ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2015)
64, *Part* 5, *pp.* 731–753

Modelling short- and long-term characteristics of follicle stimulating hormone as predictors of severe hot flashes in the Penn Ovarian Aging Study

Bei Jiang and Naisyin Wang,

University of Michigan, Ann Arbor, USA

Mary D. Sammel

University of Pennsylvania, Philadelphia, USA

and Michael R. Elliott

University of Michigan, Ann Arbor, USA

[Received October 2013. Final revision November 2014]

Summary. The Penn Ovarian Aging Study tracked a population-based sample of 436 women aged 35–47 years to determine associations between reproductive hormone levels and menopausal symptoms. We develop a joint modelling method that uses the individual level longitudinal measurements of follicle stimulating hormone (FSH) to predict the risk of severe hot flashes in a manner that distinguishes long-term trends of the mean trajectory, cumulative changes captured by the derivative of mean trajectory and short-term residual variability. Our method allows the potential effects of longitudinal trajectories on the health risks to vary and accumulate over time. We further utilize the proposed methods to narrow the critical time windows of increased health risks. We find that high residual variation of FSH is a strong predictor of hot flash risk, and that the high cumulative changes of the FSH mean trajectories in the 52.5–55-year age range also provides evidence of increased risk over that of short-term FSH residual variation by itself.

Keywords: Bayesian penalized *B*-splines; Functional regression; Increased risk window; Joint modelling; Robust inference; Short- and long-term characteristics

1. Introduction

The Penn Ovarian Aging Study (Freeman *et al.*, 2011) is a longitudinal study of a population-based sample of 436 women aged 35–47 years selected via random-digit dialling in Philadelphia County, Pennsylvania, during 1996–1997, and followed biannually through to 2010. The study goal is to explore the associations between reproductive hormone levels and symptoms in the transition to the menopause. Changes in hormone levels alter menstrual bleeding patterns, culminating in the cessation of menstruation, which marks the end of a woman's reproductive years. This course of events, which is termed perimenopause, can last for 5 or more years and coincides for a majority of women with the development of hot flashes, night sweats and other symptoms. The extent to which these symptoms are associated with reproductive hormone levels, trends over time or fluctuations is not well understood. This lack of understanding is

Address for correspondence: Bei Jiang, Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109-2029, USA.
E-mail: beijiang@umich.edu

0035–9254/15/64731

due in part to limited prospectively collected data and is also due to limitations in our ability to model various aspects of this dynamic process.

In this paper, we focus on the relationship between follicle stimulating hormone (FSH) and presence and severity of hot flashes. FSH stimulates folliculogenesis, which is an important factor in ovarian aging; thus there has been interest in using longitudinal FSH information to define menopause transition stages as discussed by Sowers *et al.* (2008). Whereas elevated FSH is an indicator of ovarian aging, Sowers *et al.* (2008) found that both acceleration and deceleration periods in FSH levels were predictive of time to final menstrual period, suggesting that features other than just the level of FSH may give rise to menopausal symptoms. Exploratory analysis of the FSH data in the Penn Ovarian Aging Study shows both acute and gradual increase periods of FSH levels in the population level and has given rise to clinical questions about whether it is the rate of increase in FSH that signals risks of severe menopausal symptoms. Moreover, identifying critical ages when women are at increased risk for symptoms would be helpful for making decisions about treatment. To understand better the association between trajectories of FSH and risk of severe menopausal symptoms in perimenopausal women, we develop a joint modelling method that

(a) makes efficient use of the available information in the longitudinal FSH trajectories, by including long-term trends captured by the mean trajectories or the time varying change rates in the long-term trends that are captured by the derivatives of the mean trajectories as potential predictors in the primary outcome submodel while also adjusting for the previously identified effect of the short-term variation that is captured by the variance of the residuals (Jiang *et al.*, 2014) and
(b) allows selection of the longitudinal FSH features within certain clinically relevant time windows to predict the risk of hot flash severities in the primary outcome submodel, where the effects outside this particular time window are assumed to be negligible.

Joint models of longitudinal and health outcome data have been extensively developed in the literature. The early developments of such joint models were mainly motivated by human immunodeficiency virus and acquired immune deficiency syndrome clinical trials and cancer research and often focused on summarizing mean longitudinal trends as time varying predictors in survival outcome models (Tsiatis *et al.* (1995), Muthén and Shedden (1999), Wang and Taylor (2001), Law *et al.* (2002), Song *et al.* (2002), Brown and Ibrahim (2003a, b), Ibrahim *et al.* (2004) and Yu *et al.* (2008), among many others). In our work, we extend the existing joint modelling approaches and shift the focus to relating scalar response and functional predictors in a functional data analysis (FDA) paradigm. Our modelling strategies are motivated by the need to account properly for three key features of the FSH trajectories in the longitudinal submodel: non-linear trajectories that are observed at unequally spaced time points, short-term elevated variation, which is shown by the residual variance, and the heterogeneous nature between individuals, which is shown by the mixture components in both the mean trajectory and the residual variance. Briefly, our work brings together advanced statistical ideas including FDA, robust and semiparametric inference, and joint longitudinal and outcome modelling in novel ways.

Unlike the typical FDA practice to smooth each individual trajectory independently of one another, we formulate a robust semiparametric mixed effect model for all trajectories, where we simultaneously model both the underlying mean and the residual variance of the longitudinal FSH trajectories. We consider the Bayesian penalized spline approach by Lang and Brezger (2004), which is a Bayesian version of the penalized splines proposed by Eilers and Marx (1996), to estimate the underlying mean FSH trajectories. In contrast with fully parametric splines, penalized splines are not as sensitive to the exact number and location of the knots as long as enough knots are being used, since 'unnecessary' knots will be smoothed away by shrinking random effects towards 0. This feature enhances the flexibility to accommodate individual curve fitting of FSH values when these subject level fitted curves may differ from each other. Examples of applications of penalized B-splines for longitudinal data include Durbán *et al.* (2005), who modelled the individual heights of children suffering from acute lymphoblastic leukaemia from a clinical trial conducted at the Dana Farber Cancer Institute, and Chen and Wang (2011), who considered modelling longitudinal systolic blood pressure data from the Framingham Heart Study. For the residual variance, instead of treating it as a nuisance parameter as many others did, we follow Elliott *et al.* (2012) and Jiang *et al.* (2014) to model the within-subject residual variance in the FSH trajectories and study its prediction ability in the primary outcome submodel. Finally, considering the bimodal nature in the FSH trajectories as suggested in Jiang *et al.* (2014), which are also shown in Figs 3 and 7 in Sections 3 and 4 respectively, we allow for mixtures for both mean trajectories and residual variances to reflect early or late rising patterns in the FSH mean trajectories, crossed with high or low level of short-term variation patterns. The structure assumed nicely reflects the heterogeneity features in the FSH observations. Besides modelling individual trajectory via spline fitting, we extend the normal errors assumptions of Jiang *et al.* (2014) by allowing for heavier tailed t-distributions for residual errors to avoid the potential influence of outlying observations.

In the primary outcome submodel, while also adjusting for the effect of the residual variance, we treat the smooth mean trajectories that are estimated from the longitudinal submodel, or the corresponding derivatives as functional predictors linked to the risk of hot flash severities through an FDA regression model in the sense of Ramsay and Dalzell (1991) and James (2002) among many others. This modelling strategy implicitly allows the effects of FSH histories (i.e. FSH values up to a particular time point) or the time varying change rates of FSH histories that are represented by functional coefficient curves to be time varying and cumulative over time. To estimate the functional coefficient curves, we also propose to use the Bayesian penalized spline approach by Lang and Brezger (2004). In addition to the desirable semiparametric features that were mentioned above, the Bayesian penalized spline approach also allows for simultaneous evaluation of the uncertainty of the estimated functional coefficient curves by providing pointwise Bayesian credible intervals, which lead to identification of critical time windows of increased risk of health outcomes of interest, whereas such intervals are typically obtained by bootstrap methods in frequentist FDA regression. To the best of our knowledge, such a modelling strategy has not been considered in the joint modelling literature. Instead, most of the joint modelling developments have focused on using

(a) a summary of important features in the longitudinal trajectories, such as the random effects and latent classes, or
(b) the last available 'true' value as a time-dependent covariate, with the earlier values being considered irrelevant to the outcome of interest. In the context of joint modelling of continuous longitudinal data and a binary outcome, Jiang *et al.* (2014) contrasted the use of random-effects and latent classes approaches and discussed how to utilize the information that they jointly provide to take advantage of each approach fully. Thorough reviews of the joint modelling of continuous longitudinal data and and time-to-event outcomes have been given by Tsiatis and Davidian (2004), Ibrahim *et al.* (2010) and Rizopoulos (2012).

The rest of this paper is organized as follows. In Section 2, we provide the statistical modelling, inference and model checking procedures that are needed to conduct the proposed analysis of the Penn Ovarian Aging Study data. In Section 3, we present the key features in the data, which have motivated the modelling and methodology strategies that are given in Section 2, as well as how we use these strategies to reach new scientific findings and discoveries in linking severe hot flashes risk to FSH longitudinal features for the Penn Ovarian Aging Study. We conclude with a discussion in Section 4. Algorithms to implement the Gibbs sampler for our proposed models are available from

http://wileyonlinelibrary.com/journal/rss-datasets

2. The model proposed

In this section, we present our joint FDA regression models for the longitudinal FSH levels to predict severity of hot flashes modelled by using ordinal multinomial probit models.

(a) Specifically, the longitudinal submodel for the FSH data is given by

$$Y_{ij}|\mathbf{b}_i = \mu(\mathbf{b}_i; t_{ij}) + \varepsilon_{ij},$$
$$\varepsilon_{ij} \sim t_v(0, \sigma_i^2),$$

which is equivalent to

$$\varepsilon_{ij} \sim N(0, \sigma_i^2/m_{ij})$$

with

$$\begin{aligned} m_{ij} &\sim \text{gamma}(v/2, v/2), \\ \mu(\mathbf{b}_i; t_{ij}) &= \sum_{l=1}^{L} b_{il}\,\phi_l(t_{ij}). \end{aligned} \tag{1}$$

Here Y_{ij} denotes the observed longitudinal FSH values for subject i, $i=1,\ldots,n$, at time t_{ij}, $j=1,\ldots,n_i$, $\mu_i(t) \equiv \mu(\mathbf{b}_i;t)$ denotes the mean of Y_{ij} at time t, and the vector $\boldsymbol{\mu}_i = (\mu(\mathbf{b}_i;t_{i1}),\ldots,\mu(\mathbf{b}_i;t_{in_i}))^{\mathrm{T}}$ defines the mean trajectory or trajectory for subject i, where $\mathbf{b}_i = (b_{i1},\ldots,b_{iL})$ is the vector of the random effects that reflects the subject level trajectory patterns, and $\phi_l(t_{ij})$, $l=1,\ldots,L$, are the B-spline basis functions.

