Statistics and Its Interface Volume 9 (2016) 183–201

Bayesian model assessments in evaluating mixtures of longitudinal trajectories and their associations with cross-sectional health outcomes

Bei Jiang*, Michael R. Elliott, Mary D. Sammel, and Naisyin Wang

In joint-modeling analyses that simultaneously consider a set of longitudinal predictors and a primary outcome, the two most frequently used response versus longitudinal-trajectory models utilize latent class (LC) and multiple shared random effects (MSRE) predictors. In practice, it is common to use one model assessment criterion to justify the use of the model. How different criteria perform under the joint longitudinal predictor-scalar outcome model is less understood. In this paper, we evaluate six Bayesian model assessment criteria: Akaike information criterion (AIC) (Akaike, 1973), Bayesian information criterion (BIC) (Schwartz, 1978), integrated classification likelihood criterion (ICL) (Biernacki et al., 1998), the deviance information criterion (DIC) (Spiegelhalter et al., 2002), the logarithm of the pseudomarginal likelihood (LPML) (Geisser and Eddy, 1979) and the widely applicable information criterion (WAIC) (Watanabe, 2010). When needed, the criteria are modified, following the Bayesian principle, to accommodate the joint modeling framework that analyzes longitudinal predictors and binary health outcome data. We report our evaluation based on empirical numerical studies, exploring the relationships and similarities among these criteria. We focus on two evaluation aspects: goodness-of-fit adjusted for the complexity of the models, mostly reflected by the numbers of latent features/classes in the longitudinal trajectories that are part of the hierarchical structure in the joint models, and prediction evaluation based on both training and test samples as well as their contrasts. Our results indicate that all six criteria suffer from difficulty in separating deeply overlapping latent features, with AIC, BIC, ICL and WAIC outperforming others in terms of correctly identifying the number of latent classes. With respect to prediction, DIC, WAIC and LPML tend to choose the models with too many latent classes, leading to better predictive performance on independent validation samples than the models chosen by other criteria do. An interesting result concerning the wrong model choice will be reported. Finally, we use the results from the simulation study to identify the suitable candidate models to link the useful features in the follicle stimulating hormone trajectories to predict risk of severe hot flash in the Penn Ovarian Aging Study.

Keywords and phrases: Bayesian model assessment, Joint models, Latent class, Shared random effect, WAIC, ICL, DIC, LPML, AIC, BIC, Out-of-sample validation.

*Corresponding author.

1. INTRODUCTION

There is a growing body of literature that models information from longitudinal data to predict risks of health outcome of interest (Taylor et al., 2005; Yu et al., 2008; Proust-Lima and Taylor, 2009; Rizopoulos, 2011; Proust-Lima et al., 2012; Elliott et al., 2012; Taylor et al., 2013; Jiang et al., 2015). An attractive feature of such predictions is that they are individualized. However, when several candidate models are available, the derived outcomes can be greatly affected by the use of different models. For example, different numbers of mixture components as well as the assumed association structure (e.g., multiple shared random effect vs. latent class structures) to link the longitudinal and primary outcome submodels can affect the target individualized predictions (Jiang et al., 2015). Various model selection criteria can be adopted to guide the selection of the proper number of components and the association structure. However, the performance of these model selection criteria have not been carefully studied in the joint modeling framework.

Choosing the number of mixture components in a finite mixture setting is a non-trivial task. The difficulty arises mainly because the estimation in finite mixture models is not a regular problem but a singular problem; hence the log-likelihood function is not well approximated by a quadratic function, and maximum likelihood estimates are not asymptotically normal. See McLachlan and Peel (2000, section 6), Frühwirth-Schnatter (2006, section 4), Steele and Raftery (2010) as well as the references therein for thorough discussions of parameter estimation in finite mixture models.

Here we consider six model selection criteria. They are Akaike information criterion (AIC) (Akaike, 1974), Bayesian information criterion (BIC) (Schwartz, 1978), integrated classification likelihood criterion (ICL) (Biernacki et al., 1998), the deviance information criterion (DIC) (Spiegelhalter et al., 2002), the logarithm of the pseudomarginal

likelihood (LPML) (Geisser and Eddy, 1979) and the widely applicable information criterion (WAIC) (Watanabe, 2010). AIC and BIC are the longest standing and most commonly used information-based model selection criteria in general; ICL is closely related to BIC with focus on the classification likelihood and entropy, whereas LPML has been most widely used in Bayesian model assessment. More details, including certain necessary modifications to accommodate the joint modeling framework, are provided in Section 3.1. DIC is often viewed as a Bayesian version of AIC with prior information on model parameters and is equivalent to AIC for non-hierarchical models with non-informative or flat priors. Many authors have proposed alternative versions of DIC. For example, Plummer (2002, 2006) and Gelman et al. (2003) proposed alternative definitions of model complexity, while Celeux et al. (2006) proposed eight variations of DIC for "missing data" problems, including hierarchical models with latent variables.

