_{i'}, \boldsymbol{\theta}, \boldsymbol{b}_{i'})}\, \pi(\boldsymbol{\theta}, \boldsymbol{b}_{i'} \mid \mathcal{D}, T_{i'} > a_{i'})\, \mathrm{d}(\boldsymbol{\theta}, \boldsymbol{b}_{i'}), \end{aligned} \tag{12}$$

Prediction of their first longitudinal measurement planned at a fixed time $t_{i'1} > a_{i'}$ is

$$\begin{aligned} P(y_{i',1} = D_k \mid T_{i'} > a_{i'}, \boldsymbol{x}_{i'}, \mathcal{D}) &= P(y^*_{i',1} \in (\gamma_{k-1}, \gamma_k] \mid T_{i'} > a_{i'}, \boldsymbol{x}_{i'}, \mathcal{D}) \\ &= \int P(y^*_{i',1} \in (\gamma_{k-1}, \gamma_k] \mid T_{i'} > a_{i'}, \boldsymbol{x}_{i'}, \boldsymbol{\theta}, \boldsymbol{b}_{i'})\ \pi(\boldsymbol{\theta}, \boldsymbol{b}_{i'} \mid \mathcal{D}, T_{i'} > a_{i'})\, \mathrm{d}(\boldsymbol{\theta}, \boldsymbol{b}_{i'}) \\ &= \int P(y^*_{i',1} \in (\gamma_{k-1}, \gamma_k] \mid \boldsymbol{x}_{i'}, \boldsymbol{\theta}, \boldsymbol{b}_{i'})\ \pi(\boldsymbol{\theta}, \boldsymbol{b}_{i'} \mid \mathcal{D}, T_{i'} > a_{i'})\, \mathrm{d}(\boldsymbol{\theta}, \boldsymbol{b}_{i'}) \end{aligned} \tag{13}$$

If as a consequence of the follow-up of this individual we would like to compute posterior probabilities of the type (10) and/or (11), the subsequent marginal posterior distribution will come from the joint posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{b}_{i'} \mid \mathcal{D}, \mathcal{Y}_{i'n_{i'}}, T_{i'} > t_{i'n_{i'}})$, which includes the common parameters and hyperparameters $\boldsymbol{\theta}$, the vector of random effects $\boldsymbol{b}$ associated to the original individuals in the study and those of that new individual considered, $b_{i'}$.

From a Bayesian point of view, the incorporation of sequential information from an individual who is already involved in the study or from the follow-up of a future subject implies the need of sequentially update the posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{b} \mid \mathcal{D})$. In the case of studies based on samples with large sample size, we would expect a minimal change in the estimation of the common parameters but possibly not in the subject specific random effects. The process of updating a posterior distribution for which we only have an approximate random sample and not an analytical distribution is conceptually easy but not so in practice. The main tools to carry out this computational process are based on Sequential MCMC methods [45, 46] and although are beyond the scope of this paper are a current aim of our research team. Rizopoulos [47] proposes, as an approximation of the subsequent posterior distribution, updating the specific random effects associated with individuals. In particular, the author uses empirical Bayesian estimation for the random effects and an asymptotic normal distribution, based on maximum likelihood estimation, for the common population parameters. Taylor *et al.* [48] also separately update the vector of random effects by using a quick MCMC based on a prior distribution for the population parameters coming from the marginal posterior $\pi(\boldsymbol{\theta} \mid \mathcal{D})$.

4. Joining longitudinal breast density and age at breast cancer detection

Let $\{D_1, D_2, D_3, D_4\}$ denote the set of BI-RADS breast density categories $\{a, b, c, d\}$, which represent low, medium, high, and very high density, respectively, y_{ij} the breast density category of woman i, $i = 1, \ldots, n$, at time t_{ij} (age $50 + t_{ij}$), $j = 1, \ldots, n_i$ and y^*_{ij} her subsequent underlying continuous latent value.

Following (3) the connection between both processes is

$$y_{ij} = D_k \ \Leftrightarrow y^*_{ij} \in (\gamma_{k-1},\ \gamma_k],\ \ k = 1, 2, 3, 4,$$

where $-\infty = \gamma_0 < \gamma_1 < \gamma_2 < \gamma_3 < \gamma_4 = \infty$ are unknown cut-points, with $\gamma_2 = 0$, and $\text{Lo}(m_{ij},\, s = 1)$ represents the corresponding logistic distribution for y^*_{ij}.

Considering the evidence of a decreasing trend of breast density with age and a linear trajectory for the longitudinal latent breast density of woman i

$$(y^*_i(t) \mid m_{it}) = m_{it} + \epsilon_{it} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t + \epsilon_{it}, \tag{14}$$

where $(\beta_0, \beta_1)^T$ and $(b_{i0}, b_{i1})^T$ are the fixed (population) and random effects (individual) for the intercept and the slope term, respectively, and ϵ_{it} the error term. Random effects $(\boldsymbol{b}_0, \boldsymbol{b}_1)^T$, where $\boldsymbol{b}_0 = (b_{01}, \ldots, b_{0n})^T$ and $\boldsymbol{b}_1 = (b_{11}, \ldots, b_{1n})^T$, are assumed conditionally i.i.d. with $(b_{i0} \mid \sigma_0) \sim \text{N}(0, \sigma_0)$ and $(b_{i1} \mid \sigma_1) \sim \text{N}(0, \sigma_1)$.

The hazard function of age at breast cancer diagnosis is defined in terms of the left truncated relative risk regression model

$$h_i(t \mid \mathbf{x}_i, \theta_{is}, t^*_i > a_i) = h_0(t \mid \lambda, \eta_0) \exp\{\eta_1 Famhist_i + \eta_2 Brstproc_i + \alpha\, m_i(t)\},\ t > a_i, \tag{15}$$

where $h_0(t \mid \lambda, \eta_0) = \lambda t^{\lambda-1}\, \text{e}^{\eta_0}$ is the baseline risk function of a Weibull distribution, $\text{We}(\lambda, \text{e}^{\eta_0})$; *family history of breast cancer (Famhist)* and *prior breast procedures (Brstproc)* are dychotomous baseline covariates with associated coefficients η_1 and η_2 respectively; α assesses the effect of the individual trajectory of breast density on breast cancer risk in terms of the latent breast density mean; and a_i is the age over 50 at which woman i enters the screening program thus providing the left truncated time [43].