To model the mean trajectory $\boldsymbol{\mu}_i$ flexibly, we use truncated power splines consisting of piecewise polynomials of certain order connected at prespecified knot locations (Ruppert *et al.*, 2003). Given the same order and knot locations, truncated power splines and B-splines are equivalent in the sense that there are unique one-to-one linear transformations between these two sets of spline basis functions (Ruppert *et al.*, 2003), leading to the same fitted values from these two splines in the regression set-up. However, the B-spline is more numerically stable than the truncated power spline because the B-spline basis functions are almost orthogonal whereas the truncated power spline basis functions are not. Therefore, we use B-spline basis functions $\phi_l(t_{ij}) \equiv \phi_{l,d}(t_{ij})$, $l=1,\ldots,L$, of degree $d=3$, where $\phi_{l,3}(t_{ij})$ is obtained by the recursion relation

$$\phi_{l,d}(t_{ij}) = \frac{t_{ij}-\kappa_l}{\kappa_{l+d}-\kappa_l}\phi_{l,d-1}(t_{ij}) + \frac{\kappa_{l+1+d}-t_{ij}}{\kappa_{l+d+1}-\kappa_{l+1}}\phi_{l+1,d-1}(t_{ij})$$

for knots at points $\kappa_1,\ldots,\kappa_{L-d-1}$, where $\phi_{l,0}(t_{ij}) = I(\kappa_l \leqslant t_{ij} \leqslant \kappa_{l+1})$. The number of interior knots is denoted by $J_{\mu(t)}$, such that $\Sigma_{l=1}^{L}\,\phi_l(t)=1$ with $L = J_{\mu(t)}+d+1$. We defer the discussion of the selection of knot points to Section 2.5.

To allow for 'heterogeneity' in the mean trajectory in the sense of growth mixture models (Verbeke and Lesaffre, 1996; Muthén and Shedden, 1999; Jiang *et al.*, 2014), we consider a finite mixture of normal distributions for the random effect $\mathbf{b}_i$:

$$\begin{aligned} D_i &\sim \text{multinomial}(\pi_1^D, \ldots, \pi_{K_D}^D); \\ \mathbf{b}_i = (b_{i1}, \ldots, b_{iL})^{\mathrm{T}} | D_i = d &\sim N(\boldsymbol{\beta}_d, \Sigma_d), \qquad d = 1, \ldots, K_D, \end{aligned} \tag{2}$$

where D_i defines the corresponding latent class membership for the mean trajectory class and $\boldsymbol{\beta}_d = (\beta_{d1}, \ldots, \beta_{dL})^{\mathrm{T}}$. Thus, the fixed effect coefficients β_{dl}, $l = 1, \ldots, L$, determine the shape and also the smoothness of the mean trajectory for the dth latent class, defined as $f_d(t_{ij}) = \Sigma_{l=1}^{L} \beta_{dl} \phi_l(t_{ij})$, $d = 1, \ldots, K_D$. Following Lang and Brezger (2004), we use Gaussian random-walk priors on $\boldsymbol{\beta}_d$ to penalize large differences between coefficients of the adjacent spline basis and therefore control the smoothness of the mean trajectory curve to avoid potential overfitting. The specific prior distributions are given in Section 2.3. The random coefficients b_{il}, $l = 1, \ldots, L$, then capture the individual deviations from the class-specific mean trajectory.

The residual ε_{ij} denotes the deviation of Y_{ij} from the subject-specific mean at t_{ij} and is assumed to follow a Student t-distribution with v degrees of freedom, assuming mean 0 and scale σ_i^2. The value of v is assumed to be known. Thus the variance of Y_{ij} is equal to $\{v/(v-2)\}\sigma_i^2$, which can be interpreted as a measurement of the short-term variability around the mean trajectory $\boldsymbol{\mu}_i$. In the case of $v = \infty$, ε_{ij} is normally distributed with mean 0, variance σ_i^2 and $m_{ij} \equiv 1$. To allow for overdispersion and 'heterogeneity' in the within-subject scale parameter σ_i^2, we assume a mixture of log-normal distributions,

$$\begin{aligned} C_i &\sim \text{multinomial}(\pi_1^C, \ldots, \pi_{K_C}^C); \\ \sigma_i^2 | C_i = c &\sim \text{log-N}(\mu_c, \tau^2), \qquad c = 1, \ldots, K_C, \end{aligned} \tag{3}$$

where C_i defines the corresponding latent class membership for the variance class and we assume that $C_i \perp\!\!\!\perp \bar{D}_i$ so that the common assumption that, for subject i, the mean trajectory $\mu_i(t)$ and the residual ε_{ij} are independent still holds.

(b) The outcome submodel for hot flash severities is defined through an ordinal probit model that assumes that there is a latent continuous variable underlying the observed ordinal outcomes. Specifically, let W_i denote this underlying latent variable. We observe the ordinal outcome $o_i = s$, $s = 0, \ldots, S$, if this latent variable W_i falls between the cut-off γ_s and γ_{s+1}, i.e.

$$o_i = s \Leftrightarrow \gamma_s < W_i \leqslant \gamma_{s+1}$$

where these cut-offs between categories are subject to the common constraint that $-\infty = \gamma_0 \leqslant \gamma_1 \leqslant \ldots \leqslant \gamma_{S+1} = \infty$ with one reference cut-off, usually γ_1, fixed at value 0. Then the distribution of this latent variable W_i is specified conditionally on individual longitudinal mean trajectories and variances as follows:

$$W_i \sim N(\eta_i^W, 1), \qquad \eta_i^W = \alpha_0 + \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\lambda}_0 + \int_T \mu_i(t)\, \theta_0(t)\, \mathrm{d}t, \tag{4}$$

where $\mathbf{x}_i$ is a vector of baseline covariates with associated (constant) parameter $\boldsymbol{\lambda}_0$, and the functional coefficient function $\theta_0(t)$ represents the effect of subject-specific mean trend $\mu_i(t)$ at time t while adjusting for the mean trends at other time points within the time window T. The purpose of considering the integral over the chosen time domain T, i.e. $\int_T \mu_i(t)\, \theta_0(t)\, \mathrm{d}t$, is to identify critical time windows of elevated outcome risks, which have several advantages over simply summing up over the observed time points t_{ij}, $j = 1, \ldots, n$.

First, longitudinal observations often have missing values and can be measured at different time points (known as unbalanced data) and hence summation over the observed time points becomes problematic. Second, $\mu_i(t)$ is a smoothed functional representation of the underlying mean function with the individual level variability 'captured' by σ_i^2. Third, since we have considered a mixed effect model to smooth all individual level curves and hence borrow strength across individuals, we obtain more stable estimates of $\mu_i(t)$ in comparison with smoothing $\mu_i(t)$ individually. Fourth, an integral over a chosen time domain implicitly uses the information at infinite time points within time window T whereas summation uses only the information at finitely observed time points. As in the mean trajectories, we let $\theta_0(t)=\Sigma_{k=1}^{K_0}\tilde{\theta}_{0k}\psi_k^0(t)$ for cubic B-spline basis functions $\boldsymbol{\psi}_k^0(t)$, with $\tilde{\theta}_{0k}$ following a random-walk prior, given in Section 2.3, to avoid overfitting. Given that we express $\mu_i(t)$ by $\mathbf{b}_i^{\mathrm{T}}\phi(t)$ and $\theta_0(t)$ by $\psi^0(t)^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_0$, thus

$$\int_T \mu_i(t)\,\theta_0(t)\,\mathrm{d}t=\int_T \mathbf{b}_i^{\mathrm{T}}\,\phi(t)\,\psi^0(t)^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_0\,\mathrm{d}t=\mathbf{b}_i^{\mathrm{T}}\mathbf{G}_T^0\tilde{\boldsymbol{\theta}}_0,$$

where $\phi(t)$ is a vector of L basis functions chosen to express $\mu_i(t)$ in the longitudinal submodel and $\psi^0(t)$ is a vector of K_0 basis functions; $\mathbf{G}_T^0=\int_T\phi(t)\,\psi^0(t)^{\mathrm{T}}\,\mathrm{d}t$. We can calculate or evaluate numerically $\mathbf{G}_T^0$ for any given spline basis functions and the estimation of unknown parameters in the outcome primary model becomes fully parametric.

Alternatively, one may postulate that the cumulative changes of the individual trajectories are potentially predictive of the outcome of interest. To accommodate such a possibility, we can consider the first derivative of $\mu_i(t)$, i.e. $\mu_i'(t)=\partial\mu_i(t)/\partial t$, as a functional predictor by taking advantage of the nice properties of B-splines of continuity and replace the specification (4) for the outcome model by the alternative form

$$W_i\sim N(\eta_i^W,1),\qquad \eta_i^W=\alpha_1+\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\lambda}_1+\int_T\mu_i'(t)\,\theta_1(t)\,\mathrm{d}t, \tag{5}$$

where, as for $\theta_0(t)$, the functional coefficient function $\theta_1(t)$ can be interpreted as the effect of the derivative of mean trend $\mu_i'(t)$ or the rate of change in $\mu_i'(t)$ at time t while adjusting for the values of $\mu_i'(t)$ at other time points within the time window T. To emphasize the fact that we can use different spline basis functions to express $\theta_1(t)$, we express $\theta_1(t)=\Sigma_{k=1}^{K_1}\tilde{\theta}_{1k}\psi_k^1(t)$ by using a different set of B-spline bases $\psi^1(t)=(\psi_1^1(t),\ldots,\psi_{K_1}^1(t))^{\mathrm{T}}$ and the associated coefficient vector $\tilde{\boldsymbol{\theta}}_1=(\tilde{\theta}_{11},\ldots,\tilde{\theta}_{1K_0})^{\mathrm{T}}$. A penalized approach was used by requiring a random-walk prior on $\tilde{\boldsymbol{\theta}}_1$, i.e. $\tilde{\theta}_{1k}\sim N(\tilde{\theta}_{1k-1},\tau_{\theta_1}^2)$, $k=2,\ldots,K_1$. Similarly, we have

$$\int_T \mu_i'(t)\,\theta_1(t)\,\mathrm{d}t=\int_T \mathbf{b}_i^{\mathrm{T}}\,\phi'(t)\,\psi^1(t)^{\mathrm{T}}\tilde{\theta}_1\,\mathrm{d}t=\mathbf{b}_i^{\mathrm{T}}\mathbf{G}_T^1\tilde{\boldsymbol{\theta}}_1,$$

where $\phi'(t)=\partial\phi(t)/\partial t$ given $\phi(t)$ is a vector of L basis functions chosen to express $\mu_i(t)$ in the longitudinal submodel and $\psi^1(t)$ is a vector of K_1 basis functions; $\mathbf{G}_T^1=\int_T\phi'(t)\,\psi^1(t)^{\mathrm{T}}\,\mathrm{d}t$.

