Biometrics 71, 487–497
June 2015
DOI: 10.1111/biom.12284

Joint Modeling of Cross-Sectional Health Outcomes and Longitudinal Predictors via Mixtures of Means and Variances

Bei Jiang,[1,*] Michael R. Elliott,[1,2] Mary D. Sammel,[3] and Naisyin Wang[4]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, U.S.A.
[2]Survey Methodology Program, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, U.S.A.
[3]Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A.
[4]Department of Statistics, University of Michigan, Ann Arbor, Michigan, U.S.A.
email: beijiang@umich.edu

Summary. Joint modeling methods have become popular tools to link important features extracted from longitudinal data to a primary event. While most modeling strategies have focused on the association between the longitudinal mean trajectories and risk of an event, we consider joint models that incorporate information from both long-term trends and short-term variability in a longitudinal submodel. We also consider both shared random effect and latent class (LC) approaches in the primary-outcome model to predict a binary outcome of interest. We develop simulation studies to compare and contrast these two modeling strategies; in particular, we study in detail the effects of the primary-outcome model misspecification. Among other findings, we note that when we analyze data from a shared random-effect using a LC model while the information from the longitudinal data is weak, the LC approach is more sensitive to such a model misspecification. Under this setting, the LC model has a superior performance in within-sample prediction that cannot be duplicated when predicting new samples. This is a unique feature of the LC approach that is new as far as we know to the existing literature. Finally, we use the proposed models to study how follicle stimulating hormone (FSH) trajectories are related to the risk of developing severe hot flashes for participating women in the Penn Ovarian Aging Study.

Key words: Joint model; Latent class; Long-term trend; Model misspecification; Predictive performance; Shared random effects and variances; Short-term variability

1. Introduction

Joint models naturally link longitudinal covariates to disease outcomes. Many joint models have been developed in the context of cancer research and HIV/AIDS clinical trials, where a mixed-effect model is outlined for the longitudinal trajectories and a primary outcome model is defined for the disease outcome. The primary outcome models are often postulated as: (1) shared random effects (SRE) models, where covariates include a functional form of the random effects in the mixed-effect submodel, and (2) latent class (LC) models, where there exists heterogeneity (latent classes) in the longitudinal mean profiles, and subjects in a particular LC share the same risk of event, conditional on other covariates.

For SRE models, the random effects are used to capture the main features in the longitudinal trajectories that predict the outcomes. The concept of "shared parameters" was first used in Wu and Carroll (1988) to model non-ignorable missing data, and later by Henderson, Diggle, and Dobson (2000) to jointly analyze longitudinal and time-to-event data; also see Tsiatis and Davidian (2004), Ibrahim, Chen, Sinha (2001), and Ibrahim, Chu, and Chen (2010) for excellent general reviews of these models. In the LC model literature, growth mixture models (Verbeke and Lesaffre, 1996; Muthén and Shedden, 1999) are extensions of random growth curve models, creating distinct subgroups where individual trajectories vary around group-specific mean trajectories. Considering time-to-event outcomes, Proust-Lima et al. (2012) studied the joint LC modeling in detail and contrasted its use in terms of goodness of fit, prediction accuracy and model performances with that of joint SRE models. Using a prostate cancer study data consisting of four well-separated classes of longitudinal mean trajectories, they illustrated that, in comparison to a LC model, the use of SRE model alone was not sufficient to fully capture the relationship between class-specific outcomes and the heterogeneity among different classes. They also reported that only a mild advantage of LC remained for prediction of outcomes from an external data set of similar nature.

In this article, we study the associations between longitudinal hormone levels and menopausal symptoms for a group of middle-aged women. The Penn Ovarian Aging Study (Freeman et al., 2011) is a longitudinal study consisting of a population-based sample of 436 women aged 35–47 years selected via random digit dialing in Philadelphia County, PA during 1996–1997. At each annual assessment, measurements and a blood sample were collected two times approximately a month apart. One goal of the study is to explore associations between reproductive hormone levels and symptoms in the transition to menopause. Changes in hormone levels alter menstrual bleeding patterns prior to menopause marking the end of a woman's reproductive years. This course of

events coincides for a majority of women with the development of hot flashes, sleep disorders, and bone loss, among other symptoms. While researchers have focused on the associations between these symptoms and hormone levels, the impact of within woman rate of change and variability in hormones, such as follicle stimulating hormone (FSH), is not well understood. To evaluate that hypothesis that subject-level hormone fluctuation may accentuate menopausal symptoms (Freeman et al., 2006), we investigate methods that model both longitudinal profiles and residual variability of the FSH and simultaneously link them with the risk of experiencing severe hot flashes (SHF). While most joint models have treated within-subject variability as a nuisance parameter, recently a small literature has developed to evaluate the associations between longitudinal within-subject variability and the primary outcomes (Sammel et al., 2001; Elliott 2007; Elliott, Sammel, and Faul, 2012).

Thus, in this dataset we have longitudinal measures with heterogeneity both in trajectory and variability that may be predictive of a binary outcome. There is evidence that these trajectories and variabilities may cluster into possibly clinically relevant groupings, so we consider a mixture model for FSH hormone that also includes latent classes for the subject-level trajectories and variability. This leads to two potential candidate models for the outcome: a "multiple shared random effects" (MSRE) model whose predictors are subject-specific random coefficients, and a LC model whose predictors are the LC memberships. Since it is not clear which approach is best, we examine the robustness and predictive accuracy of each approach via simulation study. Our key focus is not on one primary-outcome model or the other, but their contrasts and the information they jointly provide.

2. Joint Models and Corresponding Approaches

The joint modeling approach consists of a model for the longitudinal trajectories and a primary model for the outcomes.

- Let y_{ij} denote the longitudinal covariate for subject i at time t_{ij}, $j = 1, ..., n_i$, $i = 1, ..., n$, the longitudinal submodel of y_{ij} is a generalized growth mixture model (Muthén and Shedden, 1999) with subject-specific mean trajectories and residual variances:

$$\begin{aligned} D_i &\sim \text{Multinomial}(\pi_1^D, ..., \pi_{K_D}^D), \\ \boldsymbol{b}_i | D_i = d &\sim \text{N}(\boldsymbol{\beta}_d, \boldsymbol{\Sigma}_d), d = 1, ..., K_D; \\ C_i &\sim \text{Multinomial}(\pi_1^C, ..., \pi_{K_C}^C), \\ \sigma_i^2 | C_i = c &\sim \text{log-N}(\mu_c, \tau^2), c = 1, ..., K_C; \\ \text{and } y_{ij} | \boldsymbol{b}_i, \sigma_i^2 &\sim \text{N}\{f(\boldsymbol{b}_i; t_{ij}), \sigma_i^2\}, \end{aligned} \tag{1}$$

  where $\boldsymbol{b}_i$ is the r dimensional vector of subject-level random effects that reflect the subject-level trajectory patterns, and σ_i^2 is the residual variance. D_i and C_i define the latent classes for the longitudinal means and individual variance memberships, respectively.
- The primary outcome model is a probit regression model:

$$\Phi^{-1}(\text{p}(o_i = 1)) = \boldsymbol{Z}_i' \boldsymbol{\eta}, \tag{2}$$

  where the binary o_i denotes the health outcome, and $\boldsymbol{Z}_i$ the i^{th} set of covariates in the probit model. For the LC model, $\boldsymbol{Z}_i$ contains the LC memberships, D_i and C_i; while for the MSRE model, $\boldsymbol{Z}_i$ contains SRE and residual variances. Other baseline variables may be included in $\boldsymbol{Z}_i$ as well.