We assumed prior independence among all the parameters and hyperparameters as a default specification and, with the aim of giving all inferential prominence to the data, we elicited wide proper prior distributions. For the parameters in the longitudinal submodel we followed Lunn *et al.* [30] except for the standard deviations. In particular, we selected $N(0, 100)$ for the β's regression coefficients. The ordinal constraint for the cutpoints of the latent scale, $-\infty < \gamma_1 < \gamma_2 = 0 < \gamma_3 < \gamma_4 = \infty$, was expressed by truncating the subsequent prior distributions in the appropriate parametric subspace

$$\gamma_1 \sim N\left(-\log(3), \sigma_{\gamma_1} = 100\right) \mathcal{I}(-\infty, \gamma_2 = 0),$$
$$\gamma_3 \sim N\left(\log(3), \sigma_{\gamma_3} = 100\right) \mathcal{I}(\gamma_2 = 0, \infty),$$

where $\mathcal{I}(-)$ is the indicator function. Prior means for γ_1 and γ_3 respectively correspond to the first and third quartiles of a logistic distribution $\text{Lo}(0, 1)$ in order to provide the same prior probability to each response category. For the standard deviations σ_0 and σ_1 we choose a uniform distribution, $\text{Un}(0, 10)$. In the case of the survival submodel we selected $N(0, 100)$ for the η's regression coefficients as well as for the association coefficient α, and a gamma distribution $\text{Ga}(1, 1)$ for the parameter λ of the baseline hazard function because it mimics a constant baseline hazard function [49].

4.1. Posterior distribution

The posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{b} \mid \mathcal{D})$ was computed in terms of the hierarchical modelling (4) and approximated using MCMC simulation methods through the `JAGS` software [50]. In particular, we run three MCMC chains with 100 000 iterations, 10 000 of which were used for the burn-in period. The chains were thinned by only storing every 270th iteration in order to reduce autocorrelation in the saved sample. Trace plots of the simulated values of the three chains appear overlapping one another indicating stabilization. Convergence of the chains to the posterior distribution was assessed through the potential scale reduction factor, $\hat{R}$, and the effective number of independent simulation draws, n_{eff}, [51] and [52], respectively. $\hat{R}$ compares the within-chain variance to the estimated variance of the posterior distribution in such a way that $\hat{R}$ values near 1 indicate that the simulated process has reached the

posterior distribution. n_{eff} deals with the level of autocorrelation of the chains simulated values, so that $n_{eff} > 100$ indicates that sufficient MCMC samples have been obtained.

Table 2 summarizes the approximate MCMC random sample from the marginal posterior distribution $\pi(\boldsymbol{\theta} \mid \mathcal{D})$ through the mean, median, standard deviation, 2.5% and 97.5% percentiles. The last column of Table 2 contains the probability that the corresponding parameter is positive: a 0.5 probability would indicate that a positive value of the parameter is equally likely that a negative one, hence indicating little relevance of the corresponding variable (given the remaining covariates). This is not the case for the parameters of our model with probabilities that show a clear *preference* for being above or under zero.

TABLE 2 AROUND HERE

The marginal posterior distribution associated to the population intercept β_0 and slope β_1 of the mean of the latent breast density clearly states that both are negative, $P(\beta_0 < 0 \mid \mathcal{D}) = P(\beta_1 < 0 \mid \mathcal{D}) = 1$, indicating decreasing values over time of the true latent breast density, and therefore a higher probability of being in the lower breast density categories with age. The variability associated with the random intercept is important, $\mathrm{E}(\sigma_0 \mid \mathcal{D}) = 2.6067$, as a sign of high population heterogeneity with regard to initial breast density. In contrast, there is small variability in the subject-specific slopes, $\mathrm{E}(\sigma_1 \mid \mathcal{D}) = 0.0053$, which denotes that subject-specific trajectories of the true latent breast density do not differ much from the population trend. The estimation of the cut-points γ_1 and γ_3 is very stable and accurate. The posterior means of the coefficients associated to the baseline covariates, *family history of breast cancer* and *prior breast procedures*, 0.6227 and 0.4535, respectively, indicate an increase of risk of breast cancer detection when women have one or both of these risk factors. These values are consistent with the ones reported in the literature. The strength of the association between the breast density and age at breast cancer diagnosis is assessed through their posterior expectation $E(\alpha \mid \mathcal{D})$=0.149 and 95% credible interval (0.1089, 0.1887). In addition, the posterior probability 1 for that coefficient being positive provides strong support on the connection between breast density and breast cancer risk.

4.2. Probabilities associated to breast density BI-RADS categories

Figure 2 (top) shows the posterior mean and 95% credible interval of the posterior distribution $\pi(P(y_{it} = D_k \mid \theta) \mid \mathcal{D})$ associated to each BI-RADS category for a generic woman in the study. Probabilities associated to category BI-RADS b are always higher than 0.5, and grow slightly with age. Probabilities for categories a, c, and d are initially very similar, but categories c and d decrease with age following a similar pattern while category a increases (see Table **??** in Appendix). The information provided by credible intervals is very valuable thus indicating high precision in the estimated means.

FIGURE 2 AROUND HERE

Figure 2 (bottom) shows a violin plot (a combination of a kernel density plot and a boxplot) of the posterior marginal distribution of the latent breast density at ages 50, 55, 60, 65 and 70. The four categories of the ordinal breast density are marked with regard to the posterior mean of the cut points γ_1 and γ_3, and $\gamma_2 = 0$. The visual comparison between real and latent results is very interesting. We clearly appreciate that the posterior marginal distribution of the latent breast density tends towards lower values with age. In addition, the bottom tail of the distributions increase with age in detriment of the top tail thus indicating the general decreasing of breast density with age.

4.3. Assessment of the impact of the study variables on breast cancer risk

Relevant HRs arise from the combination of baseline covariate categories. Figure 3 shows the posterior distribution of the HRs of a breast cancer diagnosis for *family history of breast cancer*, *prior breast procedures*, and both risk factors, with posterior means 1.864, 1.574, and 2.934, respectively. The marginal effects of each covariate are relevant, with posterior probabilities 0.998 and 1.000 associated to HR values greater than 1 for *family history of breast cancer* and *prior breast procedures*, respectively.