2.1. Likelihood specification

Let $\boldsymbol{\phi}=(\pi_d^D,\boldsymbol{\beta}_d,\boldsymbol{\Sigma}_d,d=1,\ldots,K_D;\pi_c^C,\mu_c,c=1,\ldots,K_C;\tau^2,\alpha_0,\boldsymbol{\lambda}_0,\tilde{\boldsymbol{\theta}}_0,\gamma)$, where we assume that each parameter in $\boldsymbol{\phi}$ has an independent prior distribution, with the joint prior distribution denoted by $\pi(\boldsymbol{\phi})$, and $\mathbf{z}$ includes all unobserved latent variables, i.e. $\mathbf{z}=(\mathbf{b},\boldsymbol{\sigma},\mathbf{C},\mathbf{D})'$. The observed data $\mathbf{x}$ consist of the longitudinal trajectories $\mathbf{y}_1,\ldots,\mathbf{y}_n$ and the observed outomes $o_1,\ldots,o_n$. Then the complete-data likelihood of $\boldsymbol{\phi}$ based on $(\mathbf{x},\mathbf{z})$ is given by

$$
\begin{aligned}
f(\mathbf{x},\mathbf{z}|\phi) \propto \Bigg\{ \prod_{i=1}^{n} & \Bigg(\prod_{d=1}^{K_D} \Bigg[\pi_d^D (2\pi)^{-p/2} |\boldsymbol{\Sigma}_d|^{-1/2} \exp\Bigg\{ -\frac{1}{2} (\mathbf{b}_i - \boldsymbol{\beta}_d)' \boldsymbol{\Sigma}_d^{-1} (\mathbf{b}_i - \boldsymbol{\beta}_d) \Bigg\} \Bigg]^{I(D_i=d)} \Bigg) \\
& \times \prod_{c=1}^{K_C} \Bigg(\pi_c^C (2\pi\tau^2)^{-1/2} \sigma_i^{-2} \exp\Bigg[-\frac{\{\log(\sigma_i^2) - \mu_c\}^2}{2\tau^2} \Bigg] \Bigg)^{I(C_i=c)} \\
& \times \prod_{j=1}^{n_i} p(y_{ij}; v, \mathbf{b}_i, \sigma_i^2) \prod_{s=0}^{S} \{\Phi(\gamma_s - \eta_i^W) - \Phi(\gamma_{s+1} - \eta_i^W)\}^{I(o_i=s)} \Bigg\} \pi(\phi) \qquad (6)
\end{aligned}
$$

where $\Phi(\cdot)$ denotes the cumulative distribution function for the standard normal distribution and

$$
p(y_{ij}; v, \mathbf{b}_i, \sigma_i^2) = \begin{cases} \dfrac{1}{\sqrt{(2\pi\sigma_i^2)}} \exp\left[-\dfrac{\left\{ y_{ij} - \sum\limits_{l=1}^{L} b_{il}\phi_l(t_{ij}) \right\}^2}{2\sigma_i^2} \right] & \text{if } v = \infty, \\ \dfrac{\Gamma\left(\dfrac{v+1}{2}\right)}{\Gamma\left(\dfrac{v}{2}\right)\sqrt{(\pi v \sigma_i^2)}} \left[1 + \dfrac{1}{v} \dfrac{\left\{ y_{ij} - \sum\limits_{l=1}^{L} b_{il}\phi_l(t_{ij}) \right\}^2}{2\sigma_i^2} \right]^{-(v+1)/2} & \text{if } v < \infty. \end{cases}
$$

2.2. *Data augmentation step to impute missing data*

Given the minimum number of available repeatedly measured FSH levels in our final sample (ranging between 6 and 26 per woman), we are limited with regard to the number of knots when choosing cubic B-spline basis functions to express $\mu_i(t)$. To maximize the number of knots that we can consider, we fill in those with fewer than 26 observations based on data augmentation within each iteration of Gibbs sampling (chapter 10 in Little and Rubin (2002)). When assuming a missingness at random (MAR) missing data mechanism, this data augmentation procedure proceeds as follows:

(a) draw $\mathbf{Y}_{\text{mis}}^{(t+1)}$ from $p(\mathbf{Y}_{\text{mis}}|\phi, \mathbf{X}_{\text{obs}})$;
(b) draw $\phi^{(t+1)}$ from $p(\phi|\mathbf{X}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$.

Here ϕ denotes model parameters, $\mathbf{Y}_{\text{mis}}$ denotes the missing longitudinal observations of FSH levels and $\mathbf{X}_{\text{obs}}$ denotes all observed data including observed longitudinal observations and the primary outcome of interest. This simulation leads to draws from the joint distribution of $(\phi, \mathbf{Y}_{\text{mis}})$ given observed data $\mathbf{X}_{\text{obs}}$. Therefore, this procedure leads to the same inference about ϕ as when we focus only on the marginal distribution of ϕ given observed data $\mathbf{X}_{\text{obs}}$. This trick allows us to put in more knots to take advantage fully of the penalized spline approach that is free from knot location selection given a sufficient number of knots.

2.3. *Prior specification*

We propose a fully Bayesian approach to estimate model parameters. For the mixture normal distribution of the random effects, we assume a first-order Gaussian random-walk prior as proposed by Lang and Brezger (2004): $\beta_{dl} \sim N(\beta_{d,l-1}, \tau_{\beta d}^2)$, $l = 2, \ldots, L$, with diffuse prior $\beta_{d1} \sim N(0, 100)$ for the initial coefficient, and $\tau_{\beta d}^2 \sim \text{IG}(1, 0.005)$ to control the smoothness of the fitted curves. We do not impose restrictions on the structure of the variance–covariance matrix for the random effects $\boldsymbol{\Sigma}_d$. To avoid problems with unbounded likelihoods in normal mixture models with unstructured variance–covariance matrices (Day, 1969), we use an empirical Bayes prior proposed by Kass and Natarajan (2006): $\mathbf{\Sigma}_d \sim$ inverse-Wishart(df$=r, \mathbf{\Lambda}$), where $\mathbf{\Lambda} = r\{\Sigma_{i=1}^{n} \widehat{\mathrm{cov}}(\tilde{\mathbf{b}}_i)^{-1}/n\}^{-1}$, where $\tilde{\mathbf{b}}_i$ is given by the ordinary least squares estimator of $\mathbf{b}_i$ for subject i, and r is the dimension of $\mathbf{b}_i$.

For the mixture log-normal distribution for the residual variances, we used diffuse priors: $\mu_c \sim N(0, v)$, $\tau^2 \sim \mathrm{IG}(a, b)$ with $v = 1000$ and $a = b = 0.001$. For the class membership probabilities, we assume a conjugate Dirichlet$(4, \ldots, 4)$ distribution on both $\boldsymbol{\pi}^C = (\pi_1^C, \ldots, \pi_{K_C}^C)$ and $\boldsymbol{\pi}^D = (\pi_1^D, \ldots, \pi_{K_D}^D)$ (Frühwirth-Schnatter, 2006); this is equivalent to assuming *a priori* four observations in each class, avoiding the existence of empty classes.

Lastly, in the probit submodel we assign independent priors $N(0, 9/4)$ for the α_0 and every element of $\boldsymbol{\lambda}_0$; for the coefficients that are associated with functional coefficient function $\theta_0(t)$, $\tilde{\boldsymbol{\theta}}_0 = (\tilde{\theta}_{01}, \ldots, \tilde{\theta}_{0K_0})'$, similarly we use a first-order Gaussian random-walk prior, i.e. $\tilde{\theta}_{0k} \sim N(\tilde{\theta}_{0k-1}, \tau_{\tilde{\theta}_0}^2)$, $k = 2, \ldots, K_0$, with $\tilde{\theta}_{01} \sim N(0, 9/4)$ and $\tau_{\tilde{\theta}_0}^2 \sim \mathrm{IG}(1, 0.005)$, where the prior variance 9/4 is chosen to bound the probabilities of $o_i = s$, $s = 0, \ldots, S$, to be away from 0 and 1 (Garrett and Zeger, 2000; Elliott, 2007; Neelon *et al.*, 2011). We put flat uniform priors on γ_s for $s \notin \{0, 1, S+1\}$, i.e. $\gamma_s \sim \mathrm{uniform}(-\infty, \infty)$.

2.4. Posterior computation

Gibbs sampling is used to obtain draws from the corresponding posterior distributions. For $(\alpha_0, \lambda_0, \tilde{\boldsymbol{\theta}} | \mathbf{b}, \boldsymbol{\sigma}, \mathbf{o})$ we use the Albert and Chib (1993) data augmentation method for probit regression models. The draws of $(\sigma_i^2 | C_i, \mu_c, \boldsymbol{\gamma}, \mathbf{b}_i, o_i, W_i, \{y_{ij}\}_j)$ for $i = 1, \ldots, n$ are obtained by the inverse cumulative distribution method. The exact specification of all priors and Markov chain Monte Carlo (MCMC) sampling procedures are provided in the Web-based supporting materials.

For each model, we ran three chains of 100000 iterations from diverse starting points, discarding the first 50000 as burn-in and retaining every 10th draw to reduce auto-correlation. Gelman–Rubin statistics $\sqrt{\hat{R}}$ (Gelman *et al.*, 2003) (the square root of the total variance to within-chain variance ratio) were used to assess the convergence of the MCMC chains. For the population level parameters, the maximum $\sqrt{\hat{R}} = 1.030$ for models assuming fewer than three classes, and, when assuming three classes for either the mean trajectory or the variance class, the maximum $\sqrt{\hat{R}} = 1.184$. For the well-documented issue of 'label switching' in finite mixture modelling (Redner and Walker, 1984), various solutions have been proposed, including the relabelling algorithms by Stephens (2000), Jasra *et al.* (2005) and Rodríguez and Walker (2012). We applied the post-processing relabelling algorithm by Stephens (2000), which considers all possible permutations of class assignments at each iteration of the Gibbs sampler and chooses the one which minimizes the Kullback–Leibler divergence of the estimated *versus* true probabilities of class membership, thus maximizing the posterior probability so that the labelling of classes was consistent with the previous assignments. We post-process the MCMC chains by using Stephens's algorithm to 'untangle' the draws for model parameters.

All the calculations were performed by calling standalone C++ code in R, developed by using an open source C++ library for statistical computation, the Scythe statistical library (Pemstein *et al.*, 2007), which is available for free download from `http://scythe.wustl.edu`.

2.5. Choice of the number of classes and number of knots in penalized splines

We consider the deviance information criterion DIC, which was proposed by Spiegelhalter *et al.* (2002), both to select the number of components for the latent classes and to choose the number of knots in the penalized splines. DIC uses the discrepancy between the posterior mean of the deviance $\overline{D(\phi)} = E_\phi[-2 \log\{f(\mathbf{x}|\phi)|\mathbf{x}\}]$ and the deviance evaluated at the posterior mean $D(\bar{\phi}) = -2\log[f\{\mathbf{x}|E(\phi|\mathbf{x})\}]$ to estimate the effective number of degrees of freedom in the model p_D. DIC is then given by the analogue of the Akaike information criterion AIC:

$$\mathrm{DIC}(\mathbf{x}) = D(\bar{\phi}) + p_D = 2\,\overline{D(\phi)} - D(\bar{\phi}) = -4\,E_{\phi}[\log\{f(\mathbf{x}|\phi)|\mathbf{x}\}] + 2\log[f\{\mathbf{x}|E(\phi|\mathbf{x})\}].$$

In our setting, $f(\mathbf{x}|\phi)$, where $\mathbf{x} = (\mathbf{y}_{\mathrm{obs}}, \mathbf{o})'$, consisting of the fully observed data, is not available in closed form; instead we use the approach that was outlined in Celeux *et al.* (2006) to obtain

$$\mathrm{DIC}(\mathbf{x}) = E_{\mathbf{z}}\{\mathrm{DIC}(\mathbf{x},\mathbf{z})\} = -4\,E_{\mathbf{z},\phi}[\log\{f(\mathbf{x},\mathbf{z}|\phi)|\mathbf{x}\}] + 2\,E_{\mathbf{z}}(\log[f\{\mathbf{x},\mathbf{z}|E_{\phi}(\phi|\mathbf{x},\mathbf{z})\}]|\mathbf{x})$$

where integration over the latent variables $\mathbf{z} = (\mathbf{b}, \boldsymbol{\sigma}, \mathbf{C}, \mathbf{D}, \mathbf{y}_{\mathrm{mis}})'$ is obtained via numerical methods.