WAIC, a recently proposed approach, was derived based on singular learning theory (Watanabe, 2009) as an asymptotically unbiased approximation to the out-of-sample prediction error, and is a generalization of AIC that is applicable for both regular and singular statistical models. It is straightforward to compute based on the posterior draws, even for complex hierarchical models. Gelman et al. (2013) discussed the construction of AIC, DIC and WAIC from a Bayesian perspective using some simple examples and concluded that WAIC is a "very fast and computationally-convenient alternative" to their most preferred but often computationally expensive cross-validation approach to choose among several candidate models. However, its properties have not been studied in the setting of choosing numbers of components for finite mixture distributions.

Model assessment for joint models that incorporate mixture distributions as considered in Jiang et al. (2015) is even more challenging for various reasons. First of all, the variables that are assumed to have mixture distributions are unobserved latent features of the longitudinal trajectories. AIC, BIC or DIC based on the observed data likelihood may not be available in closed form. Secondly, the evaluation of model goodness-of-fit has to take into account the model fits of the longitudinal submodel and the primary outcome model jointly; this can be problematic, as the relatively larger gain in the fit of the primary outcome model, which contains a larger number of components, may dominate the overall model fit. This phenomenon could lead to favoring models with larger numbers of latent classes. Thirdly, when the true data generating model is a multiple shared random effect (MSRE) model, incorrectly assuming a latent class (LC) structure to link the longitudinal submodel and the primary outcome model has a high chance of creating an "outcome-informed artifact" as reported previously in Jiang et al. (2015). When the primary outcome is binary and the information about the mixture components from the longitudinal data is weak, artificial mixture-components are created to match the two outcome groups of 0 or 1 under the assumed LC structure, which could lead to over-estimation of prediction accuracy. This condition suggests two questions: (i) whether such artifacts due to the assumed structure in the primary outcome model would play any role in the performance of model selection criteria in choosing the numbers of mixture components; and (ii) whether model selection criteria would favor the assumed LC structure, which leads to seemingly better prediction over the MSRE structure in the presence of outcome-informed artifacts.

To address these issues, we conduct numerical studies to compare and contrast the performance of several commonly used model selection criteria. We consider WAIC and other modified criteria based on Bayesian principles in the joint modeling setting considered in Jiang et al. (2015). Our main goals consist of understanding the performance of these commonly used criteria under different scenarios, including when the data-generating scheme differs from the assumed structure; gaining insights on similarity of selection performance of different criteria; and uncovering the model predictive performance based on out-of-sample validation, where the performance of the selected models are further linked with model selection criteria.

The remainder of this article is organized as follows. In Section 2, we describe the joint LC and MSRE models with mixture distributions for the mean profiles and residual variances of longitudinal trajectories. In Section 3, we briefly review the Bayesian model assessment criteria as well as the overall model predictive performance measure to be included in our simulation. In Section 4, we describe our simulation study and report the outcomes. In Section 5, we describe the procedures used to validate predictive performance of the selected models by different criteria using newly generated independent samples. In Section 6, we study the performance of the selection criteria for the joint modeling of the follicle stimulating hormone trajectories and severity of hot flash for a group of middle-aged women from the Penn Ovarian Aging Study. Concluding remarks are given in Section 7.

2. JOINT LC AND MSRE MODELS

Mixture modeling is commonly used to identify unique and distinct feature subgroups (i.e., latent classes) in longitudinal trajectories, e.g., the proposal of growth mixture models (GMMs) in Muthén and Shedden (1999). Jiang et al. (2015) considered two classes of joint models for normally distributed longitudinal data and a binary health outcome data. Both models used a generalized GMM for the longitudinal data. GMMs assume latent classes for the subject-level mean profiles. Our extension in Jiang et al. (2015) considered latent classes for not only the mean profiles but also the residual variabilities of the longitudinal trajectories. Specifically, the longitudinal submodel has the form

$$y_{ij}|\boldsymbol{b}_i, \sigma_i^2 \sim \mathrm{N}\{\mu(\boldsymbol{b}_i; t_{ij}), \sigma_i^2\}, \tag{1}$$

where, y_{ij} denotes the longitudinal covariate for the i^{th} subject at time t_{ij}, $j = 1, ..., n_i$, $i = 1, ..., n$, $\boldsymbol{b}_i$ is the vector of

r random effects that reflect the subject-level mean profile/trajectory patterns and σ_i^2 is the subject-level residual variance. $\boldsymbol{D}_i = (D_{i1}, \cdots, D_{iK_D})$ and $\boldsymbol{C}_i = (C_{i1}, \cdots, C_{iK_C})$ define the latent class memberships for the individual mean profile and variance respectively:

$$
\begin{aligned}
\boldsymbol{D}_i &\sim \text{Multinomial}(1, \pi_1^D, ..., \pi_{K_D}^D), \\
\boldsymbol{b}_i | D_{id} = 1 &\sim \text{N}(\boldsymbol{\beta}_d, \boldsymbol{\Sigma}_d), d = 1, ..., K_D; \\
\boldsymbol{C}_i &\sim \text{Multinomial}(1, \pi_1^C, ..., \pi_{K_C}^C), \\
\sigma_i^2 | C_{ic} = 1 &\sim \text{log-N}(\mu_c, \tau^2), c = 1, ..., K_C.
\end{aligned} \tag{2}
$$

We consider two commonly used association structures to link these longitudinal trajectory features with the binary outcome of interest: first, an MSRE structure, where the random effects $\boldsymbol{b}_i$, σ_i^2 and their interactions $\boldsymbol{b}_i \times \sigma_i^2$ are included as linear predictors in the primary outcome model and second, an LC structure, where the main and interaction effects of latent classes $\boldsymbol{D}_i$ and $\boldsymbol{C}_i$ are included in the primary outcome model. In both cases, the primary outcome model can be written as

$$
\Phi^{-1}(\text{P}(o_i = 1)) = \boldsymbol{Q}_i' \boldsymbol{\eta}, \tag{3}
$$

where o_i denotes the binary indicator of outcome, $\boldsymbol{Q}_i$ denotes a vector of the covariates in the probit model for subject i, $i = 1, \cdots, n$ and $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution. For the LC model, $\boldsymbol{Q}_i$ contains indicators for the latent classes $\boldsymbol{D}_i$ and $\boldsymbol{C}_i$; for the MSRE model, $\boldsymbol{Q}_i$ contains the subject-specific random effects $\boldsymbol{b}_i$ and residual variance σ_i^2; other fully observed baseline covariates of interest can be included in $\boldsymbol{Q}_i$ in either model as well.

2.1 Likelihood specification

Let $\boldsymbol{\phi} = (\{\boldsymbol{\beta}_d\}_{d=1}^{K_D}, \{\boldsymbol{\Sigma}_d\}_{d=1}^{K_D}, \{\pi_d^D\}_{d=1}^{K_D}, \{\mu_c\}_{c=1}^{K_C}, \{\pi_c^C\}_{c=1}^{K_C}, \tau^2, \boldsymbol{\eta})'$. All unobserved latent variables are denoted by $\boldsymbol{Z}$, $\boldsymbol{Z} = (\boldsymbol{b}, \boldsymbol{\sigma}, \boldsymbol{C}, \boldsymbol{D})'$. The observed data $\boldsymbol{x}$ consists of the longitudinal profiles $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ and the observed outomes $o_1, ..., o_n$. Then the complete data likelihood of $\boldsymbol{\phi}$ based on the complete data $(\boldsymbol{x}, \boldsymbol{Z})$ is given by

$$
\begin{aligned}
& f(\boldsymbol{x}, \boldsymbol{Z} | \boldsymbol{\phi}) \\
&= \prod_{i=1}^{n} \left[\prod_{d=1}^{K_D} \left[\frac{\pi_d^D \exp\left\{ -\frac{1}{2} (\boldsymbol{b}_i - \boldsymbol{\beta}_d)' \boldsymbol{\Sigma}_d^{-1} (\boldsymbol{b}_i - \boldsymbol{\beta}_d) \right\}}{\sqrt{(2\pi)^r |\boldsymbol{\Sigma}_d|}} \right]^{\text{I}(D_{id}=1)} \right. \\
&\quad \times \prod_{c=1}^{K_C} \left[\frac{\pi_c^C}{\sqrt{2\pi\tau^2}\sigma_i^2} \exp\left\{ -\frac{1}{2\tau^2} \left(\log\sigma_i^2 - \mu_c \right)^2 \right\} \right]^{\text{I}(C_{ic}=1)} \\
&\quad \times \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2\sigma_i^2} \left\{ y_{ij} - \mu(\boldsymbol{b}_i; t_{ij}) \right\}^2 \right] \\
&\quad \left. \times \Phi(\boldsymbol{Q}_i' \boldsymbol{\eta})^{o_i} \left\{ 1 - \Phi(\boldsymbol{Q}_i' \boldsymbol{\eta}) \right\}^{1-o_i} \right],
\end{aligned} \tag{4}
$$

2.2 Prior specification and posterior computation

We consider the same prior distributions as considered in Jiang et al. (2015), where certain empirical data-driven priors were considered for some parameters to avoid either improper posterior or existence of empty classes during the iterations of MCMC sampling. We found that our considered model assessment criteria were not sensitive to these choices.