FIGURE 3 AROUND HERE

Following the *ad-hoc* procedure presented in subsection 3.2, Figure 4 shows the posterior mean and 95% credible intervals with regard to age of the approximate HR of a breast cancer diagnosis for women with breast densities b, c, and d compared to women with the same covariate values and breast density a. Changes in breast density from category a towards more dense categories have a strong effect on breast cancer risk. We observe posterior means of the HR around 4 for category d versus a, and HRs greater than 1 (around 1.7 and 2.6) when comparing categories b and c versus a, respectively. A gently wavy behaviour for the posterior distributions and credible intervals of the HRs with respect to age can be appreciated, as a consequence of the simulation of the logistic error in the procedure.

FIGURE 4 AROUND HERE

4.4. Prediction

Figure 5 shows the posterior mean (10) and 95% credible band for four of the women without cancer in Figure 1, at the end of follow-up. As expected, breast cancer-free probabilities are very high and decrease with age. It is worth noting the narrowness of the bands corresponding to women with a higher number of density measurements, possibly due to the precision of the random effects estimates. Women with high breast density values seem to have lower breast cancer-free survival probabilities. However, we must also consider the baseline risk factors. Thus, disease-free survival is higher for woman 942, who has a stable very high breast density, than for woman 9 672, who experiences a decrease in density. This result can be attributed to presence of prior breast procedures in woman 9 672 and absence of them in woman 942. Both women do not have family history of breast cancer.

FIGURE 5 AROUND HERE

FIGURE 6 AROUND HERE

Figure 6 shows the posterior predictive distribution of the ordinal breast density categories for the women in Figure 5. We appreciate a great variability among the predicted BI-RADS trajectories, and for most of the selected women, category b has the highest probabilities over age. But, for woman 5318 category a is always the most prevalent with an increasing trend over age, followed by categories b, c and d. These results are in line with the observed breast density trajectory for this woman: three breast density measurements with values b, a and a at late ages 65, 67 and 69, respectively. In contrast, for woman 942 category d is always the most prevalent with a decreasing trend over age. These results are also consistent with the observed breast density trajectory for this woman, stable with very high breast density at relatively late ages 61, 63, 65, 67, and 69.

5. Discussion

We propose a Bayesian joint model that combines the information provided by a longitudinal ordinal process and a left-truncated time-to-event outcome. The joint density of both processes is approached through a shared-parameter model which generates a structure of association and conditional independence between both outcomes by means of a vector of common random effects.

We chose a latent variable formulation for the longitudinal process which translated the ordinal scale to the framework of linear mixed models, with a logistic distribution for the measurement error. The latent variable approach facilitates the computational implementation of the model but introduces complexity in the interpretation of results. We assume a logistic distribution for the latent variable which implies the logistic link for cumulative probabilities. Other models might be also appropriate. The most usual alternatives are the normal and the extreme value distributions which result in the probit and complementary log-log regression links, respectively. It is widely accepted that probit and the logit links produce similar results. This also occurs in our study (results not shown), where we have implemented the probit link to assess the robustness of the model. This is not the case for the extreme value distribution which, unlike the logistic and normal ones, is not symmetrical.

We consider that the cut-points that relate the ordinal and latent variables are common for all individuals and time. This assumption may produce some stiffness in the longitudinal model. Thus, individual or time specific cut-points might endow of more flexibility to the longitudinal latent variable at the expense of a more complex model. Dealing with more than four categories in the ordinal variable is not straightforward. One of the reasons for this is that one or both of the endpoints of the truncated intervals in which the marginal prior distribution of each unknown cut-point is defined can be also unknown [53]. The estimation of these models involve important computational issues in the MCMC sampling which have provided many discussion and proposals, such as hybrid Metropolis-Hasting algorithms to sample from the subsequent posterior distribution [54, 55].

Robustness is a major statistical concern in Hierarchical Bayesian models because it can be affected by an inappropriate choice of hyperprior distributions for hyperparameters. We have tested the sensitivity of the model using other prior specifications for the hyperprior distribution of the random effects scale parameter. In particular, we have considered a wide uniform distribution, Un(0, 100), as an alternative to the elicited Un(0,10) in the paper and inverse-gamma hyperdistributions, IGa(0.01, 0.01) and IGa(0.001, 0.001), due to their common use in Bayesian applications. All of them provided almost identical results (not shown in the paper), possibly because of the large sample size.

Our proposal could be applied to a variety of real problems devoted to analyze time-to-event outcomes with temporal ordinal endogeneous covariates. We explore the role of death prior to breast cancer diagnosis as a competing risk. The cumulative incidences estimated with the Kaplan-Meier method or the competing risks with cause-specific hazards approach are very similar, even at older ages (See Figure *A*1 in the Supplementary file). Therefore, even though the censoring due to the competing risk "death" was informative, it would hardly affect our estimates. Event times have been modeled in terms of relative risk models with left-truncation as a corrective mechanism for the delayed entry bias. Left truncation is common in observational studies of risk factors, where not all the participants enter the study at time zero. We select the Weibull distribution as baseline risk function because it is a traditional model for survival data with a great flexibility in representing different types of risk. The exploration of more sophisticated baseline risk functions, which include multimodality and heavy tailed distributions, in the area of Bayesian joint modeling is a relevant subject with strong connections with the specification of prior distributions. See [56] for a detailed explanation of piecewise constant hazard models and Gamma processes. In addition, the latent linear mixed model is a flexible model that can accommodate heterogeneous trajectories, from linear to complex functions. This is the case of linear models expressed in terms of spline bases to accommodate non-linear profiles [57].