Throughout, we let $\boldsymbol{\phi}$ consist of all parameters in π_d^D, $\boldsymbol{\beta}_d$, $\boldsymbol{\Sigma}_d$, π_c^C, μ_c, τ^2, $\boldsymbol{\eta}$. We also replace $\boldsymbol{\eta}$ in (2) by $\boldsymbol{\theta}$ for the LC and by $\boldsymbol{\gamma}$ for the MSRE models to ease the task of presentation.

2.1. *Structure Specification and Posterior Computation*

We denote the prior distribution of $\boldsymbol{\phi}$ by $\pi(\boldsymbol{\phi})$, assume each parameter in $\boldsymbol{\phi}$ has independent prior and let $\boldsymbol{z} = (\boldsymbol{b}, \boldsymbol{\sigma}, \boldsymbol{C}, \boldsymbol{D})'$. The variable $\boldsymbol{x}$ consists of the longitudinal $\boldsymbol{y}$'s and the outcomes $\boldsymbol{o}$'s. The complete data likelihood of $\boldsymbol{\phi}$ based on data $(\boldsymbol{x}, \boldsymbol{z})$ is given by,

$$\begin{aligned} & f(\boldsymbol{x}, \boldsymbol{z} | \boldsymbol{\phi}) \\ & \propto \left\{ \prod_{i=1}^{n} \left[\prod_d \left[\pi_d^D (2\pi)^{-r/2} |\boldsymbol{\Sigma}_d|^{-1/2} \right. \right. \right. \\ & \quad \left. \times \exp\left\{ -\frac{1}{2} (\boldsymbol{b}_i - \boldsymbol{\beta}_d)' \boldsymbol{\Sigma}_d^{-1} (\boldsymbol{b}_i - \boldsymbol{\beta}_d) \right\} \right]^{\text{I}(D_i = d)} \\ & \quad \times \prod_c \left[\pi_c^C (2\pi\tau^2)^{-\frac{1}{2}} \sigma_i^{-2} \exp\left\{ -\frac{1}{2\tau^2} \left(\log \sigma_i^2 - \mu_c \right)^2 \right\} \right]^{\text{I}(C_i = c)} \\ & \quad \times \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2\sigma_i^2} \left\{ y_{ij} - f(\boldsymbol{b}_i; t_{ij}) \right\}^2 \right] \\ & \quad \left. \left. \times \Phi(\boldsymbol{Z}_i' \boldsymbol{\eta})^{o_i} \left\{ 1 - \Phi(\boldsymbol{Z}_i' \boldsymbol{\eta}) \right\}^{1 - o_i} \right] \right\} \pi(\boldsymbol{\phi}). \end{aligned} \tag{3}$$

We propose a Bayesian approach to estimate model parameters. For the mixture normal distribution of the random effects, we let $\boldsymbol{\beta}_d \sim \text{N}(\boldsymbol{0}, \boldsymbol{V})$, $\boldsymbol{V} = n\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\beta}}$ is the estimator in regressing $\boldsymbol{y}$ on the design matrix defined by $f(\cdot; t_{ij})$. This corresponds to a "single observation" data-driven inflated covariance prior centered at a null model, and avoids improper posteriors resulting from the possibility that some latent classes are not represented in the data (Elliott et al., 2005). For the covariance matrix of the random effects, $\boldsymbol{\Sigma}_d$, we use the prior from Kass and Natarajan (2006): $\boldsymbol{\Sigma}_d \sim \text{Inverse-Wishart}(\text{df} = m, \boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda} = r \left(\sum_{i=1}^{n} \widehat{\text{Cov}}(\tilde{\boldsymbol{b}}_i)^{-1} / n \right)^{-1}$; $\tilde{\boldsymbol{b}}_i$ is the OLS estimator of $\boldsymbol{b}_i$. We let $m = 2.5 + (r-1)/2$ as suggested by Frühwirth-Schnatter (2006, Section 6.3.2) to restrain the eigenvalues of the covariance matrices away from 0, avoiding "local maxima" that can result from the improper posterior due to unbounded likelihoods when the covariance matrix is unrestricted in normal mixture models (Day, 1969).

For the mixture log-normal distribution for the residual variances, we used diffuse priors: $\mu_c \sim \text{N}(0, v)$, $\tau^{-2} \sim \text{Gamma}(a, b)$ with $v = 1000$ and $a = b = 0.001$. For the class membership probabilities, we assume conjugate Dirichlet$(4, ..., 4)$ on both $\boldsymbol{\pi}^C$ and $\boldsymbol{\pi}^D$ (Frühwirth-Schnatter,

2006); this is equivalent to assuming a priori four observations per-class, avoiding having empty classes. Lastly, we let $\boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0}, (9/4)\boldsymbol{I})$ in the probit regression, where $(9/4)\boldsymbol{I}$ would bound the estimated outcome probabilities to be away from 0 and 1 (Garrett and Zeger, 2000).

Gibbs sampling is used to obtain draws from the posterior distributions. For $(\boldsymbol{\eta} \mid \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{O})$ we use the Albert and Chib (1993) data augmentation method for probit regression models. The draws of $(\sigma_i^2 | C_i, \{\mu_c\}_c, \tau^2, \boldsymbol{b}_i, o_i, \{y_{ij}\}_i)$ for all i are obtained by the inverse cumulative distribution method. The exact specifications of all priors and MCMC procedures are given in Web Appendix A. In the Ovarian Aging data analysis, we ran three chains from diverse starting points and use Gelman–Rubin statistics (Gelman et al., 2003) to assess MCMC convergence. In simulations, we started the chains at the initial values obtained from estimated individual parameters in longitudinal $\boldsymbol{y}_i$'s and ad hoc estimates built from them.

For the well-documented issue of "label switching" in mixture modeling (Redner and Walker, 1984), we applied the post-processing relabeling algorithm (Stephens, 2000) where class permutations and re-assignment are adopted at each MCMC iteration. In simulations, we ran Stephens's relabeling algorithm with the initial class labels on the raw MCMC output. In the data application, for models with $K_D = 2$ or $K_C = 2$, there is little evidence of label switching. For cases of larger than two K_D or K_C, label switching happens more frequently. With the convergence speed of Stephens's algorithm depending on the quality of initial labels, we re-initialize the class labels when needed, prior to a full re-run of the algorithm.

2.2. *The Choice of the Number of Classes*

The choice of the number of latent classes is known to be a challenging problem in modeling finite mixtures (McLachlan and Peel, 2000). We consider two commonly used Bayesian model assessment criteria: the deviance information criterion (DIC) of Spiegelhalter et al. (2002), and the logarithm of the pseudomarginal likelihood (LPML), proposed by Geisser and Eddy (1979). For DIC, recalling $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{o})'$, we consider

$$\begin{aligned}\mathrm{DIC}(\boldsymbol{x}) &= \overline{D(\boldsymbol{\phi})} + p_D = 2\overline{D(\boldsymbol{\phi})} - D(\bar{\boldsymbol{\phi}}) \\ &= -4\mathrm{E}_{\boldsymbol{\phi}}\left\{\log f(\boldsymbol{x} \mid \boldsymbol{\phi}) \mid \boldsymbol{x}\right\} + 2\log f\left\{\boldsymbol{x} \mid \mathrm{E}(\boldsymbol{\phi} \mid \boldsymbol{x})\right\}.\end{aligned}$$

In our setting with latent $\boldsymbol{z}$, $f(\boldsymbol{x} \mid \boldsymbol{\phi})$ is not available in closed form. We use the approach outlined in Celeux et al. (2006) and detailed in Web Appendix B, to obtain $\mathrm{DIC}(\boldsymbol{x})$ by

$$\begin{aligned}\mathrm{E}_{\boldsymbol{z}}\left\{\mathrm{DIC}(\boldsymbol{x}, \boldsymbol{z})\right\} = &-4\mathrm{E}_{\boldsymbol{z},\boldsymbol{\phi}}\left\{\log f(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\phi}) \mid \boldsymbol{x}\right\} \\ &+ 2\mathrm{E}_{\boldsymbol{z}}\left[\log f\left\{\boldsymbol{x}, \boldsymbol{z} \mid \mathrm{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \boldsymbol{x}, \boldsymbol{z})\right\} \mid \boldsymbol{x}\right],\end{aligned}$$

where integration over the latent $\boldsymbol{z}$ is obtained via numerical methods.