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 linear regression estimator of $\boldsymbol{y}$ on the design matrix of $\boldsymbol{t}$ defined by $f(\cdot; t_{ij})$. This corresponds to a prior with data-driven inflated covariance structure centered at a null model, and avoids improper posteriors resulting from the possibility that some of the latent classes are not represented in the data (Elliott et al., 2005). For the variance-covariance matrix for the random effects $\boldsymbol{\Sigma}_d$, we use the prior proposed by Kass and Natarajan (2006): $\boldsymbol{\Sigma}_d \sim$ Inverse-Wishart(df $= r, \boldsymbol{\Lambda}$), where $\boldsymbol{\Lambda} = m(\sum_{i=1}^{n} \widehat{\text{Cov}}(\tilde{\boldsymbol{b}}_i)^{-1}/n)^{-1}$, $\tilde{\boldsymbol{b}}_i$ is given by OLS estimator of $\boldsymbol{b}_i$ for subject i and r is the dimension of $\boldsymbol{b}_i$. We let $m = 2.5 + (r-1)/2$ as suggested by Frühwirth-Schnatter (2006, Sec. 6.3.2) to restrain the eigenvalues of $\boldsymbol{\Sigma}_d$ away from 0, avoiding the improper posterior that can result from unbounded likelihoods when the variance-covariance matrix is unrestricted in normal mixture models (Day, 1969).

We used diffuse priors: $\mu_c \sim \text{N}(0, v)$, $\tau^{-2} \sim \text{Gamma}(a, b)$ with $v = 1{,}000$ and $a = b = .001$ for the mixture of log normal distributions for the residual variances. For the class membership probabilities, we assume conjugate Dirichlet$(4, ..., 4)$ on both $\boldsymbol{\pi}^C = (\pi_1^C, ..., \pi_{K_C}^C)$ and $\boldsymbol{\pi}^D = (\pi_1^D, ..., \pi_{K_D}^D)$ (Frühwirth-Schnatter, 2006), which is equivalent to assuming *a priori* 4 or more observations in each class, avoiding the existence of empty classes. Lastly, we used $\boldsymbol{\eta} \sim \text{N}(\boldsymbol{0}, (9/4)\boldsymbol{I})$ as the prior for $\boldsymbol{\eta}$ in the probit regression, where $(9/4)\boldsymbol{I}$ is chosen to bound the estimated outcome probabilities to be away from 0 and 1 (Garrett and Zeger, 2000; Elliott et al., 2007, and Neelon et al., 2011).

Gibbs sampling is utilized to obtain the draws from the posterior distributions. The detailed MCMC sampling procedures are provided in the Appendix A.1.

3. MODEL SELECTION AND ASSESSMENT CRITERIA

3.1 Model selection criteria

We consider several commonly used model selection criteria that are both computationally feasible and stable for our proposed joint models. For a comprehensive review of Bayesian model selection criteria, please refer to Vehtari and Ojanen (2012).

3.1.1 Log-pseudo marginal likelihood criterion

The log-pseudo marginal likelihood (LPML) (Geisser and Eddy, 1979) corresponds to a Bayesian cross-validation mea-

sure, defined as LPML $= \sum_{i=1}^{n} \log(\text{CPO}_i)$, where $\text{CPO}_i = f(\boldsymbol{y}_i, o_i | \boldsymbol{y}_{(-i)}, \boldsymbol{o}_{(-i)})$ represents a cross-validated posterior predictive density for $\boldsymbol{x}_i = (\boldsymbol{y}_i, o_i)$ given the data excluding $(\boldsymbol{y}_i, o_i)$ (denoted by $(\boldsymbol{y}_{(-i)}, \boldsymbol{o}_{(-i)}) = \boldsymbol{x}_{(-i)}$). The model with higher value of LPML provides better fit to the data (Ibrahim et al., 2001). Details of the LPML computation are provided in the Appendix A.2.

3.1.2 Deviance information criterion

DIC (Spiegelhalter et al., 2002) is a Bayesian analog of the original AIC (Akaike, 1974), but DIC uses the discrepancy between the posterior mean of the deviance $\overline{D(\boldsymbol{\phi})} = E_{\boldsymbol{\phi}}\{-2\log f(\boldsymbol{x} \mid \boldsymbol{\phi}) \mid \boldsymbol{x}\}$ and the deviance evaluated at the posterior mean $D(\overline{\boldsymbol{\phi}}) = -2\log f\{\boldsymbol{x} \mid E(\boldsymbol{\phi} \mid \boldsymbol{x})\}$ to estimate the effective number of degrees of parameters in the model p_D:

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

In our setting, the observed data likelihood $f(\boldsymbol{x} \mid \boldsymbol{\phi})$ is not available in closed form, where $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{o})'$; instead we use the approach outlined in Celeux et al. (2006) to obtain $\text{E}_{\boldsymbol{Z}}\{\text{DIC}(\boldsymbol{x}, \boldsymbol{Z})\} = -4\text{E}_{\boldsymbol{Z},\boldsymbol{\phi}}\{\log f(\boldsymbol{x}, \boldsymbol{Z} \mid \boldsymbol{\phi}) \mid \boldsymbol{x}\} + 2\text{E}_{\boldsymbol{Z}}[\log f\{\boldsymbol{x}, \boldsymbol{Z} \mid \text{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \boldsymbol{x}, \boldsymbol{Z})\} \mid \boldsymbol{x}]$, where the complete data likelihood $f(\boldsymbol{x}, \boldsymbol{Z} \mid \boldsymbol{\phi})$ has a closed form as specified in (4), and $\text{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \boldsymbol{x}, \boldsymbol{Z})$ is obtained via numerical methods. The detail is provided in the Appendix A.3.

3.1.3 Modified AIC

Although the original AIC proposed by Akaike (1974) is developed for "regular" models and hence is not directly defined for Bayesian hierarchical model, we consider AIC modified based on the complete data likelihood using Bayesian principle. Specifically, the modified AIC is defined using the deviance based on the complete data likelihood with a penalty term to account for the number of model parameters as follows:

$$\begin{aligned}\text{AIC}_1 &= \overline{D(\boldsymbol{\phi})} + 2p \\ &= -2\text{E}_{\boldsymbol{Z},\boldsymbol{\phi}}\left\{\log f(\boldsymbol{x}, \boldsymbol{Z} \mid \boldsymbol{\phi}) \mid \boldsymbol{x}\right\} + 2p, \\ \text{AIC}_2 &= D(\overline{\boldsymbol{\phi}}) + 2p \\ &= -2\text{E}_{\boldsymbol{Z}}\left[\log f\left\{\boldsymbol{x}, \boldsymbol{Z} \mid \text{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \boldsymbol{x}, \boldsymbol{Z})\right\} \mid \boldsymbol{x}\right] + 2p,\end{aligned}$$

where for joint LC model, $p = [K_C + (r+1)(1+r/2)]K_D + 2K_C - 1$ and for joint MSRE model, $p = (r+1)[2 + (1 + r/2)K_D] + 2K_C - 1$, where r is the dimension of random effect $\boldsymbol{b}_i$. For both models, there are $K_D - 1$ parameters for π_d, rK_D for $\boldsymbol{\mu}_d$, $r(r+1)K_D/2$ for $\boldsymbol{\Sigma}_d$ in the mean profile; there are $K_C - 1$ parameters for π_c, K_C for μ_c, 1 for τ^2 in the variance profile. For the LC structure, since we consider a saturated model with all possible main and interaction effects between the mean and the variance profiles, there are $K_C K_D$ parameters in the primary outcome model; for the MSRE structure, since we consider all possible main and interaction effects between random effects $\boldsymbol{b}_i$ and variances σ_i^2, there are $2(r+1)$ parameters in the primary outcome model.

3.1.4 Modified BIC and ICL

Accordingly, we consider the following modified BIC's that correspond to the above definition of AIC's:

$$\begin{aligned}\text{BIC}_1 &= \overline{D(\boldsymbol{\phi})} + p\log(n) \\ &= -2\text{E}_{\boldsymbol{Z},\boldsymbol{\phi}}\left\{\log f(\boldsymbol{x}, \boldsymbol{Z} \mid \boldsymbol{\phi}) \mid \boldsymbol{x}\right\} + p\log(n), \\ \text{BIC}_2 &= D(\overline{\boldsymbol{\phi}}) + p\log(n) \\ &= -2\text{E}_{\boldsymbol{Z}}\left[\log f\left\{\boldsymbol{x}, \boldsymbol{Z} \mid \text{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \boldsymbol{x}, \boldsymbol{Z})\right\} \mid \boldsymbol{x}\right] + p\log(n).\end{aligned}$$