2.6. Goodness-of-fit evaluation

We assessed the model goodness of fit to the data in two ways: pivotal discrepancy measures (PDMs) (Johnson, 2007; Yuan and Johnson, 2012), which yield an overall goodness-of-fit measure for the longitudinal predictor component, and the area under the receiver operating characteristic (ROC) curve, AUC, which is a goodness-of-fit measure focusing on prediction of the ordinal outcome of interest.

In contrast with more general posterior predictive distribution measures of fit (Gelman *et al.*, 1996), PDMs are defined to depend only on the data and the model parameters with a known distribution. If the model is correctly specified, the PDMs that are evaluated at the true parameter value and the draws from the posterior distribution should have the same sampling distribution. Therefore, model adequacy can be tested by treating the PDMs as a test statistic to obtain a uniformly distributed p-value. However, the posterior samples of PDMs are not independent as they are all derived from the observed data (Johnson, 2004); thus p-value calculation is difficult. Instead, Johnson (2007) and Yuan and Johnson (2012) focused on the upper bound of p-values, and hence the upper bound of a p-value being less than 0.05 definitely provided strong evidence of model inadequacy.

To examine the fit of the longitudinal trajectories, we consider subject level PDMs where, for subject i, we let $D_i = \Sigma_{j=1}^{n_i} m_{ij}\{y_{ij} - \mu_i(t_{ij})\}^2/\sigma_i^2$. When the assumed longitudinal submodel defined in expression (1) is correct, the PDM D_i is $\chi^2_{n_i-1}$ distributed. We use repeated posterior draws to obtain the sampling distribution of PDMs and compute the upper bounds of the p-values based on the ordered statistics of PDMs by using the approach by Yuan and Johnson (2012).

Second, we assessed the prediction of the outcome by using ROC curves, in particular the area under the ROC curve, AUC. ROC curves plot the true positive rate TP *versus* the false positive rate FP for all possible cut-offs based on predicted $P(o_i = s) = \Phi(\mathbf{Z}_i'\eta)$ obtained from expression (4) for $s = 0, \ldots, S$. The ROC curve and AUC were computed at each MCMC iteration by using the ROCR package in R (Sing *et al.*, 2005). The ROC is computed by ordering the observations $(i) = 1, \ldots, n$ so that $\hat{P}(o_{(i)} = 1) \geqslant \hat{P}(o_{(i+1)} = 1)$, computing change points $c = 2, \ldots, n_c$, $n_c \leqslant n$, where the observations change from positive to negative (i.e. $o_{(c-1)} = 1$ and $o_{(c)} = 0$), and plotting $\Sigma_{(i)=1}^{c}(1 - o_{(i)})/\Sigma_{(i)=1}^{n}(1 - o_{(i)})$ on the horizontal axis against $\Sigma_{(i)=1}^{c} o_{(i)}/\Sigma_{(i)=1}^{n} o_{(i)}$ on the vertical axis. The area under the ROC curve is then computed by using a trapezoidal approximation. The posterior mean AUC is calculated as the average AUCs across MCMC iterations. To obtain the posterior mean and the pointwise 95% credible interval of the ROC curve, we choose 250 points equally spaced along the FP-axis and take the vertical average or 95% quantiles of TPs at the 250 chosen points. This approach was referred to as vertical averaging of ROC curves at fixed FP-rates by Fawcett (2006).

3. Predicting risks of hot flash severities from longitudinal follicle stimulating hormone data

In the Penn Ovarian Aging Study, participating women had their hormone measures taken annually during the early follicular phase of a menstrual cycle for two sequential menstrual cycles, with up to 13 years of follow-up available at the time of our analysis. We focus our analysis on the 234 women who

(a) had not experienced hot flash symptoms at baseline,
(b) had baseline measurements of body mass index BMI and smoking status (0 or 1) that are to be included as baseline covariates in the outcome submodel and
(c) had at least six measurements of FSH levels.

Among this restricted sample, 144 (62%) women had fully participated in the study. Among the remaining 90 (38%) women, 42 of them dropped out after at least six assessment periods, whereas 48 of them had either sporadically skipped the assessments or dropped out of the study at the very beginning but came back to the study later when increased incentives were offered. Nelson *et al.* (2004) examined the factors that may predict the participation after six assessment periods and concluded that dropping out was probably random; for those who came back to the study because of increased incentives, their initial dropout was probably due to personal reasons that were not symptom related. FSH values could be missing because of laboratory errors or missing blood samples (7.1%), which are likely to be missing at random. Further, FSH values were censored if a woman

(a) was pregnant and/or breastfeeding (0.3%),
(b) had a hysterectomy with or without oophorectomy (3.0%),
(c) was taking exogenous hormone replacement therapy (1.4%),
(d) was taking oral contraceptives (2.5%),
(e) was taking cancer treatment medications (0.6%) or
(f) was taking other oestrogen (0.2%) during the follow-up.

The average number of available FSH levels per woman is 18.7 (range: 6–26) in our final sample.

We let y_{ij} denote the natural log-transformed FSH levels, i.e. log(FSH), and o_i denote the ordinal outcome of interest, i.e. the severity of hot flashes (0, 1 and 2), defined as $o_i = 0$ if never had severe hot flashes (severity score less than 2 throughout the follow-up period), $o_i = 1$ if had severe but not more severe hot flashes (severity score at least once equal to 2 or once equal to 3 that occurred before 40 years old) and $o_i = 2$ if had more severe hot flashes (severity score at least once equal to 3 after 40 years old). In our final sample, 117 (50%) never experienced any severe hot flashes during follow-up (severity score 0), 80 (34%) had a severity score of 1 and 37 (16%) had a severity score of 2. Since most women start to experience menopausal-related symptoms between the age of 45 and 50 years and reach the menopause by the age of 55 years, we consider $T = [45, 55]$ as a potential risk time window in our analysis for the effect of changes in FSH levels on risk of severe hot flashes.

We use the longitudinal submodel defined in expression (1) to describe longitudinal measured FSH and the outcome model defined in expression (4) to relate long- and short-term FSH characteristics to the risk of severe hot flashes. Preliminary analysis suggested using cubic B-spline basis functions with 1–3 inner knots to express $\mu_i(t_{ij})$ and cubic B-spline basis functions with 1–5 inner knots to express the functional coefficient function $\theta_0(t)$. Thus we consider models with one, three or five knots, putting these knots at the equally spaced quantiles of the

Table 1. DIC from various joint models for the analysis of the Penn Ovarian Aging study data, assuming normal, t_7- and t_4-distribution for the longitudinal submodel and using $\mu_i(t)$, $i=1,\ldots,n$, within the time window $T=[45,55]$ as a functional predictor in the primary outcome submodel†

Model	*Results for* $K_C=1$			*Results for* $K_C=2$			*Results for* $K_C=3$		
	$K_D=1$	$K_D=2$	$K_D=3$	$K_D=1$	$K_D=2$	$K_D=3$	$K_D=1$	$K_D=2$	$K_D=3$
Normal									
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=1$	11439.0	11477.2	11492.9	*11333.6*	11369.1	11399.3	11511.6	11545.1	11560.7
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=3$	11437.5	11487.9	11501.8	*11327.7*	11364.9	11386.7	11506.8	11542.9	11561.5
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=5$	11435.0	11480.5	11493.3	*11330.6*	11369.1	11385.7	11500.7	11552.2	11574.9
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=1$	11923.4	11912.4	11924.6	11809.6	11788.7	11798.9	12000.1	11977.5	11984.4
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=3$	11923.8	11901.3	11915.5	11807.0	11803.5	11799.8	11995.0	11971.6	11997.1
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=5$	11924.7	11892.4	11919.2	11799.7	11788.2	11801.4	11993.1	11965.6	11991.5
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=1$	12419.3	12400.5	12418.6	12319.9	12308.2	12316.5	12506.2	12489.0	12499.3
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=3$	12421.8	12398.8	12412.5	12317.6	12306.7	12320.6	12506.5	12486.7	12489.2
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=5$	12416.6	12399.3	12409.5	12317.0	12298.1	12307.5	12504.7	12472.7	12485.0
t_4									
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=1$	10335.0	10257.5	10271.0	10303.3	*10215.4*	10246.8	10425.0	10326.3	10347.2
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=3$	10333.2	10255.7	10272.5	10308.8	*10210.8*	10235.5	10419.9	10330.3	10374.1
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=5$	10331.2	10260.0	10273.9	10298.5	*10230.4*	10228.3	10432.3	10322.7	10371.9
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=1$	10831.8	10823.6	10826.4	10803.1	10774.6	10778.2	10947.6	10906.7	10889.1
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=3$	10830.0	10821.0	10833.2	10821.3	10776.0	10812.1	10929.6	10897.9	10934.2
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=5$	10828.0	10818.8	10822.3	10818.0	10780.1	10791.6	10936.8	10914.8	10922.0
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=1$	11280.6	11259.2	11256.8	11287.8	11255.8	11257.5	11406.5	11369.9	11397.4
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=3$	11275.4	11251.5	11256.8	11276.3	11251.4	11271.0	11393.9	11356.3	11382.0
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=5$	11278.3	11250.5	11265.0	11298.1	11253.6	11264.5	11409.9	11381.4	11384.1
t_7									
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=1$	10626.5	10585.0	10606.3	10566.9	*10518.2*	10533.3	10679.8	10603.3	10652.2
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=3$	10624.0	10584.2	10600.6	10567.8	*10511.5*	10532.0	10694.9	10633.9	10648.5
$J_{\mu(t)}=1$, $J_{\theta_0(t)}=5$	10622.5	10579.8	10598.3	10558.1	*10512.0*	10536.6	10670.4	10615.5	10628.7
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=1$	11127.3	11114.8	11125.2	11065.8	11051.9	11067.9	11214.9	11205.2	11201.2
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=3$	11123.7	11116.2	11132.3	11074.7	11062.0	11061.8	11210.6	11195.2	11207.4
$J_{\mu(t)}=2$, $J_{\theta_0(t)}=5$	11126.5	11115.4	11128.0	11069.1	11055.4	11056.6	11225.2	11185.2	11206.9
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=1$	11604.1	11582.4	11585.9	11570.0	11550.0	11544.7	11652.8	11651.3	11661.6
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=3$	11601.5	11577.1	11588.5	11572.0	11541.7	11547.4	11687.8	11644.0	11672.1
$J_{\mu(t)}=3$, $J_{\theta_0(t)}=5$	11600.6	11586.8	11587.9	11569.2	11540.3	11548.9	11672.2	11671.7	11651.9

†Designs with the lowest DIC are given in italics.

distinctly observed ages of these women (Ruppert *et al.*, 2003). This is equivalent to assuming piecewise cubic orthogonal polynomials connected at those chosen knot locations. Next, we consider the number of components for both mean trajectory and variance classes. Previous analysis of fitting mixture distributions for both the random effects and the variances (Jiang *et al.*, 2014) successfully identified one mean trajectory class and two variance classes under the normality assumption for ε_{ij}. However, our current approach assumes a t-distribution for ε_{ij} that will potentially impact the effect of any outliers on estimation of the mean trajectories, which may alter the optimal numbers of components for the mean trajectory and variance classes. With all these considerations, we consider $K_D=1,2,3$ and $K_C=1,2,3$ in our analysis. We attempted to estimate the degrees of freedom ν of the t_ν-distribution by treating it as a true parameter in our model, but we found that its estimation was unstable without the use of a strongly informative prior. Hence we performed a sensitivity analysis, comparing results