LPML corresponds to a Bayesian cross-validation measure and is defined as $\mathrm{LPML} = \sum_{i=1}^{n} \log(\mathrm{CPO}_i)$, where $\mathrm{CPO}_i = f(\boldsymbol{y}_i, o_i | \boldsymbol{y}_{(-i)}, \boldsymbol{o}_{(-i)})$ represents a leave-one-out cross-validated posterior predictive density for $(\boldsymbol{y}_i, o_i)$ given the data excluding $(\boldsymbol{y}_i, o_i)$ (denoted by $(\boldsymbol{y}_{(-i)}, \boldsymbol{o}_{(-i)})$). The model with higher value of LPML provides a better fit to the data (Ibrahim et al., 2001). Details of the LPML computation are also provided in Web Appendix B.

2.3. *Goodness of Fit Evaluation*

We assessed the model goodness of fit to the data in two ways. First, we examined the posterior predictive distributions (PPDs; Gelman et al., 2003), where a PPD p value close to 0.5 implies a satisfactory fit of the model to the data. For the longitudinal trajectories, we draw $\boldsymbol{y}^{\mathrm{rep}}$ from the posterior predictive distribution $f(\boldsymbol{y}^{\mathrm{rep}} | \boldsymbol{y}, \boldsymbol{o}) = \int f(\boldsymbol{y}^{\mathrm{rep}} | \boldsymbol{\phi}, \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{o}) f(\boldsymbol{\phi}, \boldsymbol{z} | \boldsymbol{y}, \boldsymbol{o}) d\boldsymbol{\phi} d\boldsymbol{z}$ to compute the PPD p values $\mathrm{P}\{T_i(\boldsymbol{y}_i; \boldsymbol{b}_i, \sigma_i^2) < T_i(\boldsymbol{y}_i^{\mathrm{rep}}; \boldsymbol{b}_i, \sigma_i^2) \mid \boldsymbol{y}\}$, where for subject i, we consider a χ^2-like statistic, $T_i(\boldsymbol{y}_i; \boldsymbol{b}_i, \sigma_i^2) = \sum_j \left\{y_{ij} - f(\boldsymbol{b}_i; t_{ij})\right\}^2 / \sigma_i^2$. For the outcome indicator o_i, we compute $\mathrm{P}\left(T^{\mathrm{rep}} < T^{\mathrm{obs}} \mid \boldsymbol{y}\right)$, where $T^{\mathrm{obs}} = n^{-1} \sum_i \mathrm{I}(o_i = 1)$, and $T^{\mathrm{rep}} = n^{-1} \sum_i o_i^{\mathrm{rep}}$ with o_i^{rep} drawn from the posterior predictive distribution, a Bernoulli distribution with the success probability $P(o_i^{\mathrm{rep}} = 1 | \boldsymbol{y}, \boldsymbol{o}) = \int \Phi(\boldsymbol{Z}_i' \boldsymbol{\eta}) f(\boldsymbol{\phi}, \boldsymbol{z} | \boldsymbol{y}, \boldsymbol{o}) \mathrm{d}\boldsymbol{\phi} \mathrm{d}\boldsymbol{z}$.

Second, we assessed the discriminatory ability of the model using receiver-operating characteristic (ROC) curves, in particular the area under the ROC curve (AUC). ROC curves plot true positive rate (TP) versus false positive rate (FP) for all possible cutoffs based on predicted $\mathrm{P}(o_i = 1) = \Phi(\boldsymbol{Z}_i' \boldsymbol{\eta})$ obtained from (2). The ROC curve and AUC were computed at each MCMC iteration using the ROCR package in R (Sing et al., 2005). To obtain the posterior mean and the pointwise 95% credible interval of ROC curve, we select 250 points equally spaced along the FP axis and take the vertical average or 95% quantiles of TP's at the 250 chosen points. This approach is referred to as vertical averaging of ROC curves at fixed FP rates by Fawcett (2006).

3. Simulations

We conduct simulation studies to evaluate the properties of the LC and MSRE modeling when the true and the assumed models may or may not be the same; i.e., the data could be generated under an LC model but analyzed using an MSRE model, and vice versa. We consider four scenarios for the longitudinal model with different levels of overlapping mixtures in both mean profiles and variance patterns, crossed with two primary-outcome models.

3.1. *Simulation Study Design*

For the longitudinal observations, we generate data for subject i from the following model with two mean profiles and two variance classes:

$$\begin{aligned}\mathrm{y}_{ij} | \boldsymbol{b}_i, \sigma_i^2 &\sim \mathrm{N}(b_{0i} + b_{1i} t_{ij}, \sigma_i^2); \\ \boldsymbol{b}_i &\sim \pi_d \mathrm{N}(\boldsymbol{\beta}_1, \boldsymbol{\Sigma}_1) + (1 - \pi_d)\mathrm{N}(\boldsymbol{\beta}_2, \boldsymbol{\Sigma}_2) \text{ and} \\ \log \sigma_i^2 &\sim \pi_c \mathrm{N}(\mu_1, \tau^2) + (1 - \pi_c)\mathrm{N}(\mu_2, \tau^2), \qquad (4)\end{aligned}$$

where $t_{ij} = 0, 1, ..., n_i$; $n_i \equiv 20$. For $k = 1, 2$, we let $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2})'$ and $\boldsymbol{\Sigma}_k$ have diagonal elements $(\omega_{k1}^2, \omega_{k2}^2)$ and correlation ρ_k. We let $\boldsymbol{\beta}_1 = (0, 0)'$ and $\boldsymbol{\beta}_2 = (2\sqrt{2}, 2\sqrt{2})'$, $\rho_1 = 0$, $\boldsymbol{\mu}_1 = -2$ and $\boldsymbol{\mu}_2 = -0.5$ in all scenarios. Thus the means of the two bivariate normals differ by 4 throughout, while the mean log-variances are separated by 1.5. Our four longitudinal model scenarios are defined by $(\rho_2, \omega^2, \tau^2)' =$

$(0.6, 2, 0.25)$, $(-0.6, 1, 0.25)$, $(0.6, 2, 0.06)$, and $(-0.6, 1, 0.06)$, respectively, where $\omega = \omega_{11} = \omega_{12} = \omega_{21} = \omega_{22}$.

Figure 1 shows the 95% contours for the two components in the mean profiles and the density plots of the log-variance classes in each of the four scenarios: both mean and variance classes heavily overlapping (scenario # 1), only the variance classes heavily overlapping (scenario # 2), only the mean classes heavily overlapping (scenario # 3), neither the mean nor the variance classes heavily overlapping (scenario # 4). In all scenarios, $\pi_d = 0.35$ and $\pi_c = 0.65$.