To identify the correct number of components for finite mixture distributions, Biernacki et al. (1998) also suggested an integrated classification likelihood criterion (ICL), which was shown by McLachlan and Peel (2006, page 216) to be approximately equal to BIC plus two times the entropy of classification probability into assumed number of clusters. Here, we adopt this approximated version of ICL. Further, given that we have two mixture distributions for the random effects and the residual variances, respectively, we have the following two forms of ICL,

$$\begin{aligned}\text{ICL}_1 &= \text{BIC}_1 + 2\text{E}_{\boldsymbol{Z},\boldsymbol{\phi}}\left\{\text{EN}(\boldsymbol{\phi}, \boldsymbol{Z}) \mid \boldsymbol{x}\right\}, \\ \text{ICL}_2 &= \text{BIC}_2 + 2\text{E}_{\boldsymbol{Z}}\left[\text{EN}\left\{\text{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \boldsymbol{x}, \boldsymbol{Z}), \boldsymbol{Z}\right\} \mid \boldsymbol{x}\right],\end{aligned}$$

with

$$\begin{aligned}&\text{EN}(\boldsymbol{\phi}, \boldsymbol{Z}) \\ &\quad = -\sum_{d=1}^{K_D}\sum_{i=1}^{n} \text{P}(D_{id} = 1 \mid \boldsymbol{\phi}, \boldsymbol{Z}_i)\log \text{P}(D_{id} = 1 \mid \boldsymbol{\phi}, \boldsymbol{Z}_i) \\ &\qquad -\sum_{c=1}^{K_C}\sum_{i=1}^{n} \text{P}(C_{ic} = 1 \mid \boldsymbol{\phi}, \boldsymbol{Z}_i)\log \text{P}(C_{ic} = 1 \mid \boldsymbol{\phi}, \boldsymbol{Z}_i),\end{aligned}$$

where, let $\tilde{\pi}_d^D = \text{P}(D_{id} = 1 \mid \boldsymbol{\phi}, \boldsymbol{Z}_i)$ and $\tilde{\pi}_c^C = \text{P}(C_{ic} = 1 \mid \boldsymbol{\phi}, \boldsymbol{Z}_i)$ with their expressions for LC and MSRE models respectively are given in equations (11) and (12) in the Appendix A.1.

3.1.5 WAIC

Following Gelman et al. (2013), we consider the following two forms of WAIC, defined based on the conditional data likelihood $f(\boldsymbol{x}_i \mid \boldsymbol{Z}_i, \boldsymbol{\phi})$:

$$\begin{aligned}\text{WAIC}_k = &-2\sum_{i=1}^{n}\log\left[\text{E}_{\boldsymbol{Z},\boldsymbol{\phi}}\left\{f(\boldsymbol{x}_i \mid \boldsymbol{Z}_i, \boldsymbol{\phi}) \mid \boldsymbol{x}\right\}\right] \\ &+ 2p_{\text{WAIC}_k},\ k = 1, 2,\end{aligned}$$

with

$$p_{\text{WAIC}_1} = 2\sum_{i=1}^{n}\left[\log\left[\text{E}_{\boldsymbol{Z},\boldsymbol{\phi}}\left\{f(\boldsymbol{x}_i \mid \boldsymbol{Z}_i, \boldsymbol{\phi}) \mid \boldsymbol{x}\right\}\right]\right.$$

$$- \mathrm{E}_{\boldsymbol{Z},\boldsymbol{\phi}} \left[\log f(\boldsymbol{x}_i \mid \boldsymbol{Z}_i, \boldsymbol{\phi}) \mid \boldsymbol{x} \right]],$$

$$p_{\mathrm{WAIC}_2} = \sum_{i=1}^{n} \mathrm{Var}_{\boldsymbol{Z},\boldsymbol{\phi}} \left[\log f(\boldsymbol{x}_i \mid \boldsymbol{Z}_i, \boldsymbol{\phi}) \mid \boldsymbol{x} \right],$$

where

$$f(\boldsymbol{x}_i \mid \boldsymbol{Z}_i, \boldsymbol{\phi}) = \left[\Phi(\boldsymbol{Q}_i^T \boldsymbol{\eta})^{o_i} \{1 - \Phi(\boldsymbol{Q}_i^T \boldsymbol{\eta})\}^{1-o_i} \right] \times \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2\sigma_i^2} \{y_{ij} - \mu(\boldsymbol{b}_i; t_{ij})\}^2 \right].$$

3.2 Overall model performance measure

For each model under consideration, we also summarize predictive performance to link with different model selection criteria. There are many choices to quantify the performance of a predictive model for binary prediction (Taylor et al., 2008). Here, we consider the widely-used area under the curve (AUC) based on the receiver operating characteristic (ROC) curve to assess the overall model discrimination ability averaged across all predictive cutoffs. For out-of-sample prediction validation in Section 5, we consider the Brier-score based posterior predictive mean squared error as an additional performance measurement; details are provided therein.

Briefly, the ROC curve plots sensitivity versus 1-sensitivity for all possible cutoffs based on predicted $\mathrm{P}(o_i = 1) = \Phi(\boldsymbol{Q}_i' \boldsymbol{\eta})$ obtained from (3). The ROC curve and AUC were computed at each MCMC iteration using the ROCR package in R (Sing et al. 2005). The reported AUC is calculated as the posterior mean AUC averaged across all MCMC iterations.