We have used our joint model for analyzing the relationship between mammographic breast density and breast cancer risk in women attending a public screening program. A linear subject-specific trajectory of the latent variable is included in a relative risks survival model together with two of the most known breast cancer risk factors, *family history of breast cancer* and *previous breast procedures*. Our joint model for breast cancer and breast density is a good starting point that provides results consistent with the literature [16, 18, 19, 21]. They are the basis for a rationale for extending the model and assessing its adequacy and accuracy. Evaluating the ability of breast density to predict time to breast cancer diagnosis, under our joint model, by means of calibration and discrimination measures [58] is a major concern of our research. The discriminative power of our model should offset its complexity. Discrimination measures based on Receiver Operating Characteristic (ROC) curves are commonly used for assessing predictive accuracy. We are currently exploring a general latent variable approach that could be appropriate.

In contrast to studies published to date, our study is the first to have used the complete longitudinal history of breast density for assessing breast cancer risk over time, at population and individual level. Potential benefits of the proposed joint model include obtaining individual predictions of time-free of breast cancer at age $u > t$, given the observed responses up to age t, and also individual longitudinal predictions of future breast density values. Thus, a joint model similar to that shown here could be used for surveillance of breast cancer risk over time, for scheduling screening exams based on individual dynamic predictions, and also in discussing prevention strategies for those at high risk [16].

Acknowledgements

This paper was partially supported by the research grants Combination and Propagation of Uncertainties (ComPro_UN, MTM2013-42323-P), Statistical methods for clinical trials, complex censoring schemes and integrative omics data analysis (MTM2015-64465-C2-1-R), and Women participation in decisions and strategies on early detection of breast cancer (PI14/00113) from the Spanish Ministry of Economy and Competitiveness, ACOMP/2015/202 from the Generalitat Valenciana, and GRBIO-2014-SGR464 and GRAES-2014-SGR978 from the Generalitat de Catalunya.

We thank Dimitris Rizopoulos and Arindom Chakraborty for their advice on joint modeling. We are also indebted to the research group Grup de Recerca en Anàlisi Estadística de la Supervivència (GRASS), specially to Carles Serrat, for their support and fruitful discussions. We also thank Núria Torà and Marina Pont for their work on data management, and JP Glutting for review and editing. We wish to acknowledge three anonymous referees and the Associate Editor for their useful comments, that improved very much the original version of the paper.

We also would like to thank the staff of the Cancer Screening Office and the radiologists of the BCEDP-VOE, for their valuable help in obtaining the data we analyzed. And, last but not least, we appreciate all women that kindly answered the questionnaire.

References

1. Onega T, Beaber E, Sprague B, Barlow W, Haas J, Tosteson A, Schnall M, Armstrong K, Schapira M, Geller B, *et al.*. Breast cancer screening in an era of personalized regimens: a conceptual model and National Cancer Institute initiative for risk-based and preference-based approaches at a population level. *Cancer* 2014; **19**:2955–2964.
2. Vilaprinyo E, Forné C, Carles M, Sala M, Pla R, Castells X, Domingo L, Rue M, Interval Cancer Study Group (INCA). Cost-effectiveness and harm-benefit analyses of risk-based screening strategies for breast cancer. *PLoS One* 2014; **9**:e86 858.
3. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* 1989; **81**:1879–1886.

4. Gail MH. Personalized estimates of breast cancer risk in clinical practice and public health. *Stat. Med.* 2011; **30**(10):1090–1104, doi:10.1002/sim.4187. URL http://dx.doi.org/10.1002/sim.4187.
5. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. *Stat. Med.* 2004; **23**(7):1111–1130, doi:10.1002/sim.1668. URL http://dx.doi.org/10.1002/sim.1668.
6. Anothaisintawee T, Teerawattananon Y, Wiratkapun C, Kasamesup V, Thakkinstian A. Risk prediction models of breast cancer: a systematic review of model performances. *Breast Cancer Res. Treat.* 2012; **133**(1):1–10, doi:10.1007/s10549-011-1853-z. URL http://dx.doi.org/10.1007/s10549-011-1853-z.
7. Tice JA, Cummings SR, Ziv E, Kerlikowske K. Mammographic breast density and the Gail model for breast cancer prediction in a screening population. *Breast Cancer Res. Treat.* 2005; **94**:115–122.
8. Chen J, Pee D, Ayyagari R, Graubard B, Schairer C, Byrne C, Benichou J, Gail MH. Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. *J. Natl. Cancer Inst.* 2006; **98**:1215–1226.
9. Barlow WE, White E, Ballard-Barbash R, Vacek PM, Titus-Ernstoff L, Carney PA, Tice JA, Buist DS, Geller BM, Rosenberg R, *et al.*. Prospective breast cancer risk prediction model for women undergoing screening mammography. *J. Natl. Cancer Inst.* 2006; **98**:1204–1214.
10. Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K. Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Ann. Intern. Med.* 2008; **148**:337–347.
11. Byrne C, Schairer C, Wolfe JN, Parekh N, Salane M, Brinton LA, Hoover RN, Haile R. Mammographic features and breast cancer risk: effects with time, age, and menopause status. *J. Natl. Cancer Inst.* 1995; **87**:1622–1629.
12. Boyd NF, Lockwood GA, Byng JW, Tritchler DL, Yaffe MJ. Mammographic densities and breast cancer risk. *Cancer Epidemiol. Biomarkers Prev.* 1998; **7**:1133–44.
13. Vacek PM, Geller BM. A prospective study of breast cancer risk using routine mammographic breast density measurements. *Cancer Epidemiol. Biomarkers Prev.* 2004; **13**:715–722.
14. Vachon CM, van Gils CH, Sellers TA, Ghosh K, Pruthi S, Brandt KR, Pankratz VS. Mammographic density, breast cancer risk and risk prediction. *Breast Cancer Res.* 2007; **9**:217.
15. van Gils CH, Hendriks JHCL, Holland R, Karssemeijer N, Otten JDM, Straatman H, Verbeek ALM. Changes in mammographic breast density and concomitant changes in breast cancer risk. *Eur. J. Cancer Prev.* 1999; **8**:509–515.
16. Kerlikowske K, Ichikawa L, Miglioretti DL, Buist DS, Vacek PM, Smith-Bindman R, Yankaskas B, Carney PA, Ballard-Barbash R, Consortium NIoHBCS. Longitudinal measurement of clinical mammographic breast density to improve estimation of breast cancer risk. *J. Natl. Cancer Inst.* 2007; **99**:386–395.
17. Vachon CM, Pankratz VS, Scott CG, Maloney SD, Ghosh K, Brandt KR, Milanese T, Carston MJ, Sellers TA. Longitudinal trends in mammographic percent density and breast cancer risk. *Cancer Epidemiol. Biomarkers Prev.* 2007; **16**:921–928.
18. Lokate M, Stellato RK, Veldhuis WB, Peeters PHM, van Gils CH. Age-related changes in mammographic density and breast cancer risk. *Am. J. Epidemiol.* 2013; **178**:101–109.
19. Work ME, Reimers LL, Quante AS, Crew KD, Whiffen A, Beth Terry M. Changes in mammographic density over time in breast cancer cases and women at high risk for breast cancer. *Int. J. Cancer* 2014; **135**:1740–1744.
20. Huo CW, Chew GL, Britt KL, Ingman WV, Henderson MA, Hopper JL, Thompson EW. Mammographic density - a review on the current understanding of its association with breast cancer. *Breast Cancer Res. Treat.* 2014; **144**:479–502.
21. Kerlikowske K, Gard CC, Sprague BL, Tice JA, Miglioretti DL. One versus two breast density measures to predict 5- and 10-year breast cancer risk. *Cancer Epidemiol. Biomarkers Prev.* 2015; **24**(6):889–897, doi:10.1158/1055-9965.EPI-15-0035.
22. Kalbfleisch JD, Prentice RL. *The statistical analysis of failure time data. (Second edition)*. Wiley, 2002.
23. Tsiatis AA, DeGruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *J. Am. Stat. Assoc.* 1995; **90**:27–37.
24. Faucett CL, Thomas DC. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Stat. Med.* 1996; **15**:1663–1685.
25. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error . *Biometrics* 1997; **53**:330–339.
26. Rizopoulos D. *Joint models for longitudinal and time-to-event data with applications in R*. CRC Press, Biostatistics Series: Boca Raton, FL, 2012.
27. Rizopoulos D. JM: An R package for the joint modelling of longitudinal and time-to-event data. *J. Stat. Softw.* 2010; **35**:1–33.
28. Rizopoulos D. *JMbayes: Joint modeling of longitudinal and time-to-event data under a Bayesian approach* 2015. URL http://CRAN.R-project.org/package=JMbayes, r package version 0.7-2.
29. Serrat C, Rué M, Armero C, Piulachs X, Perpiñán H, Forte A, Páez A, Gómez G. Frequentist and bayesian approaches for a joint model for prostate cancer risk and longitudinal prostate-specific antigen data. *Journal of Applied Statistics* 2015; **42**(6):1223–1239.