The following two underlying probit models are considered for health outcome:

(1) *LC probit submodel:*

$$\Phi^{-1}\left\{\mathrm{P}(o_i = 1)\right\} = \theta_0 + \theta_1 \mathrm{I}(D_i = 2) + \theta_2 \mathrm{I}(C_i = 2) + \theta_3 \mathrm{I}(D_i = 2, C_i = 2); \quad (5)$$

(2) *Multiple shared random effect (MSRE) probit submodel:*

$$\Phi^{-1}\left\{\mathrm{P}(o_i = 1)\right\} = \gamma_0 + \gamma_1 b_{0i} + \gamma_2 b_{1i} + \gamma_3 \sigma_i^2 + \gamma_4 b_{0i}\sigma_i^2 + \gamma_5 b_{1i}\sigma_i^2, \quad (6)$$

where $D_i = 1$ corresponds to the mean class $\mathrm{N}(\mathbf{(0, 0)'}, \mathbf{\Sigma}_1)$, and $C_i = 1$, the variance class $\mathrm{N}(-2, \tau^2)$ in the longitudinal model (4). We choose $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ for each scenario so that the outcome prevalence is approximately 50%.

To investigate the robustness of each approach under primary model-misspecification, we generated data from LC and MSRE primary models from equations (5) and (6) under each of the four longitudinal mixture scenarios, and then applied the approaches assuming the LC and MSRE structure to all generated data sets regardless of how the data were generated. For scenarios in which the true and assumed model differ, we generated observations from 10,000 subjects, obtained the corresponding maximum likelihood estimates (MLE) constructed under the assumed model, and repeated the process 1000 times to obtain the averages of the estimated parameters. We then used these average estimates as if they were the "true" parameters for the assumed structure under that simulation scenario. This practice allows us to compare the robustness for the two different modeling considerations under the same data-generation mechanism. For each scenario, we simulate 100 data sets of $n = 200$.

3.2. *Estimates of the Longitudinal Model*

First, we report the findings on regression associations and classification of LC membership, two aspects that play an explanatory role in accuracy of health-outcome prediction. The performances of estimation of the longitudinal parameters are reported in Tables A.1–A.4, Web Appendix C. When fitting true underlying models, we find that the performance of the LC approach is affected by how difficult it is to separate the mixture components in LC, though they tend to do better than the MSRE approach. When fitting misspecified models, both approaches are quite robust when there is sufficient information in the longitudinal data to separate classes. When the information from the longitudinal data is weak, the LC approach is more sensitive to model misspecification. Model misspecification also tends to damage the estimation of the mixture proportions in scenarios #1 and #3; even fitting a correctly assumed model still yields somewhat biased and under covered estimates of the mixture proportions. The variance components of the longitudinal model were generally well estimated under all scenarios.

3.3. *Estimates of the Primary Outcome Model*

For the study of regression association, we focus on the best (scenario #4) and the worst (scenario #1) scenarios in terms of the levels of mixture overlapping. Table 1 gives the Monte Carlo bias, standard deviation (SD), mean squared error (MSE), and 95% credible interval coverage (95% COV) for the corresponding association parameters under the correctly specified and mis-specified primary-outcome models. Recall that when the true and assumed models differ, the values reported under the "True" and "Bias" columns in the table refer to the corresponding large-sample MLEs and their discrepancies to the estimates given by fitting the assumed models. Such discrepancies can reflect how much the association between the longitudinal data and the health outcome of interest can be affected by model-misspecification. We clearly observed the association-correspondence from the correctly and mis-specified assumed models in scenario #4. For example, θ_2 and γ_3 always shared the same sign, indicating how the binary outcomes associate with the magnitude of subject-level residual variances, or a positive association between the outcome and a D_2 class in a true LC model is reflected by the positive values of targeted γ_1 and γ_2 in the assumed MSRE fit.

Examining the Monte Carlo bias and coverage probability of the 95% credible intervals for each parameter, we find that outcomes from fitting an MSRE model are not affected much by the levels of mixture overlapping nor by model-misspecification. In contrast, under scenario #1, if the true model is MSRE, the estimates of association parameters obtained by assuming LC can be far away from the targeted values and result in reduction of credible-sets coverage. The complete simulation results for all four scenarios are given in Tables A.1–A.4 (Web Appendix C).

3.4. *Misclassification Rates*

As the true class labels are known in our simulated data sets, we also consider the misclassification rates defined as the percentages of misclassified subjects when the classifications are based on $\tilde{C}_i = \mathrm{argmax}_c \hat{\pi}_{ic}^C$ and $\tilde{D}_i = \mathrm{argmax}_d \hat{\pi}_{id}^D$, where $\hat{\pi}_{ic}^C$ and $\hat{\pi}_{id}^D$ are the posterior means of $\mathrm{P}(D_i = d|\boldsymbol{y}, \boldsymbol{o})$ and $\mathrm{P}(C_i = c|\boldsymbol{y}, \boldsymbol{o})$, respectively.

In Table 2, we report the mis-classification rates for both mean and variance classes under scenarios # 1–4 and different combinations of true and assumed LC and MSRE models. The LC model tends to perform reasonably well when correctly specified. When the two mixture components are well separated, both approaches perform well regardless of model specification. Variance classes are generally well estimated, with some modest reduction in accuracy for overlapping components. Such results are robust against model-misspecification. For the overlapping mean classes in scenarios #1 and #3, the