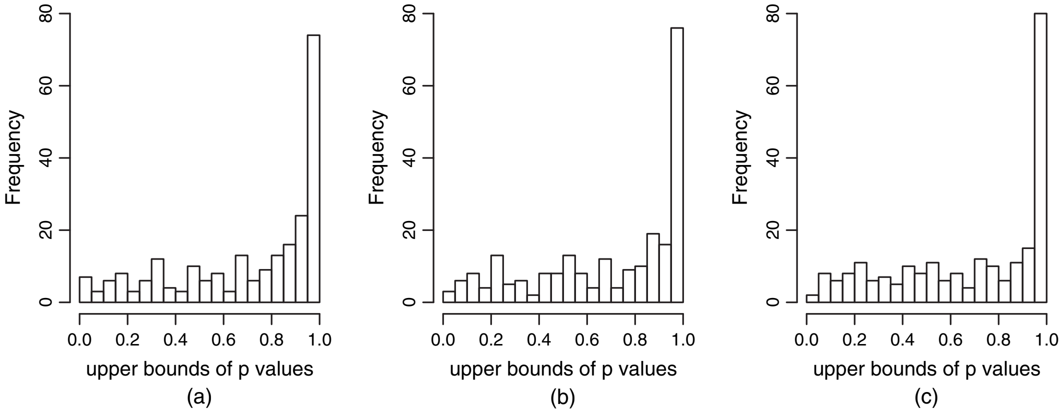


Fig. 1. Upper bounds of p-values based on PDMs for individual trajectories fitted by our best-fitting models with $\mu_i(t)$, $i=1,\ldots,n$, within the time window $T=[45,55]$ years as a functional predictor in the primary outcome submodel: (a) best-fitting normal model; (b) best-fitting t_7-model; (c) best-fitting t_4-model

from a normal model with a submodel with t_4- and t_7-assumptions based on Jeffreys's (1973), page 65, suggestion to replace the normality assumption with a t-distribution with degrees of freedom in the range 4–15. We chose these three scenarios as representative settings to reflect the assumptions of presence of extreme outliers, mild outliers or absence of outliers relative to a normal distribution in the FSH data.

Table 1 presents the DIC-statistics for all models considered: one, two or three latent classes for the mean trajectories and variances, normal, t_7- and t_4-assumptions for the errors in the longitudinal submodel, and one, three or five knots for the longitudinal trajectories or functional varying-coefficient function. In general, DIC suggests that joint models with the t_4-assumption for the longitudinal submodel fit the data better than with t_7 and much better than the normal model. $K_D=K_C=2$ is selected for both the t_4- and the t_7-assumption. Given these selected numbers of components for both the mean trajectory and the variance classes for each model, DIC further suggests that one knot (i.e. $J_{\mu(t)}=1$) at 46.6 years of age for the longitudinal trajectories and three knots (i.e. $J_{\theta_0(t)}=3$) at 41.6, 46.6 and 51.5 years of age for the functional varying-coefficient function offer the best balance between goodness of fit and smoothness under all these three longitudinal submodel assumptions. Thus we shall focus on these three best-fitting models:

(a) best-fitting normal model—$K_D=1$, $K_C=2$ with $J_{\mu(t)}=1$ at 46.6 years of age and $J_{\theta_0(t)}=3$ at 41.6, 46.6 and 51.5 years of age;

(b) best-fitting t_7- and t_4-models—$K_D=K_C=2$ with $J_{\mu(t)}=1$ at 46.6 years of age and $J_{\theta_0(t)}=3$ at 41.6, 46.6 and 51.5 years of age

For these best-fitting models, PDMs also confirmed our previous finding based on model selection criterion DIC that the t_4-model fits the longitudinal FSH trajectories better than the t_7- and normal distribution. Fig. 1 shows the upper bounds of the p-values based on PDMs for longitudinal trajectories fitted by all three final models. If the upper bound of a p-value is less than 0.05, there is strong evidence of inadequate fit. We see that the normal model fits the large majority of subjects well, with seven individual trajectories being considered to have inadequate fit by PDMs. Out of these seven individual trajectories, assuming a t-distribution with 7 degrees of freedom improved the fits of four individual trajectories, leaving three individual trajectories with inadequate fit; among the three individual trajectories, assuming a t-distribution with 4 degrees of freedom resulted in only two individual trajectories with inadequate fit. Fig. 2(a) shows the two trajectories that are considered to have inadequate fits by all three best-fitting models based on PDMs. Fig. 2(b) shows the four trajectories that have upper bounds of p-values less than 0.05 by our best-fitting normal model but upper bounds of p-values greater than 0.05 by both our best-fitting t_7- and t_4-models. Clearly, these plots suggest that t-models with 4 and 7 degrees of freedom show considerably less influence by outlying observations than the normal model and they both have almost identical fits visually. Finally, Fig. 2(c) shows random selected four trajectories that have upper bounds of p-values that are greater than 0.05 by all three of our best-fitting models: the normal and t_7- and t_4-models show very similar fits. Therefore, the inadequate fit of longitudinal FSH trajectories that was identified by PDMs is probably due to these varying degrees of extreme outliers. Although we could consider even smaller degrees of freedom of t-distribution or more heavily tailed distributions for the longitudinal submodel to accommodate these extreme outlying observations, the t-model with either 4 or 7 degrees of freedom already shows almost identical robustness to them and seems to provide a reasonably good fit to more than 99% of the FSH data.

Next, we contrast the estimation results from these models to demonstrate the influence of not appropriately accommodating outlying observations. Fig. 3 presents the mean trajectory components and two variance components that were identified by the three best-fitting models. Consistent with the finding that was reported in Jiang *et al.* (2014), under the normal model assumption, a single-component mean trajectory is favoured by DIC. In contrast, under both the t_7- and the t_4-model assumptions, a two-component mean trajectory is favoured by DIC: the major mean class (86% of women) whose FSH levels begin to increase in their late 40s and the minor mean class (14% of women) with increasing FSH levels starting around age 40 years capturing a proportion of women who might transition into the menopause at an earlier age. The variance class has different meanings under the t- and normal assumptions but in both scenarios measure the short-term variations in FSH levels: according to their magnitudes, both t- and normal models would classify them to either 'low' or 'high' variance classes. On the basis of the posterior estimates of these component-specific parameters given in Table 1 in the Web-based supporting material, we can see more subtle differences in these estimated mixture components under varying assumptions.

Table 2 shows that all three models reach the same broad conclusions: high short-term variability (its effect is represented by θ_3) in the FSH levels is strongly associated with increased risks of more severe hot flashes, smoking (its effect is represented by θ_2) is marginally associated with more severe hot flashes, and there was no association with BMI (its effect is represented by θ_1) or the individual mean trajectories between age 45 and 55 years (its cumulative time varying effect is represented by $\theta_0(t)$). The most dramatic difference between the different degrees-of-freedom models occurs for the estimated functional coefficient $\theta_0(t)$ that captures the cumulative time varying effect of the mean trajectory $\mu_i(t)$. Figs 4(a), 4(b) and 4(c) show the estimated functional coefficient $\theta_0(t)$ by our best-fitting normal, t_7- and t_4-models respectively. The estimated $\theta_0(t)$ under our best-fitting normal model tends to have larger effect size (larger magnitude in $\theta_0(t)$) before age 53 years and an overall wider pointwise 95% credible interval than the estimated $\theta_0(t)$s under our best-fitting t_4- and t_7-models. All three coefficient curves suggest that, when adjusting for the whole history of mean FSH levels over the age range of age 45 to age 55 years, higher mean FSH levels before age 53 years reduce the risk of severe hot flashes, whereas higher mean FSH levels between age 53 and age 55 years increase this risk, but there is no conclusive evidence of a true association between the FSH trajectory histories and the risk of more severe hot flashes.