30. Lunn DJ, Wakefield J, Racine-Poon A. Cumulative logit models for ordinal data: a case study involving allergic rhinitis severity scores. *Stat. Med.* 2001; **20**:2261–2285.
31. Chakraborty A, Das K. Inferences for joint modelling of repeated ordinal scores and time to event data. *Comput. Math. Methods Med.* 2010; **11**:281–295.
32. Li N, Elashoff RM, Li G, Saver J. Joint modeling of longitudinal ordinal data and competing risks survival times and analysis of the NINDS rt-PA stroke trial. *Stat. Med.* 2010; **29**:546–557.
33. Luo S. A Bayesian approach to joint analysis of multivariate longitudinal data and parametric accelerated failure time. *Stat. Med.* 2014; **33**:580–594.
34. Hogan JW, Laird N. Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* 1997; **28**:239–257.
35. Chakraborty A. Bounded influence function based inference in joint modelling of ordinal partial linear model and accelerated failure time model. *Stat. Methods Med. Res.* 2014; **11**, doi:10.1177/0962280214531570.
36. Proust-Lima C, Dartigues JF, Jacqmin-Gadda H. Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: a latent process and latent class approach. *Statistics in Medicine* 2016; **35**:382–398.
37. He B, Luo S. Joint modeling of multivariate longitudinal measurements and survival data with applications to Parkinson disease. *Stat. Methods. Med. Res.* 2013; doi:10.1177/0962280213480877.
38. Liu F, Li Q. A Bayesian model for joint analysis of multivariate repeated measures and time to event data in crossover trials. *Stat. Methods. Med. Res.* 2014; doi:10.1177/0962280213519594.
39. Baré ML, Montes J, Florensa R, Sentís M, Donoso L. Factors related to non-participation in a population-based breast cancer screening programme. *Eur. J. Cancer Prev.* 2003; **12**:487–94.
40. Baré M, Bonfill X, Andreu X. Relationship between the method of detection and prognostic factors for breast cancer in a community with a screening programme. *J. Med. Screen.* 2006; **13**:183–91.
41. Baré M, Sentís M, Galceran J, Ameijide A, Andreu X, Ganau S, Tortajada L, Planas J. Interval breast cancers in a community screening programme: frequency, radiological classification and prognostic factors. *Eur. J. Cancer Prev.* 2008; **17**:414–21.
42. American College of Radiology. *Breast Imaging Reporting and Data System®(BI-RADS®)*. 5th edn., American College of Radiology: Reston, VA, 2013.
43. Uzunoğullari a, Wang JL. Comparison of hazard rate estimators for left truncated and right censored data. *Biometrika* 1992; **79**:297–310.
44. Armero C, Forte A, Perpiñán H, Sanahuja MJ, Agustí S. Bayesian joint modeling for assessing the progression of Chronic Kidney Disease in children. *Stat. Methods. Med. Res.* 2016; doi:10.1177/0962280216628560.
45. Del Moral P, Doucet A, Jasra A. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2006; **68**:411–436.
46. Andrieu C, Doucet A, Holenstein R. Particle Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2010; **72**:269–342.
47. Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* 2011; **67**:819–829.
48. Taylor JMG, Park Y, Ankerst DP, Proust-Lima C, Williams S, Kestin L, Bae K, Pickles T, Sandler H. Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics* 2013; **69**(1):206–213.
49. Guo X, Carlin BP. Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician* 2004; **58**:1–9.
50. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)* 2003; .
51. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**:457–511.
52. Kass RE, Carlin BP, Gelman A, Neal R. Markov Chain Monte Carlo in Practice: A Roundtable Discussion. *The American Statistician* 1998; **52**:93–100.
53. Albert JH, C S. Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association* 1993; **88**:669–679.
54. Cowles M. Accelerating Monte Carlo Markov chain convergence for cumulative-link generalizad linear models. *Statistics and Computing* 1996; **6**:101–111.
55. Nandram B, Chen MH. Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence. *Stat. Med.* 1996; **54**:129–144.
56. Ibrahim JG, Chen MH, Sinha D. *Bayesian Survival Analysis*. New York: Springer, 2001.
57. Andrinopoulou ER, Rizopoulos D, Takkenberg JJM, Lesaffre E. Joint modeling of two longitudinal outcomes and competing risk data. *Stat. Med.* 2014; **33**(18):3167–3178.