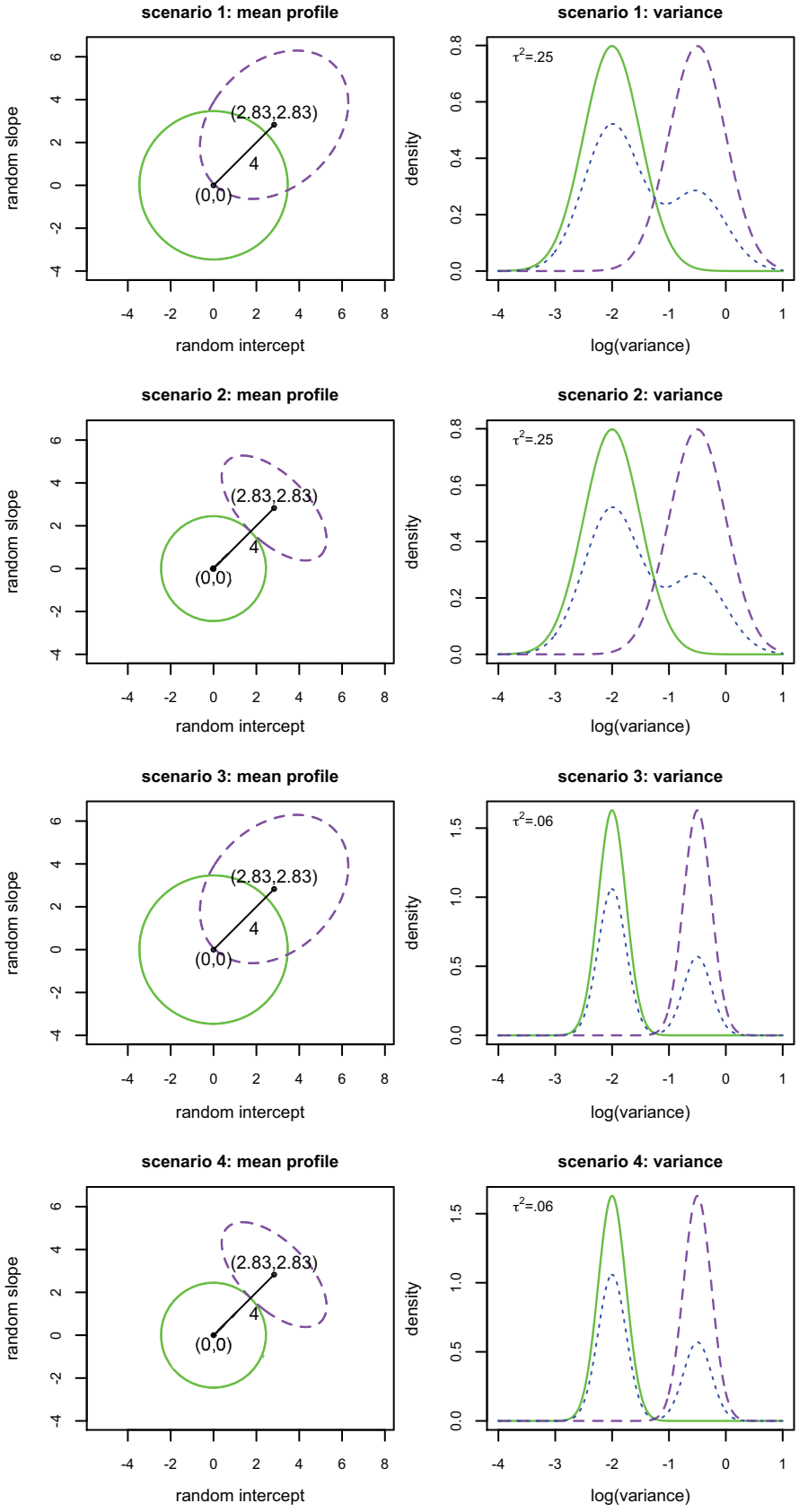

Figure 1. Simulation setup for the mean profiles and variance classes: left column: 95% contour plots of the two components for mean profile class; right column: density plots of the two components for variance class (dotted curves are the density curves for the variances).

Table 1
Estimates of the association parameters in the primary outcome model from the simulation study based on 100 datasets of size, $n = 200$

Assumed structure		True LC structure					True MSRE structure				
		TRUE	BIAS	SD	RMSE	95% COV	TRUE	BIAS	SD	RMSE	95% COV
(a) Generated from longitudinal scenario # 1											
LC	θ_0	−0.80	−0.65	0.50	0.82	0.89	−0.40	−1.00	1.13	1.51	0.35
	θ_1	1.80	0.61	0.66	0.90	0.88	−0.11	2.40	2.34	3.35	0.35
	θ_2	−0.20	0.20	0.69	0.72	0.98	0.53	1.64	1.76	2.41	0.36
	θ_3	−0.30	−0.28	0.82	0.87	0.97	0.16	−3.69	3.73	5.24	0.35
MSRE	γ_0	−0.32	0.00	0.21	0.21	0.95	−1.00	0.19	0.24	0.30	0.92
	γ_1	0.20	0.01	0.11	0.11	0.96	1.00	−0.09	0.16	0.18	0.95
	γ_2	0.18	−0.01	0.11	0.11	0.96	−1.00	0.04	0.17	0.18	0.96
	γ_3	−0.22	−0.15	0.60	0.62	0.92	2.00	−0.52	0.58	0.78	0.87
	γ_4	−0.04	0.01	0.32	0.32	0.93	−2.00	0.29	0.36	0.46	0.90
	γ_5	−0.04	0.06	0.30	0.30	0.94	2.00	−0.14	0.38	0.41	0.95
(b) Generated from longitudinal scenario # 2											
LC	θ_0	−0.80	−0.06	0.25	0.25	0.98	−0.48	−0.19	0.30	0.36	0.89
	θ_1	1.80	0.13	0.37	0.39	0.98	0.06	0.03	0.45	0.45	0.92
	θ_2	−0.20	−0.07	0.52	0.52	0.96	0.65	0.52	0.66	0.84	0.84
	θ_3	−0.30	−0.05	0.69	0.69	0.96	−0.08	−0.09	0.91	0.92	0.87
MSRE	γ_0	−0.66	0.01	0.24	0.24	0.98	−1.00	0.07	0.24	0.25	0.97
	γ_1	0.28	−0.02	0.12	0.12	0.96	1.00	−0.05	0.17	0.17	0.94
	γ_2	0.28	0.02	0.11	0.11	0.94	−1.00	0.02	0.17	0.17	0.93
	γ_3	−0.22	−0.28	0.56	0.62	0.97	2.00	−0.33	0.59	0.68	0.96
	γ_4	−0.05	0.12	0.33	0.35	0.89	−2.00	0.17	0.37	0.40	0.95
	γ_5	−0.05	−0.02	0.29	0.30	0.94	2.00	−0.06	0.39	0.39	0.97
(c) Generated from longitudinal scenario # 3											
LC	θ_0	−0.80	−0.60	0.42	0.74	0.87	−0.41	−1.02	1.15	1.54	0.34
	θ_1	1.80	0.53	0.55	0.77	0.91	−0.12	2.52	2.28	3.40	0.31
	θ_2	−0.20	0.10	0.62	0.63	1.00	0.57	1.30	1.50	1.98	0.34
	θ_3	−0.30	−0.11	0.77	0.77	1.00	0.15	−3.42	3.15	4.65	0.36
MSRE	γ_0	−0.28	−0.01	0.20	0.20	0.96	−1.00	0.16	0.22	0.28	0.96
	γ_1	0.19	0.02	0.13	0.13	0.89	1.00	−0.09	0.15	0.18	0.92
	γ_2	0.20	0.00	0.14	0.14	0.93	−1.00	0.04	0.16	0.17	0.92
	γ_3	−0.37	−0.12	0.47	0.49	0.98	2.00	−0.42	0.53	0.68	0.94
	γ_4	−0.07	−0.10	−0.03	0.34	0.92	−2.00	0.25	0.35	0.43	0.94
	γ_5	−0.07	−0.04	0.04	0.36	0.93	2.00	−0.10	0.35	0.37	0.96
(d) Generated from longitudinal scenario # 4											
LC	θ_0	−0.80	−0.01	0.19	0.19	0.97	−0.50	−0.05	0.24	0.25	0.91
	θ_1	1.80	0.00	0.25	0.25	0.99	0.06	0.05	0.30	0.30	0.90
	θ_2	−0.20	−0.08	0.51	0.52	0.95	0.69	0.10	0.38	0.40	0.94
	θ_3	−0.30	0.09	0.57	0.58	0.95	−0.08	−0.09	0.49	0.49	0.95
MSRE	γ_0	−0.62	−0.01	0.23	0.23	0.98	−1.00	0.14	0.25	0.29	0.93
	γ_1	0.29	−0.01	0.12	0.12	0.97	1.00	−0.08	0.15	0.17	0.90
	γ_2	0.29	0.03	0.13	0.13	0.94	−1.00	0.05	0.17	0.18	0.93
	γ_3	−0.36	−0.16	0.69	0.71	0.95	2.00	−0.44	0.58	0.73	0.93
	γ_4	−0.09	0.07	0.35	0.35	0.95	−2.00	0.22	0.36	0.43	0.90
	γ_5	−0.08	−0.02	0.30	0.30	0.97	2.00	−0.10	0.42	0.43	0.94

use of MSRE tends to result in high-misclassification rates even if the model is well specified, but these rates are higher when fitting an LC model under an MSRE structure. However, a high mis-classification rate obtained under an assumed MSRE model, such as those in scenario #1, does not associate with deteriorated performances in estimating association parameters in Table 1. In contrast, a high mis-classification rate obtained by an assumed LC model, particularly under model-misspecification, does.

3.5. *Predictive Accuracy*

We next turn our attention to evaluating the predictive accuracy of outcome using the same setups. We evaluate the true AUC (i.e., the AUC for the true model computed with the known parameters) and the corresponding values predicted by assuming an LC or an MSRE model, respectively. The means and 2.5/97.5 percentiles of the posterior mean AUCs based on repeated samples are given in Table 3a. Besides the true AUC, the rows in Table 3a summarize outcomes from

Table 2
Misclassification rates (%) for the mean profile and variance class memberships from the simulation study based on 100 datasets of size, $n = 200$

Assumed model	True LC model				True MSRE model			
	Scenario				Scenario			
	# 1	# 2	# 3	# 4	# 1	# 2	# 3	# 4
Mean profile class								
LC	12	0	11	0	50	1	50	1
MSRE	33	0	34	1	33	0	34	1
Variance class								
LC	11	11	3	3	12	13	3	3
MSRE	10	11	3	3	11	11	3	3

within-sample (training) and out-of-sample (testing) predictions. Additional independent data sets of size $\tilde{n} = 50$ were generated from the same model as the testing sets. Under model-misspecification, we also reported AUC obtained when the true LC membership/random effects are used to build predictions for the assumed MSRE/LC models (i.e., "assumed" AUC at the last two tables of Web Appendix C). The differences between the "training" and "assumed" AUC reflect the effects attributed to the estimated class-memberships.