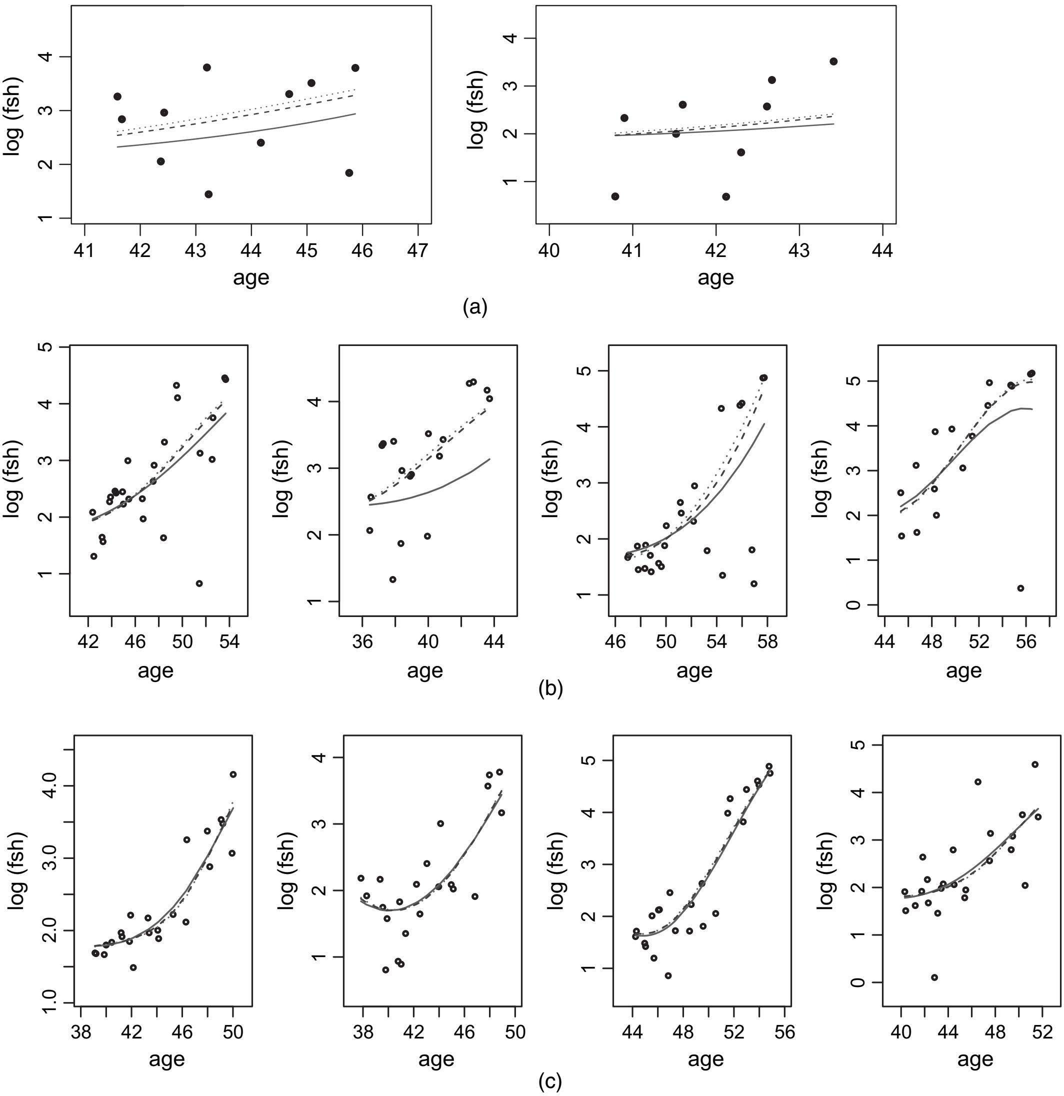
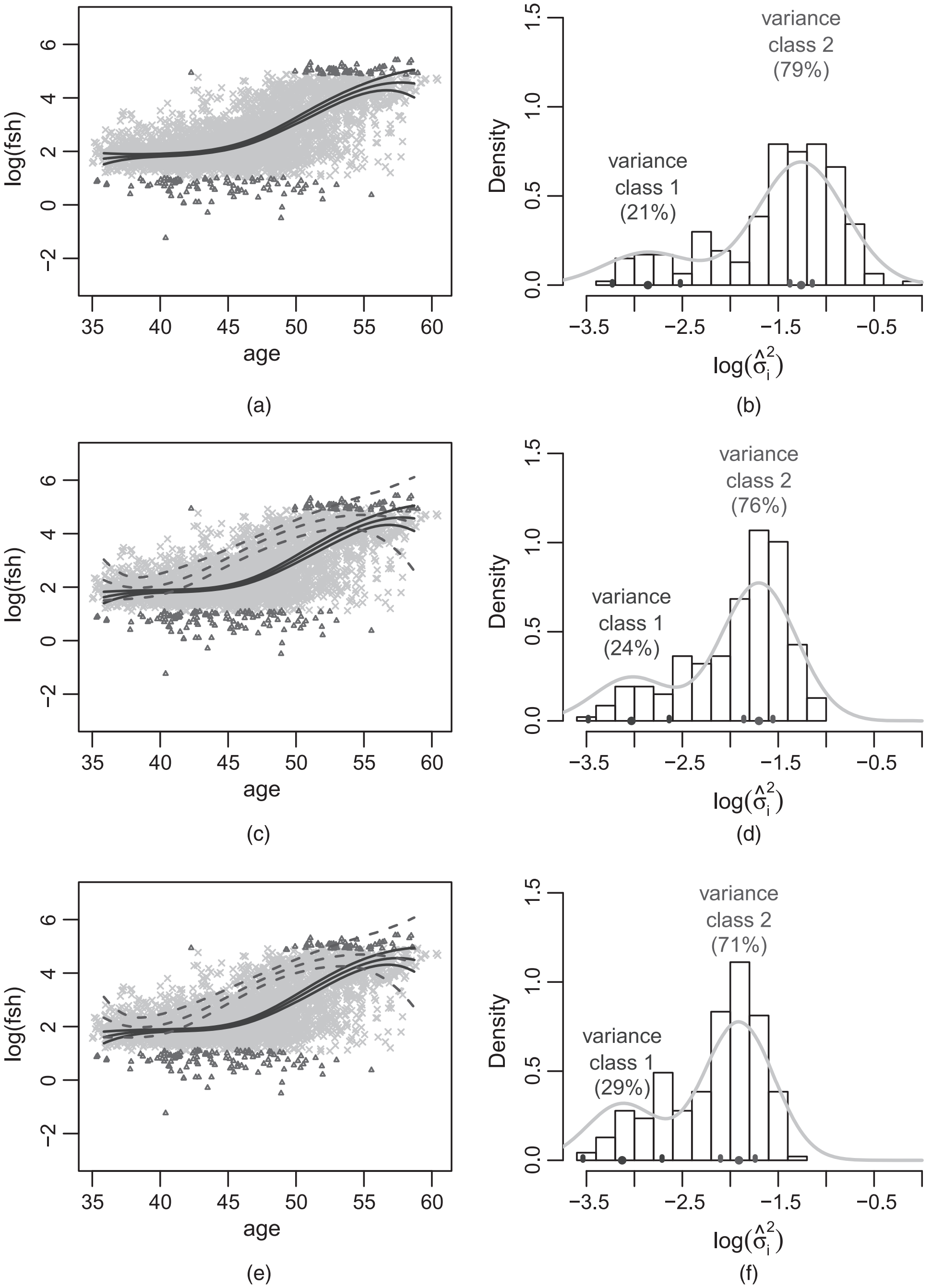


Fig. 2. Selected individual FSH trajectories from the Penn Ovarian Aging Study fitted by our best-fitting joint models with $\mu_i(t)$, $i=1,\ldots,n$, within the time window $T=[45,55]$ years as a functional predictor in the primary outcome submodel (———, normal model; — — —, t_7-model; · · · · · ·, t_4-model): (a) the fitted trajectories by all three models have upper bounds of p-values based on PDMs for individual trajectories of less than 0.05; (b) fitted trajectories by the normal model have upper bounds of p-values based on PDMs for individual trajectories less than 0.05 under the normality assumption but upper bounds of p-values greater than 0.05 under t_4- and t_7-model assumptions; (c) fitted trajectories under all three models have upper bounds of p-values based on PDMs for individual trajectories of greater than 0.05

Finally, to consider the effect of the derivative of the mean trajectory $\mu_i'(t)$, or the rate of change in the mean trajectory $\mu_i(t)$, we focus on the best-fitting t_4-model. Fig. 5(a) considers the effect of cumulative changes in the mean trajectories across the age range $T=[45,55]$, whereas Fig. 5(b) considers the equivalent effect across the age range $T=[50,55]$, which is potentially a more clinically relevant age range since the median age of menopause is 51 years and therefore the hormone dynamics in this time window are more likely to play a role in the menopause-



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 3. Longitudinal mean trajectories for the Penn Ovarian Aging Study from our final models with $J_\mu = 1$ and $K_D = K_C = 2$ in the longitudinal submodel, $\mu_i(t)$ as functional predictor with time window $T = [45, 55]$ years and $J_{\theta_0} = 3$ in the primary outcome submodel with different assumptions for the longitudinal submodel: (a) normal assumption (2.58% not covered by the 95% prediction interval); (b) normal assumption; (c) t_7-assumption (2.93% not covered by the 95% prediction interval; ———, major mean class (86%); ------, minor mean class (14%)); (d) t_7-assumption; (e) t_4-assumption (2.78% not covered by the 95% prediction interval); ———, major mean class (86%); ------, minor mean class (14%)); (f) t_4-assumption

Table 2. Estimates of the regression coefficients in the outcome model for the Penn Ovarian Aging Study by our best-fitting models with $\mu_i(t), i = 1, \ldots, n$, within the time window $T = [45, 55]$ years as a functional predictor in the primary outcome submodel†

Parameter	*Results for normal model*			*Results for t_7-model*			*Results for t_4-model*		
	Mean	*Standard error*	*95% credible interval*	*Mean*	*Standard error*	*95% credible interval*	*Mean*	*Standard error*	*95% credible interval*
α_0 (intercept)	0.305	0.995	(−1.631, 2.329)	0.279	0.972	(−1.637, 2.268)	0.012	0.985	(−1.886, 1.979)
λ_{01} (log(BMI))	0.068	0.277	(−0.501, 0.607)	0.039	0.264	(−0.497, 0.573)	0.101	0.273	(−0.449, 0.627)
λ_{02} (smoking)	*0.386*	0.170	(0.052, 0.717)	*0.370*	0.170	(0.039, 0.708)	*0.371*	0.171	(0.036, 0.708)
λ_{03} (variance)	*1.576*	0.565	(0.498, 2.703)	*1.887*	0.747	(0.451, 3.394)	*1.960*	0.723	(0.579, 3.403)

†Values in italics have 95% statistical significance.

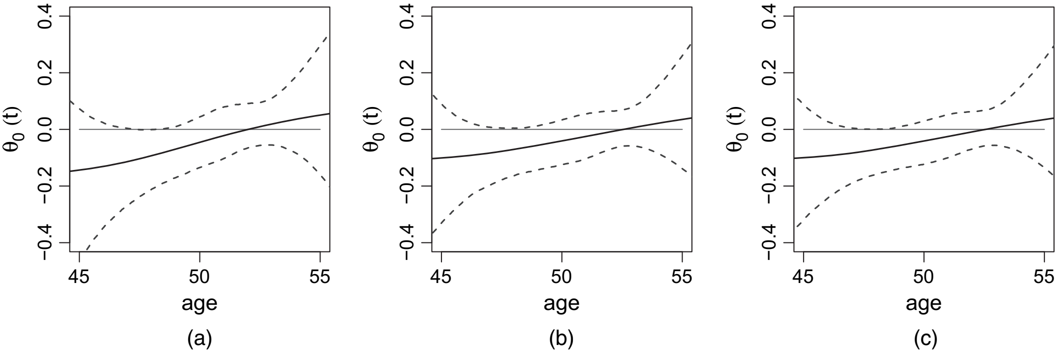


Fig. 4. Functional coefficient $\theta_0(t)$ for the Penn Ovarian Aging Study from our best-fitting (a) normal, (b) t_7- and (c) t_4-models with $\mu_i(t), i=1,\ldots,n$, within the time window $T=[45,55]$ years as a functional predictor in the primary outcome submodel

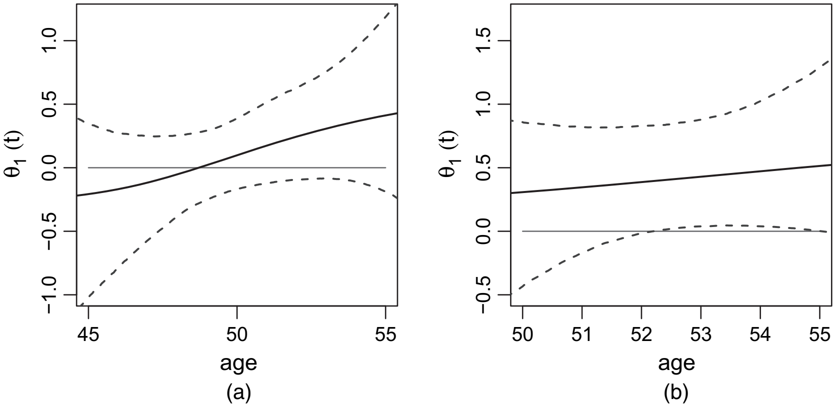
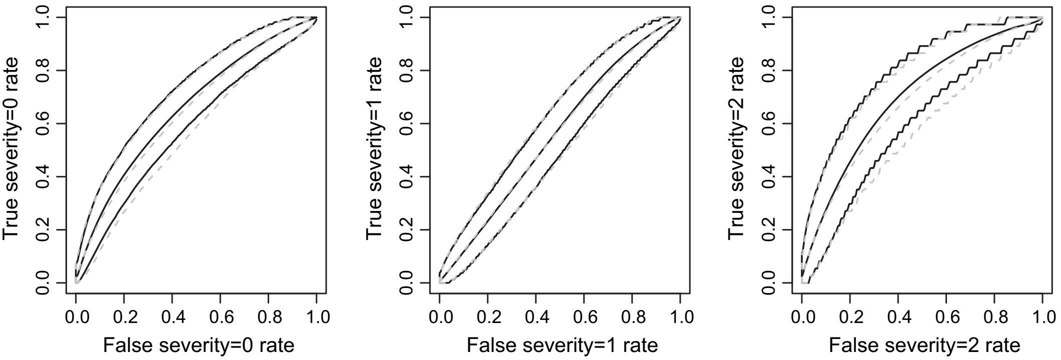


Fig. 5. Functional coefficient $\theta_1(t)$ for the Penn Ovarian Aging Study from our best-fitting model with $J_\mu=1$ and $K_D=K_C=2$ in the longitudinal submodel with t_4-assumption, and $J_{\theta_1}=3$ in the primary outcome submodel: (a) $\mu_1'(t)$ as functional predictor with $T=[45,55]$ years; (b) μ_1' as functional predictor with $T=[50,55]$ years

related symptoms. When fitted over the wider age range, higher values of $\mu_i'(t)$ decrease risk slightly before age 50 years and increase it over age 50 years, although the 95% credible intervals include 0 by a wide margin. In contrast, a more narrowly focused age range of $T=[50,55]$ years suggested a significantly increased risk of severe hot flashes that is associated with higher values of $\mu_i'(t)$ in the age range of 52.5–55 years, with $\hat{\theta}_1(52.5)=0.408$ (95% credible interval 0.019, 0.843) and $\hat{\theta}_1(55)=0.514$ (95% credible interval 0.003, 1.290).