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

6. Tables and Figures

Table 1. Baseline risk factors according to breast cancer diagnosis status at the end of follow-up.

	No breast cancer N=13 254		Breast cancer N=431	
Family history of breast cancer				
No	12 539	(94.8%)	388	(90.2%)
Yes	686	(5.2%)	42	(9.8%)
Prior breast procedures				
No	12 318	(92.9%)	374	(86.8%)
Yes	936	(7.1%)	57	(13.2%)
Breast density at first examination (baseline breast density)				
a: Almost entirely fatty	2959	(23.4%)	56	(13.9%)
b: Scattered fibroglandular densities	5353	(42.3%)	138	(34.2%)
c: Heterogeneously dense	2301	(18.2%)	103	(25.6%)
d: Extremely dense	2037	(16.1%)	106	(26.3%)
Breast density at last examination[a] (women with at least two examinations)				
a: Almost entirely fatty	2284	(18.1%)	35	(9.4%)
b: Scattered fibroglandular densities	7957	(63.0%)	201	(54.0%)
c: Heterogeneously dense	1475	(11.7%)	71	(19.1%)
d: Extremely dense	919	(7.3%)	65	(17.5%)

[a] If breast cancer was diagnosed within 6 months following the last mammography, the last breast density considered was the previous one, whenever it was not coincident with the baseline measure.

Table 2. Posterior summaries of the parameters and hyperparameters of the breast cancer joint model.

	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	
λ	1.5366	0.1044	1.3287	1.5387	1.7386	
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

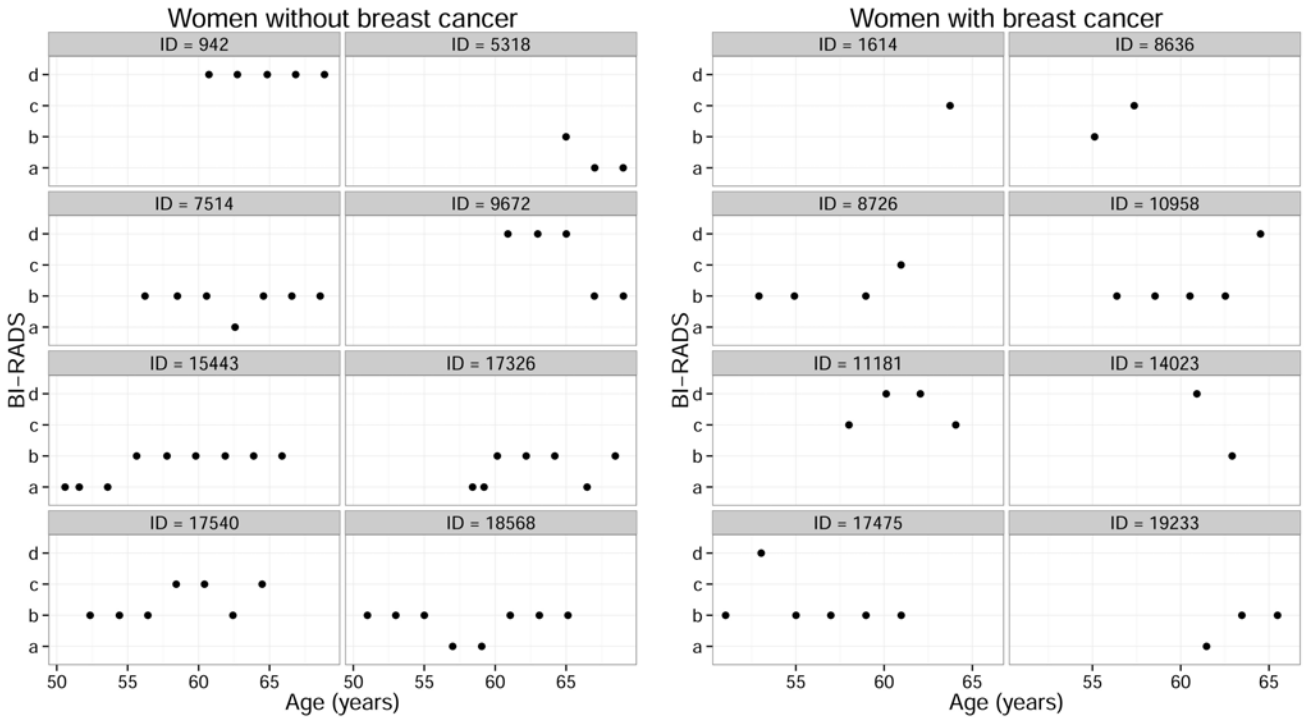

Figure 1. Subject-specific profiles of BI-RADS measures for sixteen randomly selected women. The left panel corresponds to eight women without breast cancer, and the right panel corresponds to eight women with breast cancer.