When the MSRE model is assumed, the AUC outcomes either show a little loss of predictive power (only under misspecification) or the results are close to the true AUC. The slightly lower AUC values of testing samples, in comparison to those of training, are as expected. When fitting correctly specified LC model, the empirical 95% credible intervals of AUC under scenarios #1 and #3 are wider than the truth, while such intervals under scenarios #2 and #4 are of similar length to the truth, reflecting the larger variabilities in the predictive power for settings under overlapping mean components.

When the MSRE model is the truth and the LC model is used, the outcomes in "LC-testing" suggest that the average posterior means given in "LC-training" could be overly optimistic, except for scenario #4. In addition, there again exists considerably large variation, indicated by the wide credible intervals in scenarios #1 and #3, corresponding to the deteriorated performances we observed in Table 1. Under scenario #1, Figure A.1 (Web Appendix C) presents two typical data examples that have either very high (top panel) or very low (bottom panel) AUC estimated by the LC model when the truth is the joint MSRE model. In both examples, the AUC's by the correctly specified MSRE model are very close to the truth. However, the high AUC by LC suggests that the LC model has some ability to create "outcome-informed clusters" and deliver overly optimistic within-sample prediction under model-misspecification. This finding is also revealed by the differences between the values of "LC-training" and "LC-assumed" (measuring the effects due to estimated cluster-memberships) reported in Web Appendix C.

The phenomenon is a unique feature of joint LC modeling, and is partly due to the difficulty in determining cluster-memberships and partly due to the fact that the mixture classification is done given both the longitudinal $\boldsymbol{y}$ and the outcome $\boldsymbol{o}$. When the information to divide clusters in $\boldsymbol{y}$ is relatively weak, the binary outcome $\boldsymbol{o}$ tends to dominate in determining the latent classes to boost the posterior density. With the outcome $\boldsymbol{o}$ being binary, classes were created to match the two groups of $o = 0$ and $o = 1$. This results in the predictive power of future longitudinal data being over-estimated, as the prediction under the current data only weakly relies on it. Figure A.1 illustrates this phenomenon. This phenomenon for joint LC modeling also happens when the data are generated from the LC model, but the effect is much less prominent. To our knowledge this phenomenon has not been previously noted in the literature and it could have strong implications for outcome interpretation. On the other hand, when almost all subjects are being assigned to one mean class by the LC model, prediction of the outcome is solely dependent on the variance class and consequently the LC model had low predictive performance. The existence of these two typical cases in Figure A.1 leads to overly inflated variation for LC estimated AUC's. We also report the corresponding outcomes for Brier Score (Brier, 1950) in Table 3b, which re-enforce the findings obtained using AUC.

Finally, all simulations are repeated with $n = 500$; see Tables A.5–A.10 (Web Appendix C) for results. The outcomes are consistent with the findings of $n = 200$, with notably reduced bias and RMSE of the estimates of all model parameters and reduce mis-classification rates of class memberships when the true and assumed models are the same.

4. Analysis of Penn Ovarian Aging Study Data

One goal of the Penn Ovarian Aging Study is to determine to what extent the annually FSH levels are predictive of the risk of SHF. Out of the 436 women in the study, we restrict our analysis to the 245 who (a) had not experienced SHF at baseline and (b) had at least three measurements of FSH. Hormone values were treated as missing if a woman was pregnant, breast feeding or taking exogenous hormones during the study period. A total of 4244 FSH values were observed, ranging from 3 to 26 per woman. Of the 245 women without SHF symptoms at baseline, 118 (48.2%) had experienced SHF at least once during the study.

After removing the population level non-linear trend by subtracting the loess estimate of mean FSH by age, we seek to evaluate whether each individual's deviation from it, postulated by the subject-level random coefficients in an orthogonal polynomial model, is associated with SHF. We let y_{ij} denote the detrended log(FSH) lowess residuals (Figure A.2, Web Appendix D) and o_i denote the SHF indicator: $o_i = 1$ if any SHF score ≥ 2 during study. Preliminary analysis by linear mixed effects (LME) modeling indicates that a random intercept and random slope model is sufficient to capture the trends in the residual trajectories. Thus we let $f(\boldsymbol{b}_i; t_{ij}) = b_{i0} + b_{i1} t_{ij}$, where t_{ij} is the linear term in the orthogonal polynomial used in the LME modeling, and b_{i0} and b_{i1} are the subject-level random intercepts and slopes, respectively. We then jointly model the FSH mean profile and residual variance to predict the risk of SHF using models in (1) and (2). We examine the use of the primary probit LC and MSRE models under the joint modeling framework, as

Table 3

(a) Mean area under the ROC curves and (b) Brier score for the prediction of outcome from the simulation study based on 100 datasets of size, $n = 200$. Left columns: data generated from the LC model; right columns: data generated from the MSRE model. "Percentile" refers to the 2.5 and 97.5 percentiles of the results computed under the true parameters across simulations; "95% CI" refers to mean of the lower and upper 95% credible intervals across simulations. LC/MSRE-testing refers to results obtained for the validation sample of size $\tilde{n} = 50$, while LC/MSRE-training gives within-sample prediction outcomes

	TRUE: joint LC model				TRUE: joint MSRE model			
	Scenario				Scenario			
	# 1	# 2	# 3	# 4	# 1	# 2	# 3	# 4
(a) Area under the ROC curves								
Truth								
Mean	0.80	0.81	0.81	0.81	0.84	0.85	0.83	0.84
Percentile	(0.75, 0.86)	(0.75, 0.86)	(0.75, 0.87)	(0.75, 0.86)	(0.79, 0.89)	(0.80, 0.90)	(0.77, 0.88)	(0.78, 0.89)
LC-training								
Mean	0.80	0.82	0.80	0.81	0.85	0.69	0.83	0.64
95% CI	(0.58, 0.91)	(0.75, 0.88)	(0.63, 0.92)	(0.75, 0.86)	(0.63, 0.97)	(0.58, 0.82)	(0.60, 0.96)	(0.56, 0.72)
LC-testing								
Mean	0.67	0.79	0.68	0.79	0.64	0.59	0.66	0.61
95% CI	(0.54, 0.79)	(0.69, 0.9)	(0.59, 0.78)	(0.67, 0.89)	(0.53, 0.73)	(0.49, 0.7)	(0.58, 0.74)	(0.49, 0.77)
MSRE-training								
Mean	0.76	0.80	0.77	0.81	0.84	0.85	0.83	0.83
95% CI	(0.69, 0.83)	(0.73, 0.85)	(0.71, 0.85)	(0.74, 0.86)	(0.79, 0.89)	(0.79, 0.90)	(0.76, 0.88)	(0.77, 0.89)
MSRE-testing								
Mean	0.74	0.78	0.75	0.79	0.78	0.8	0.78	0.79
95% CI	(0.59, 0.89)	(0.65, 0.9)	(0.61, 0.88)	(0.67, 0.89)	(0.66, 0.88)	(0.68, 0.89)	(0.64, 0.89)	(0.65, 0.9)
(b) Brier score								
Truth								
Mean	0.16	0.16	0.16	0.16	0.16	0.15	0.16	0.16
Percentile	(0.13, 0.2)	(0.13, 0.19)	(0.13, 0.19)	(0.13, 0.19)	(0.13, 0.18)	(0.13, 0.18)	(0.14, 0.19)	(0.13, 0.19)
LC-training								
Mean	0.15	0.16	0.15	0.16	0.12	0.2	0.14	0.22
95% CI	(0.1, 0.23)	(0.13, 0.19)	(0.09, 0.19)	(0.13, 0.19)	(0.04, 0.23)	(0.15, 0.24)	(0.05, 0.23)	(0.2, 0.24)
LC-testing								
Mean	0.26	0.27	0.25	0.26	0.26	0.27	0.25	0.25
95% CI	(0.23, 0.33)	(0.22, 0.34)	(0.23, 0.28)	(0.2, 0.31)	(0.22, 0.31)	(0.22, 0.32)	(0.22, 0.29)	(0.22, 0.28)
MSRE-training								
Mean	0.19	0.17	0.19	0.17	0.16	0.15	0.16	0.16
95% CI	(0.16, 0.22)	(0.14, 0.2)	(0.16, 0.21)	(0.14, 0.2)	(0.14, 0.18)	(0.12, 0.18)	(0.14, 0.2)	(0.14, 0.19)
MSRE-testing								
Mean	0.26	0.26	0.26	0.26	0.25	0.26	0.26	0.25
95% CI	(0.22, 0.3)	(0.21, 0.31)	(0.22, 0.31)	(0.22, 0.31)	(0.21, 0.32)	(0.21, 0.32)	(0.21, 0.31)	(0.22, 0.29)