4. SIMULATION STUDY

In this section, we conduct several simulation studies to evaluate the performance of the proposed model selection criteria under both the LC or MSRE data generating schemes and apply both models LC and MSRE to all the generated data. In what follows, we refer to the model where the observations are generated from as the "true model". Two representative data-generating structures, $K_D = K_C = 2$ and $K_D = K_C = 1$, are considered. The former, with different combinations of relative mixture locations, represents a simple but informative mixture structure; while the latter represents the null model. We report the number of components selected by each criterion under each scenario. We also report the within and out-of samples predictive performance.

4.1 Simulation setup

We specify a combination of sub-structures for our simulation studies below.

4.1.1 $K_D = K_C = 2$

For the longitudinal observations, we generate data from the following models with two components within both the mean and the variance profiles,

$$\begin{aligned} \mathrm{y}_{ij} | \boldsymbol{b}_i, \sigma_i^2 &\sim \mathrm{N}(b_{i0} + b_{i1} t_{ij}, \sigma_i^2), \\ \boldsymbol{b}_i &\sim \pi \mathrm{N}(\boldsymbol{\beta}_1, \boldsymbol{\Sigma}_1) + (1-\pi)\mathrm{N}(\boldsymbol{\beta}_2, \boldsymbol{\Sigma}_2), \\ \log \sigma_i^2 &\sim \pi \mathrm{N}(\mu_1, \tau^2) + (1-\pi)\mathrm{N}(\mu_2, \tau^2), \end{aligned} \tag{5}$$

where $i = 1, ..., 200$ and $t_{ij} = 0, 1, 2, ..., n_i$ with $n_i \equiv 20$. For $k = 1, 2$, we denote $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2})'$ and $\boldsymbol{\Sigma}_k = \begin{pmatrix} \omega^2 & \rho_k \omega^2 \\ \rho_k \omega^2 & \omega^2 \end{pmatrix}$. For all model scenarios, 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 $\boldsymbol{\mu}_2 = -.5$. Thus the mean of the two bivariate normals differs by an Euclidean distance of 4 throughout, while the mean log of the variances are separated by 1.5. In our investigation, we consider the case of "overlapped" versus "separated" mixture components, crossed with "balanced" 50:50 versus "unbalanced" 20:80 mixing proportions for both the mean and the variance profiles. Besides the separation in mixture components, we anticipate the mixing proportion of 50:50 to yield more difficult to separate latent classes. Figure 1 shows the corresponding 95% contours and density plots of the two "overlapped" versus "separated" components for the mean and the variance profile, respectively. Finally, our eight longitudinal model scenarios are defined by $(\rho_2, \omega^2, \tau^2, \pi) = (.6, 2, .4, .5)$, $(-.6, 1, .4, .5)$, $(.6, 2, .06, .5)$, $(-.6, 1, .06, .5)$, $(.6, 2, .4, .2)$, $(-.6, 1, .4, .2)$, $(.6, 2, .06, .2)$, and $(-.6, 1, .06, 0.2)$, respectively.

For each scenario, we simulate 100 data sets and report the models (i.e., K_D and K_C) selected by each selection criteria. We apply each assumed model to each data scenario without regard to the true mechanism generating the data, with the assumed numbers of components being $K_D = 1, 2, 3$ and $K_C = 1, 2, 3$.

For each of the simulation scenarios proposed for the longitudinal observations, the following two underlying probit models are considered for the health outcome (we replace $\boldsymbol{\eta}$ in the general models (3) by $\boldsymbol{\theta}$ for the LC and by $\boldsymbol{\gamma}$ for the MSRE probit primary models to simplify the task of presentation):

LC probit submodel:

$$\begin{aligned} \Phi^{-1}\{\mathrm{P}(o_i = 1)\} &= \theta_0 + \theta_1 \mathrm{I}(D_{i2} = 1) + \theta_2 \mathrm{I}(C_{i2} = 1) \\ &\quad + \theta_3 \mathrm{I}(D_{i2} = 1, C_{i2} = 1), \end{aligned} \tag{6}$$

MSRE probit submodel:

$$\begin{aligned} \Phi^{-1}\{\mathrm{P}(o_i = 1)\} &= \gamma_0 + \gamma_1 b_{i0} + \gamma_2 b_{i1} + \gamma_3 \sigma_i^2 \\ &\quad + \gamma_4 b_{i0} \sigma_i^2 + \gamma_5 b_{i1} \sigma_i^2, \end{aligned} \tag{7}$$

where $D_{i1} = 1$ corresponds to the mean component $\mathrm{N}(\boldsymbol{(0, 0)'}, \boldsymbol{\Sigma}_1)$, and $C_{i1} = 1$ corresponds to the variance component $\mathrm{N}(-2, \tau^2)$ in the longitudinal submodel (5). We let $\boldsymbol{\theta} = (-0.8, 1.8, -.2, -.3)$ and $\boldsymbol{\gamma} = (-1, 1, -1, 2, -2, 2)'$ for each model scenario so that the outcome prevalence is approximately 50 percent.