Fig. 6 shows the ROC curves for the best-fitting t_4-model, comparing the use of the $\mu_i(t)$ and $\mu_i'(t)$ between ages 45 and 55 years to discriminate each of the hot flash severities (0, 1 and 2), along with the other predictors (residual variance, BMI and smoking status). These ROCs and their associated AUCs suggest that using either functional predictors led to moderately accurate classifications of different hot flash severities. Visually, there is not much difference in these ROC curves; a further comparison of AUCs also suggests that the predictive performances by using both $\mu_i(t)$ and $\mu_i'(t)$ have negligible differences (the ΔAUCs for severity 0, 1 and 2 are −0.012 (−0.097, 0.070), −0.002 (−0.073, 0.071) and −0.020 (−0.131, 0.091) respectively).



(a) (b) (c)

Fig. 6. ROC curves for the Penn Ovarian Aging Study from our final t-model (AUC0 is obtained by using $\mu_i(t)$ with $J_{\theta_0(t)} = 3$ within the time window $T = [45, 55]$ years as a functional predictor in the outcome submodel and AUC1 is obtained by using $\mu_i'(t)$ with $J_{\theta_1(t)} = 3$ within the time window $T = [45, 55]$ years as a functional predictor within the outcome submodel): (a) AUC0 (———) = 0.657 (0.606, 0.708) and AUC (– – –) = 0.645 (0.581, 0.709); (b) AUC0 (———) = 0.559 (0.512, 0.609) and AUC (– – –) = 0.557 (0.505, 0.616); (c) AUC0 (———) = 0.697 (0.627, 0.765) and AUC (– – –) = 0.678 (0.589, 0.757)

4. Conclusions and discussion

In this paper we develop a novel joint modelling approach to answer the scientifically important research question of how long-term history of FSH values or their rate of change affects the risk of hot flash severity, which is a symptom that almost every woman experiences during the menopausal transition. Although many joint models have been developed in the context of cancer research and human immunodeficiency virus or acquired immune deficiency syndrome clinical trials in the past decade, most methods focus on the features in the true underlying longitudinal process (i.e. mean trajectory) that take the forms of random effects or latent classes; or alternatively the last available true underlying value as a time-dependent covariate. Following Elliott *et al.* (2012) and Jiang *et al.* (2014), we seek the useful longitudinal features in both the mean trajectories and the short-term variability. Further we allow the mean of the longitudinal process and the corresponding derivatives to be time varying, and their effects on the responses to be cumulative over time. To summarize, we propose a broadly applicable joint modelling approach.

(a) The approach extends conventional functional data analysis to the framework of joint modelling of both the longitudinal (functional predictor) and outcome data, which allows us to study different aspects of the features in the dynamics of a longitudinal process as functional predictors. In particular, we focus on the values and derivatives of the mean trajectories at certain time windows as potential functional predictors. This will allow us to identify ages of vulnerability and to test hypotheses about the association between the functional predictors (FSH level and rate of change) and our outcome, severe hot flashes, while also adjusting for the previously identified effect of short-term variability captured by the variance of the residuals (Jiang *et al.*, 2014).

(b) It uses flexible mixed effects models with a Bayesian penalized B-spline basis and latent classes in the longitudinal submodel, which relaxes assumptions about the specific form of the trajectories and allows uneven spacing and unequal length that are densely or sparsely measured to be used as functional predictors.

(c) The approach allows the effects of FSH histories (the mean value or derivative) to be time varying and to accumulate over time. Statistical tests of these functional coefficient functions in the primary outcome submodel for hot flashes can then be used to identify critical time windows where the association exists. Using a Bayesian approach allows easy calculation of pointwise credible intervals for the functional coefficient functions in comparison with frequentist approaches.

(d) Finally, it uses a robust model to decrease the influence of outlying observations in the FSH data.

To realize these modelling goals, we use a penalized spline approach to allow the flexible modelling of longitudinal features and the functional coefficient curve representing the time varying effect of the longitudinal features. Since the ultimate goal is to model simultaneously both the mean trajectories and the residual variability but to distinguish between their effects in the outcome submodel, we choose a t-distribution to model residual variability properly to avoid the effect of outlying FSH values. In particular, we demonstrate the importance of assuming this robust distribution assumption instead of the typical normal assumption that is used in most of the joint modelling literature. However, because of the limited number of longitudinal observations for some women (i.e. ranging from 6 to 26), there is insufficient information in the data to assume individually varying degrees of freedom in the t-distribution; thus we are limited to assuming a global degrees of freedom that is common to all trajectories. In addition, our attempts to use the data to estimate even the global degrees-of-freedom parameter using the informative exponential distribution that was proposed by Geweke (1993), the truncated uniform prior on the inverse of the degrees of freedom that was suggested in Lange *et al.* (1989) and Gelman and Hill (2007) and the Jeffreys prior that was derived by Fonseca *et al.* (2008) all failed: the estimated global degrees of freedom were always close to a prior cut-off value, implying extreme outliers in the FSH data that tend to drive the degrees of freedom in the t-distribution to low values. Given that the fitted values are only modestly affected by different values of the degrees of freedom in the t-distribution (Lange *et al.*, 1989), we chose to fix the degrees-of-freedom parameter at a small number of fixed values and to conduct a sensitivity analysis using DIC to choose between the models.

The model proposed also allows latent heterogeneities in both the individual level mean trajectories and the residual variability as in Jiang *et al.* (2014). Under our best-fitting t_4-model, as shown in Fig. 3(e), the mean FSH trajectories can be separated into two classes: one minor class with 14% of trajectories and the other major class with 86% of trajectories. Both classes are reflective of three typical FSH change patterns for women in the transition to the menopause (Burger *et al.*, 1999) in that the FSH level is relatively flat before the menopause transition, has an increasing period during the menopause transition and eventually plateaus once women are 2 years post menopause; but women in the minor class tend to have earlier increases in FSH along with higher FSH values than the women in the major class. Fig. 7 plots the fitted mean FSH curves for the 28 women who were assigned to the minor class and a random sample of 20 women who were assigned to the major class on the basis of the posterior mode. This once again shows the heterogeneous nature in the mean FSH trajectories that is supported by our model selection criterion DIC and implies that the women in the minor class tend to reach the menopause at a much earlier age. Also, as shown in Fig. 3, even with the use of the t-distribution to account for extreme outlying observations, it seemed that there is still a true mixture in residual variability, with a low variance class consisting of one in three to one in five women, with the remainder in a high variance class.

In summary, the model proposed gives added insights about hormone changes in the menopausal transition and their associations with severe hot flashes. First, whether the robust

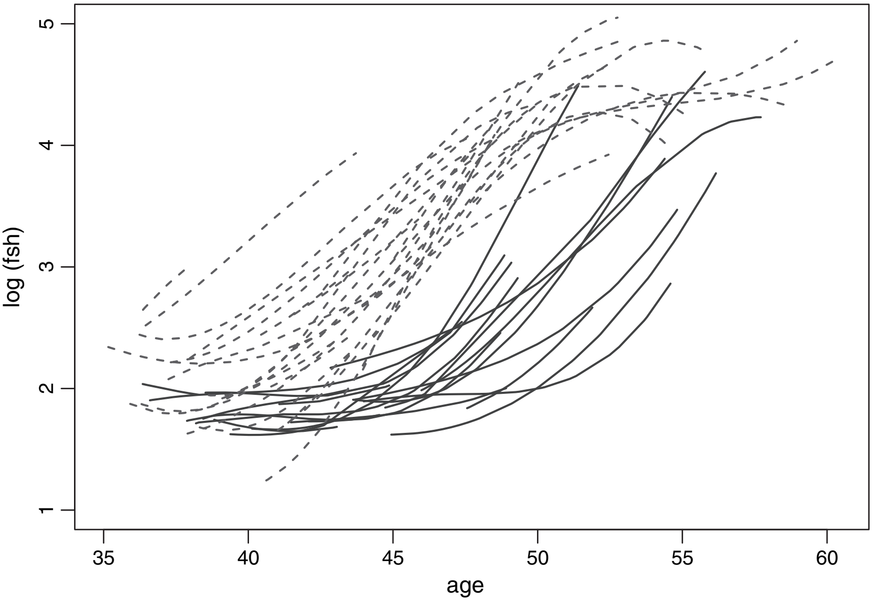


Fig. 7. Individual FSH trajectories from the Penn Ovarian Aging Study that are assigned to the minor (– – –) and major (———) mean trajectory classes by our best-fitting t_4-model with $\mu_i(t), i = 1, \ldots, n$, within the time window $T = [45, 55]$ years as a functional predictor in the primary outcome submodel

or normal models were used, we identified a strong association between residual variability in FSH and hot flashes as in Jiang *et al.* (2014), and similarly to what has been reported for depressive symptoms (Freeman *et al.*, 2006). In addition, we identified latent heterogeneities in both the individual level mean trajectories. Under our best-fitting t_4-model, as shown in Fig. 3(e), the mean FSH trajectories can be separated into two classes: one minor class with 14% of trajectories and the other major class with 86% of trajectories. Both classes are reflective of three typical FSH change patterns for women in the transition to the menopause (Burger *et al.*, 1999) in that the FSH level is relatively flat before the menopause transition, has an increasing period during the menopause transition and will eventually plateau once women are about 2 years post menopause; but women in the minor class tend to have an earlier increase in their FSH trajectories along with higher FSH values than the women in the major class. As shown in Fig. 7, the fitted mean FSH curves for the total 28 women who were assigned to the minor class and a random sample of 20 women who were assigned to the major class on the basis of the posterior mode were plotted. This once again shows the heterogeneous nature in the mean FSH trajectories that is supported by our model selection criterion DIC and implies that the women in the minor class tend to the reach the menopause at a much earlier age. Also, as shown in Fig. 3, even with the use of the t-distribution to account for extreme outlying observations, it seemed that there is still a true mixture in residual variability. Another interesting finding is illustrated in Fig. 5(b) depicting the association between increases in hot flashes and the functional coefficient which describes the rate of change in FSH between the ages of 50 and 55 years. This age window corresponds precisely to when hot flashes are reported to be most likely (Harlow *et al.*, 2012). These findings have important ramifications for treatment of hot flashes with hormone replacement therapy. These medications affect the levels of FSH and oestradiol, and reduce variability. The current recommendation is for women to take these medications for no more than 3–5 years; however, the optimal time frame and duration for treatment are unknown.