presented in Section 2. We also adjust for additional baseline covariates log(BMI) and smoking status in both models.

For all models, we ran three MCMC chains of 50,000 iterations, discarded the first 10,000 iterations as burn-in, and only retained every 10^{th} draw to reduce autocorrelation. We assessed chain-convergence by the Gelman–Rubin statistic $\hat{R}$. The maximum value among all parameters was less than 1.1, indicating convergence. Given the moderate sample size $n = 245$, we considered the models in (1) with K_D and K_C being 1–3. The K_D and K_C selected by DIC and LPML differed, with LPML preferring more mixture components (Table A.12, Web Appendix D); a typical behavior of LPML in our additional simulation outcomes (not shown). The best model selected by DIC had $K_D = 1, K_C = 2$, with a two-class model of a $K_D = 2, K_C = 2$ close second for both the MSRE and LC models.

Figure 2 shows the mean and variances for the $K_D = 1, K_C = 2$ model (left: MSRE; right: LC), indicating the bimodal nature of the posterior means of the individual variances. Table 4 reports the results for fitting both MSRE and LC models when $K_D = 1, K_C = 2$; see Table A.13, Web Appendix D, for results assuming $K_D = 2, K_C = 2$. The estimation of the longitudinal submodel differs little between an LC and an MSRE fit. The two-class mean model separates the mean trajectories into two approximately equal-sized classes, with one a "null class" with slope and intercept near zero, the other a "high and rising" (Figure A.3, Web Appendix D) class with the slope/intercept being 0.21/0.16 and 0.22/0.17 under

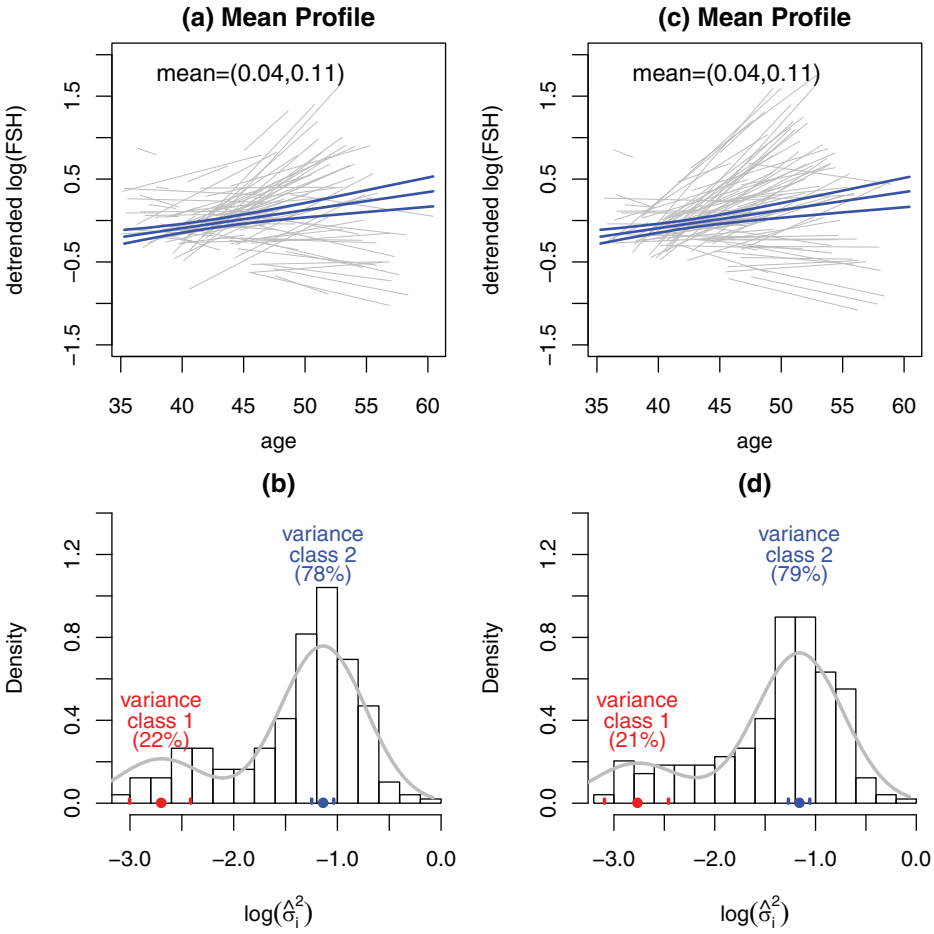

Figure 2. Posterior pointwise 95% credible intervals for the mean profile classes and the histograms of log-variances in the analysis of Penn Ovarian Aging data with $K_D = 1, K_C = 2$: (a) and (b): under the joint MSRE model and (c) and (d): under the joint LC model.

the MSRE and LC models, respectively. Both LC and MSRE modeling implies a reduced risk of SHF for the "high and rising" class of FSH, albeit not being significant. The MSRE outcomes further indicate that the subject-level random intercept for the residual FSH measures is non-significant.

All models suggest that a little more than one in five women (22% under MSRE, 21% under LC) belong to a low residual variance class, centered at 0.07(MSRE)/0.06(LC), while the remainder belong to a higher variance class, centered at 0.32(MSRE)/0.31(LC). Both MSRE and LC models suggest a positive and highly significant association between subject-level variance and risk of SHF while adjusting for baseline covariates of smoking and BMI.

For a non-smoking woman at mean BMI of 27.7 with FSH slope and intercept at the population mean, the probability of experiencing a SHF under the MSRE model with $K_D = 1, K_C = 2$ is 30.5% (19.7%, 42.0%) and 45.9% (38.2%, 54.1%), respectively, if her residual variance is at the Class 1/Class 2 mean. The difference is greater under the LC model with outcome probabilities become 17.9% (5.0%, 32.9%) and 51.5% (43.0%, 60.1%). No significant interactions between subject-level means and residual variances were found among models with $K_D = 1$ or 2. (Table A.14, Web Appendix D). All models provide marginal evidence to support smoking at baseline as contributing to higher risk of SHF, while the effect from baseline BMI is non-significant.