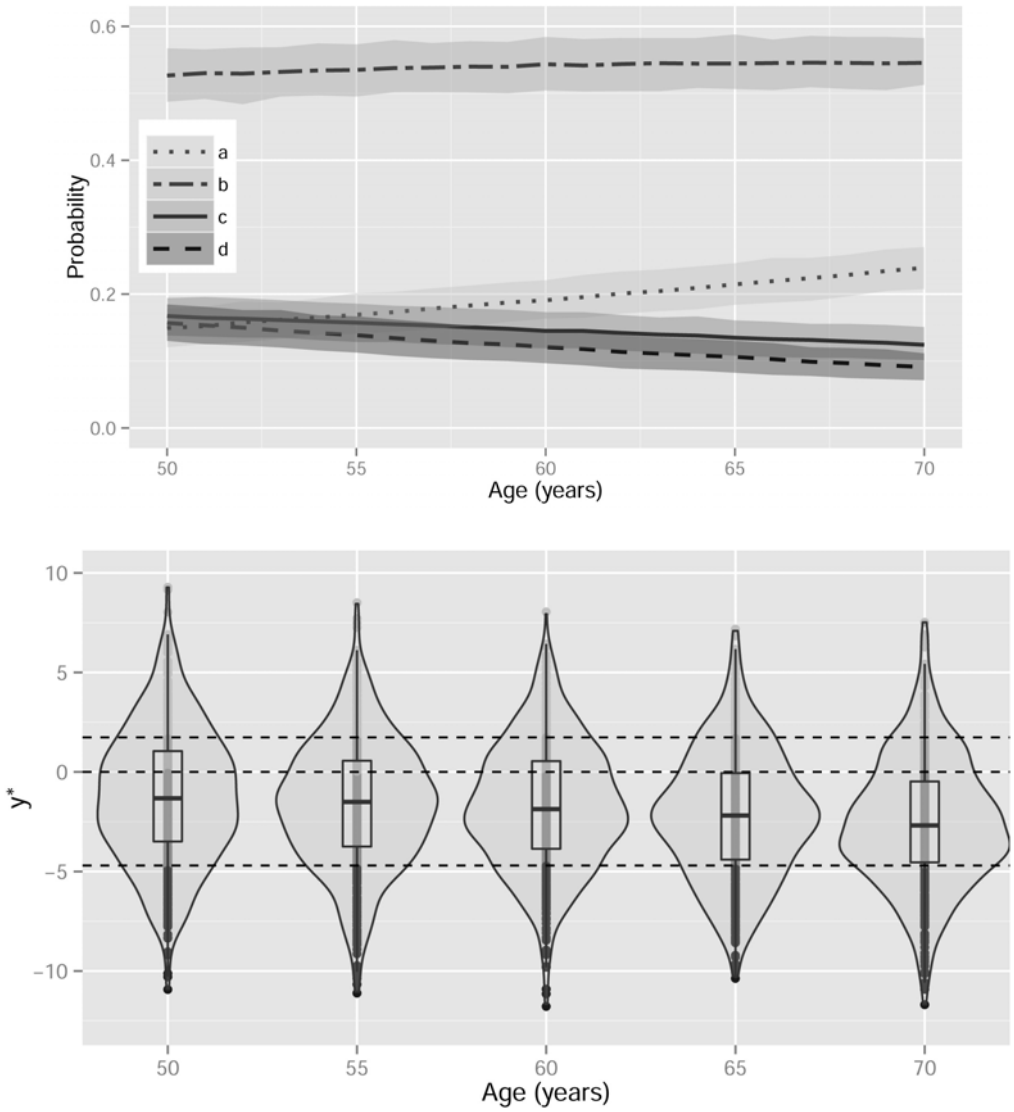

Figure 2. Age-specific population distribution of breast density. Posterior mean and 95% credible interval of the probability associated to each BI-RADS category with respect to age (top) and violin plot of the posterior marginal distribution of the latent breast density of an average woman at ages 50, 55, 60, 65 and 70 (bottom). Horizontal dotted lines represent the posterior mean of the cut-points thus approximately indicating the region of the latent density corresponding to ordinal BD categories *a*, *b*, *c*, and *d* (from bottom to top).

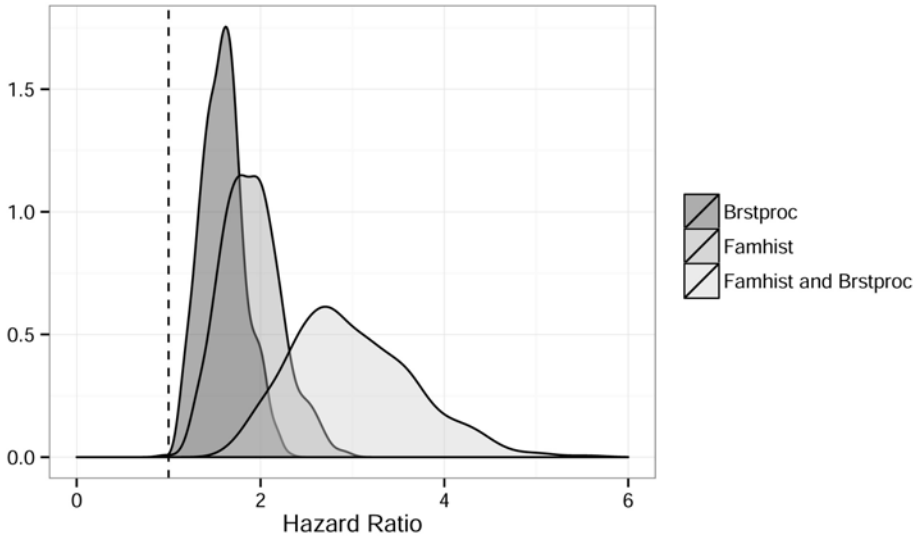

Figure 3. Posterior distribution of the hazard ratios associated to family history of breast cancer and prior breast procedures.

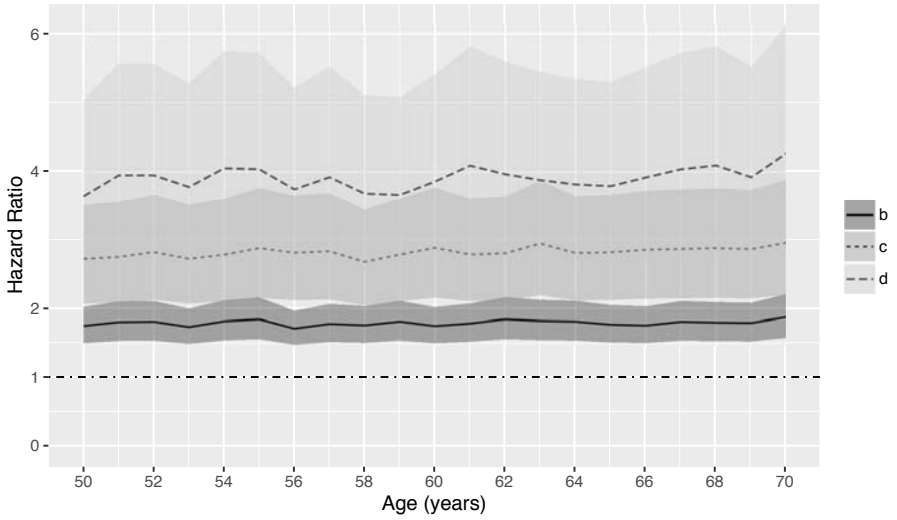

Figure 4. Approximate posterior mean and 95% credible interval with regard to age of the HRs of a BC diagnosis for women with breast density b, c and d as compared with women with the same covariates and BD measurement a.