Generally, the functional coefficient curves $\theta_0(t)$ and $\theta_1(t)$ can be fitted by any spline basis with or without penalty parameters. In particular, if the shape of $\theta_0(t)$ or $\theta_1(t)$ is known—e.g. $\theta_0(t)$ is a linear function—then we can let $\psi^0(t) = (1, t)$ and assume a regular normal prior on the coefficients that are associated with basis function 1 and t. When the true shape of $\theta_0(t)$ or $\theta_1(t)$ is unknown, we recommend starting the analysis by using a more flexible penalized approach to obtain some idea of the shape of $\theta_0(t)$ or $\theta_1(t)$, which may be further reduced to simple parametric form to stabilize estimation of model parameters and to reduce the length of pointwise credible or confidence intervals for $\theta_0(t)$ or $\theta_1(t)$.

The methods that were presented for data augmentation of unobserved FSH values assumes MAR. For the FSH values that were missing because of age at enrolment or reasons such as a subject did not deliver a blood sample at a certain visit, we can reasonably assume MAR. One known non-random source of missingness would be when women went on hormone replacement therapy for relief of menopausal symptoms. These hormone values during hormone replacement therapy were censored. In this subset who were symptom free at baseline, $31/234 = 13\%$ reported any hormone therapy use over the 13 years of follow-up, and the majority, $26/31 = 84\%$, reported use at only one or two visits. Among the remaining five women, three reported use of hormone replacement therapy at six visits, one woman reported use at four visits, and one at three visits. However, skipped visits or dropout during the first 5 years (i.e. 10 visits) for women were less likely to be due to menopausal symptoms (Nelson *et al.*, 2004). Furthermore, when fitting the individual's FSH trajectory assuming MAR, we did not observe noticeable irregular residual patterns from the FSH values that were collected before and after skipped visits; therefore the effect from assuming MAR for the sporadic missingness should be minimal. We may underestimate the short-term variation if the missingness is associated with a high level of FSH fluctuation and this could be a worthy future research topic. For dropout, we may expect an effect if those who dropped out had different profiles after they left from those who stayed. There are a total of 26 women who dropped out after being in the programme for more than 5 years. Among them, 10 women contribute 20 or more observations before the dropout and five women dropped out at age 54 years or older. A preliminary study that examined FSH patterns and values in the visits before the dropout did not reveal a reason behind the dropout. Nor could we find an explanation behind their dropouts based on factors such as their history of hot flash severity, menopausal stage or HRT use. Future work will develop methods to examine thoroughly the sensitivity to different missing data mechanisms through pattern mixture models or selection models within our modelling framework, although the sensitivity of our results to failures of the MAR assumption as expected would be relatively minor given the limited amount of missing data.

Another direction for future work is to make use of the fact that longitudinal studies often measure several variables repeatedly; for example, in the Penn Ovarian Aging Study several other hormone trajectories are available. Developing methods to model these potentially correlated longitudinal trajectories simultaneously while also using this information effectively to predict or relate to the outcome of interest is a key area for future research.

Acknowledgements

This work was supported in part by grant R03AG031980 from the US National Institute of Aging and by grant R01CA74552 from the US National Cancer Institute. The authors thank Dr Ellen W. Freeman, for sharing her data, as well as two reviewers and the Associate Editor for comments that helped to improve the manuscript.

References

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass.*, **88**, 669–679.

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

Brown, E. R. and Ibrahim, J. G. (2003b) Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics*, **59**, 686–693.

Burger, H. G., Dudley, E. C., Hopper, J. L., Groome, N., Guthrie, J. R., Green, A. and Dennerstein, L. (1999) Prospectively measured levels of serum follicle-stimulating hormone, estradiol, and the dimeric inhibins during the menopausal transition in a population-based cohort of women. *J. Clin. Endcrin. Metablm*, **84**, 4025–4030.

Celeux, G., Forbes, F., Robert, C. P. and Titterington, D. M. (2006) Deviance information criteria for missing data models. *Baysn Anal.*, **1**, 651–673.

Chen, H. and Wang, Y. (2011) A penalized spline approach to functional mixed effects model analysis. *Biometrics*, **67**, 861–870.

Day, N. E. (1969) Estimating the components of a mixture of normal distributions. *Biometrika*, **56**, 463–474.

Durbán, M., Harezlak, J., Wand, M. P. and Carroll, R. J. (2005) Simple fitting of subject-specific curves for longitudinal data. *Statist. Med.*, **24**, 1153–1167.

Eilers, P. H. C. and Marx, B. D. (1996) Flexible smoothing with B-splines and penalties. *Statist. Sci.*, **11**, 89–121.

Elliott, M. R. (2007) Identifying latent clusters of variability in longitudinal data. *Biostatistics*, **8**, 756–771.

Elliott, M. R., Sammel, M. D. and Faul, J. (2012) Associations between variability of risk factors and health outcomes in longitudinal studies. *Statist. Med.*, **31**, 2745–2756.

Fawcett, T. (2006) An introduction to ROC analysis. *Pattn Recogn Lett.*, **27**, 861–874.

Fonseca, T. C., Ferreira, M. A. and Migon, H. S. (2008) Objective Bayesian analysis for the student-t regression model. *Biometrika*, **95**, 325–333.

Freeman, E. W., Sammel, M. D., Lin, H., Liu, Z. and Gracia, C. R. (2011) Duration of menopausal hot flushes and associated risk factors. *Obstetr. Gyn.*, **117**, 1095–1104.

Freeman, E. W., Sammel, M. D., Lin, H. and Nelson, D. B. (2006) Associations of hormones and menopausal status with depressed mood in women with no history of depression. *Arch. Gen. Psychiatr.*, **63**, 375–382.

Frühwirth-Schnatter, S. (2006) *Finite Mixture and Markov Switching Models*. New York: Springer.

Garrett, E. S. and Zeger, S. L. (2000) Latent class model diagnosis. *Biometrics*, **56**, 1055–1067.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003) *Bayesian Data Analysis*, 2nd edn. London: CRC Press.

Gelman, A., Goegebeur, Y., Tuerlinckx, F. and Van Mechelen, I. (2000) Diagnostic checks for discrete data regression models using posterior predictive simulations. *Appl. Statist.*, **49**, 247–268.

Gelman, A. and Hill, J. (2007) *Data Analysis using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Gelman, A., Meng, X.-L. and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sin.*, **6**, 733–760.

Geweke, J. (1993) Bayesian treatment of the independent student-t linear model. *J. Appl. Econmetr.*, **8**, S19–S40.

Harlow, S. D., Gass, M., Hall, J. E., Lobo, R., Maki, P., Rebar, R. W., Sherman, S., Sluss, P. M. and de Villiers, T. J. (2012) Executive summary of the stages of reproductive aging workshop+ 10: addressing the unfinished agenda of staging reproductive aging. *Climacteric*, **15**, 105–114.

Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2004) Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. *Statist. Sin.*, **14**, 863–884.

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

James, G. M. (2002) Generalized linear models with functional predictors. *J. R. Statist. Soc.* B, **64**, 411–432.

Jasra, A., Holmes, C. and Stephens, D. (2005) Markov chain Monte Carlo methods and the label switching problem in bayesian mixture modeling. *Statist. Sci.*, **20**, 50–67.

Jeffreys, H. (1973) *Scientific Inference*, 3rd edn. New York: Cambridge University Press.

Jiang, B., Elliott, M. R., Sammel, M. D. and Wang, N. (2014) Joint modeling of cross-sectional health outcomes and longitudinal predictors via mixtures of means and variances. *Biometrics*, to be published, doi 10.1111/biom.12284.

Johnson, V. E. (2004) A Bayesian χ^2 test for goodness-of-fit. *Ann. Statist.*, **32**, 2361–2384.

Johnson, V. E. (2007) Bayesian model assessment using pivotal quantities. *Baysn Anal.*, **2**, 719–734.

Kass, R. E. and Natarajan, R. (2006) A default conjugate prior for variance components in generalized linear mixed models (comment on article by Browne and Draper). *Baysn Anal.*, **1**, 535–542.

Lang, S. and Brezger, A. (2004) Bayesian P-splines. *J. Computnl Graph. Statist.*, **13**, 183–212.

Lange, K. L., Little, R. J. and Taylor, J. M. (1989) Robust statistical modeling using the t distribution. *J. Am. Statist. Ass.*, **84**, 881–896.

Law, N. J., Taylor, J. M. and Sandler, H. (2002) The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics*, **3**, 547–563.

Little, R. J. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*. New York: Wiley.

Muthén, B. and Shedden, K. (1999) Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, **55**, 463–469.

Neelon, B., O'Malley, A. J. and Normand, S.-L. T. (2011) A Bayesian two-part latent class model for longitudinal medical expenditure data: assessing the impact of mental health and substance abuse parity. *Biometrics*, **67**, 280–289.

Nelson, D. B., Sammel, M. D., Freeman, E. W., Liu, L., Langan, E. and Gracia, C. R. (2004) Predicting participation in prospective studies of ovarian aging. *Menopause*, **11**, 543–548.

Pemstein, D., Quinn, K. M. and Martin, A. D. (2007) The scythe statistical library: an open source C++ library for statistical computation. *J. Statist. Softwr.*, **42**, no. 12, 1–26.

Ramsay, J. O. and Dalzell, C. J. (1991) Some tools for functional data analysis (with discussion). *J. R. Statist. Soc.* B, **53**, 539–572.

Redner, R. A. and Walker, H. F. (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*, **26**, 195–239.

Rizopoulos, D. (2012) *Joint Models for Longitudinal and Time-to-event Data: with Applications in R*. Boca Raton: CRC Press.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. and Müller, M. (2011) proc: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.*, no. 12, article 77.

Rodríguez, C. E. and Walker, S. G. (2012) Label switching in Bayesian mixture models: deterministic relabeling strategies. *J. Computnl Graph. Statist.*, **23**, 25–45.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. New York: Cambridge University Press.

Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

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

Sowers, M. R., Zheng, H., McConnell, D., Nan, B., Harlow, S. and Randolph, J. F. (2008) Follicle stimulating hormone and its rate of change in defining menopause transition stages. *J. Clin. Endcrin. Metablm*, **93**, 3958–3964.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.

Stephens, M. (2000) Dealing with label switching in mixture models. *J. R. Statist. Soc.* B, **62**, 795–809.

Tsiatis, A., Degruttola, V. and Wulfsohn, M. (1995) Modeling the relationship of survival to longitudinal data measured with error: applications to survival and cd4 counts in patients with aids. *J. Am. Statist. Ass.*, **90**, 27–37.

Verbeke, G. and Lesaffre, E. (1996) A linear mixed-effects model with heterogeneity in the random-effects population. *J. Am. Statist. Ass.*, **91**, 217–221.

Wang, Y. and Taylor, J. M. G. (2001) Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *J. Am. Statist. Ass.*, **96**, 895–905.

Yu, M., Taylor, J. M. G. and Sandler, H. M. (2008) Individual prediction in prostate cancer studies using a joint longitudinal survival–cure model. *J. Am. Statist. Ass.*, **103**, 178–187.

Yuan, Y. and Johnson, V. E. (2012) Goodness-of-fit diagnostics for Bayesian hierarchical models. *Biometrics*, **68**, 156–164.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

> 'Web-based supporting materials for "Modeling short- and long-term characteristics of follicle stimulating hormone as predictors of severe hot flashes in Penn Ovarian Aging Study" '.