For the joint MSRE and LC models, we conducted model-checking via PPD p values (PPD-p's). The corresponding histograms are given in Figure A.4 (Web Appendix D). For $K_D = 1, K_C = 2$, the longitudinal detrended log(FSH), the ranges and medians of PPD-p's are $(0.06, 0.93)$ and 0.54 (MSRE) and $(0.09, 0.92)$ and 0.54 (LC), respectively. The contrasts between the individual fits from the top ($0.1 \leq$ PPD-p's ≤ 0.9) and bottom (otherwise) panels of Figure A.5 suggest that the small PPD-p's appear to be driven by the individual outlying points and large PPD-p's are caused by the "almost perfect" fits. The goodness of fit for the FSH trajectories and the risk of SHF, are further supported by Figure A.6, which shows that only about 4% of the FSH values are not covered by the 95% subject-level posterior predictive intervals and by the PPD-p of 0.497 and 0.498, under primary-outcome MSRE and LC models, respectively. Finally, we found that the MSRE model had somewhat greater predictive power than the LC model, with the posterior means of AUC = 0.682 for former, and 0.645 for latter; the ROC curves are provided in Figure A.7. A comparison of AUCs suggests that the difference in

Table 4
Posterior estimates of the model parameters under the joint MSRE and LC models in the analysis of Penn Ovarian Aging data with $K_D = 1, 2$ and $K_C = 1, 2$

	MSRE model			LC model		
	Mean	SE	95% CI	Mean	SE	95% CI
β_{11}	0.040	0.031	$(-0.020, 0.101)$	0.038	0.031	$(-0.024, 0.099)$
β_{12}	0.110	0.025	(0.061, 0.158)	0.109	0.025	(0.060, 0.157)
ω^2_{11}	0.200	0.023	(0.160, 0.250)	0.200	0.023	(0.160, 0.249)
ω^2_{12}	0.102	0.013	(0.079, 0.130)	0.103	0.013	(0.080, 0.132)
ρ_1	0.668	0.055	(0.552, 0.767)	0.668	0.056	(0.551, 0.768)
μ_1	-2.699	0.149	$(-3.004, -2.416)$	-2.767	0.162	$(-3.094, -2.459)$
μ_2	-1.138	0.054	$(-1.247, -1.035)$	-1.160	0.054	$(-1.269, -1.057)$
τ^2	0.171	0.040	(0.105, 0.262)	0.191	0.043	(0.120, 0.287)
π_1^C	0.225	0.040	(0.150, 0.305)	0.212	0.039	(0.140, 0.292)
γ_0 (intercept)	-0.457	0.953	$(-2.327, 1.430)$			
γ_1 (log(BMI))	-0.065	0.284	$(-0.625, 0.493)$			
γ_2 (smoking)	0.375	0.186	(0.011, 0.746)			
$\gamma_3(b_{0i})$	-0.889	0.322	$(-1.546, -0.286)$			
$\gamma_4(b_{1i})$	0.753	0.467	$(-0.137, 1.694)$			
$\gamma_5(\sigma_i^2)$	1.627	0.592	(0.515, 2.831)			
θ_0 (intercept)				-0.826	0.946	$(-2.670, 1.011)$
θ_1 (log(BMI))				-0.041	0.280	$(-0.587, 0.498)$
θ_2 (smoking)				0.330	0.184	$(-0.036, 0.691)$
θ_3 $(D = 2)$						
θ_4 $(C = 2)$				1.000	0.326	(0.437, 1.717)

the performance was not clearly delineated (ΔAUC is 0.037 $(-0.039, 0.114)$).

5. Concluding Remarks

In this article, we study two joint modeling approaches, LC and MSRE, to link the important characteristics or features in the longitudinal trajectories to the primary health outcome when the underlying true model may or may not be the model used to analyze the data. Both LC and MSRE models are built upon certain modeling assumptions whose violations may not be easily detected using popular model-selection/diagnostic approaches. However, relatively little attention has been paid to the potential impact of model misspecification in the joint modeling framework. This work provides guidance concerning the potential impact of choosing one of the LC and MSRE modeling strategies to link longitudinal measurements and health outcome while the other model generates the data.

Our simulation study showed that the MSRE model had several strengths over that of the LC model. First, it was not as sensitive to model misspecification as the LC approach. In addition, the MSRE approach was not as sensitive as the LC approach to failures to clearly separate the latent classes because correct class assignment is more critical to estimating the outcome-model association parameters under LC modeling strategy. In terms of prediction, the misspecified MSRE AUC measure was almost identical to the truth while LC approach suffered considerable loss of predictive power when misspecified. Furthermore, for overlapping mixture components, the misspecified LC AUC computed based on within-sample classification could lead to an over-optimistic impression of prediction power because of the creation of outcome-informed clusters. This phenomenon is a consequence of difficulties in identifying cluster memberships. The LC model did have several strengths relative to the MSRE model. For the estimation of longitudinal parameters themselves, the LC approach could outperform MSRE, which performed poorly when the components of the latent classes are not well separated. Also, the LC model has the advantages of summarizing complex multivariate prediction features into a much simpler form. When the resulting latent classes are easily interpretable, the LC model allows one to relate the outcome risk to meaningful features identified by the various latent classes. A final feature of note from our simulation study was that the LC model was more sensitive to LC misclassification and outcome parameter estimation bias when the mean classes were not well separated than when the variance classes were not well separated.

Both modeling strategies gave similar results when applied to the Penn Ovarian Aging study. There was no strong evidence of clustering among the mean FSH hormone trajectories, nor strong evidence that subject-level variability in these trajectories was associated with risk of severe hot flash. In contrast, residual variances did group into a low- (∼20%) and high- (∼80%) variance class, with both the MSRE and LC models showing that lower variances were associated with very substantial declines in risk of SHF.

This work can be extended in a variety of ways. For example, the assumption of a low-order polynomial function for the longitudinal predictors could be relaxed to allow for a penalized spline or functional regression model. This may provide a more non-parametric parsing of "short term" and "long term" subject-level variability, if sufficient data are available at the subject-level to allow estimation of

such terms. Also, developing methods to compensate for missing data in both the longitudinal predictors and outcome measures, particularly under non-missing-at-random mechanisms, will have practical application as well.

6. Supplementary Materials

Web Appendices A–D referenced in Sections 2.1, 2.2, 3, and 4; and C++/R codes to implement our LC and MSRE methods are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

This work was supported in part by Grants R03AG031980 and R01CA74552 from the National Institutes of Health. The authors thank Prof Ellen Freeman for sharing her data.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1–3.

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

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

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

Elliott, M. R., Gallo, J. J., Ten Have, T. R., Bogner, H. R., and Katz, I. R. (2005). Using a Bayesian latent growth curve model to identify trajectories of positive affect and negative events following myocardial infarction. *Biostatistics* **6**, 119–143.

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

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861–874.

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

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. *Archives of General Psychiatry* **63**, 375.

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.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**, 153–160.

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

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modeling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–480.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian survival analysis*. New York: Springer-Verlag.

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

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). *Bayesian Analysis* **1**, 535–542.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

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

Proust-Lima, C., Séne, M., Taylor, J. M., and Jacqmin-Gadda, H. (2012). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research*.

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

Sammel, M. D., Wang, Y., Ratcliffe, S. J., Freeman, E. W., and Propert, K. J. (2001). Models for within-subject heterogeneity as predictors for disease. In *Proceedings of the Joint Meeting of the American Statistical Association*, Atlanta, GA, August 6.

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **64**, 583–639.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **62**, 795–809.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* **14**, 809–834.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91**, 217–221.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the precence of informative right censoring by modeling the censoring process. *Biometrics* **44**, 175–188.

Received August 2013. *Revised November* 2014.
Accepted December 2014.