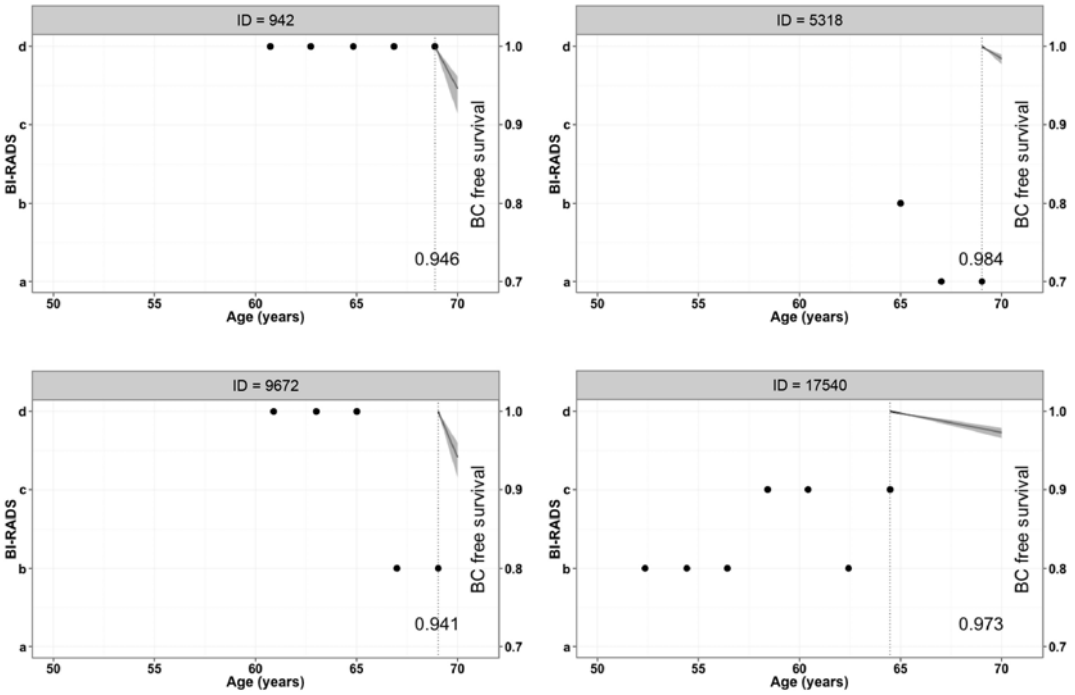

Figure 5. Posterior mean and 95% credible band of the probability of a breast cancer-free diagnosis for women with IDs 942, 5 318, 9 672 and 17 540 without breast cancer at the end of the follow-up. The value of the probability at the lower right of each graphic is the subsequent posterior mean at 70 years.

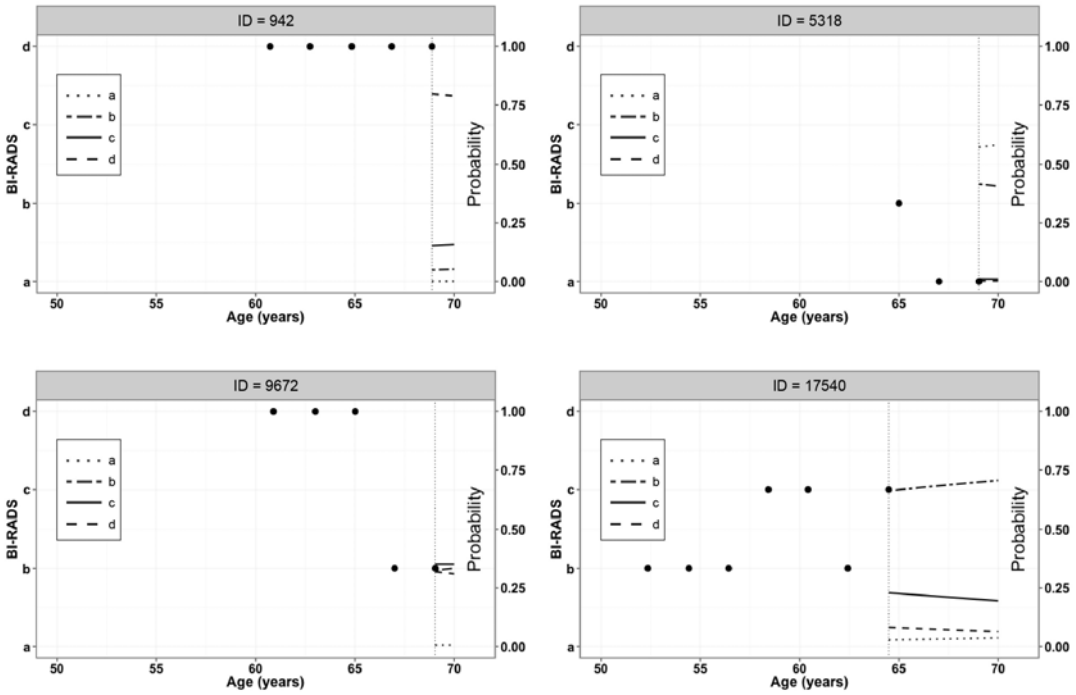

Figure 6. Posterior predicted mean of the breast density in the BI-RADS scale over age for women with IDs with IDs 942, 5 318, 9 672 and 17 540 without breast cancer at the end of the follow-up.