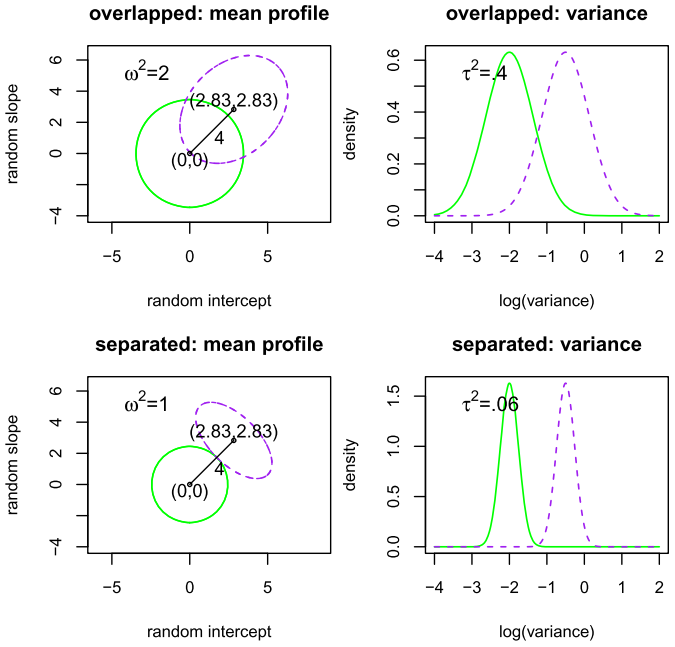

Figure 1. Simulation setup for the mean and variance profiles representing low versus high levels of separation. Left panels: 95% contour plots of the two components for the mean profiles; right panels: density plots of the two components for the variance profiles.

4.1.2 $K_D = K_C = 1$

Further, we consider the null case where there are no mixture/latent class for either the mean or the variance profiles by retaining only the first of the mixture components in (8). The primary probit models for the health outcomes are:

1. **latent class (LC) probit submodel:**

$$\Phi^{-1}\{\mathrm{P}(o_i = 1)\} = \theta_0, \tag{8}$$

2. **multiple shared random effect (MSRE) probit submodel:**

$$\Phi^{-1}\{\mathrm{P}(o_i = 1)\} = \gamma_0 + \gamma_1 b_{i0} + \gamma_2 b_{i1} + \gamma_3 \sigma_i^2 + \gamma_4 b_{i0}\sigma_i^2 + \gamma_5 b_{i1}\sigma_i^2, \tag{9}$$

For each scenario, we simulate 100 data sets and report the models (i.e., K_D and K_C) selected by each selection criteria under various scenarios, equivalently to those in Section 4.1.1, but only consider the fitted models with $K_D = 1, 2$ and $K_C = 1, 2$.

4.2 Simulation results

4.2.1 $K_D = K_C = 2$

Among the 100 simulated data sets, in Tables 1 to 2, we report the number of times each model, indicated by particular numbers of mixture components (K_D, K_C), is selected by one of the criteria given in the first column. Table 1 shows the results when the true data-generating model has LC structure while Table 2 reports those under the MSRE structure. In general, separation in mixture components plays an important role in the performance of these criteria in identifying the correct number of components: when there is a sufficiently large degree of separation in either the mean or the variance profiles, it is generally easier to choose the correct number of components regardless of which criteria is utilized.

Scenarios (a)–(d) represent different levels of separation of mean (or variance) components, as indicated in the Tables. Mixing proportion might also play a role in selecting correct numbers of mixture components. We use an unbalanced 20% vs. 80% mixing design to create some asymmetry in the mixture density. In our study, all criteria seem to perform slightly better in the cases of the unbalanced design (results not shown). Incorrectly assuming the outcome structure has some impacts on the performance of these model criteria criteria, and the degree of impact depends on the criteria and hence the goal in model selection, as well as the true association structure in the outcome model. In particular, the outcome-informed artifacts due to fitting an LC structure to the data generated under a true MSRE model, reported in Jiang et al. (2015), also have some connection with the performance of these criteria, which will be addressed in this section.

Overall, under the correctly assumed association structure in the primary outcome model, the modified AIC, BIC and ICL perform very well in selecting the correct numbers of components, even when the mixture components are harder to separate. When the true structure in the primary outcome model is LC, the modified AIC, BIC and ICL still perform equally well regardless of fitting LC or MSRE models. However, when the true structure is MSRE, fitting the LC structure can affect the performance of the modified AIC, BIC and ICL. This phenomenon is most prominent for scenario (a) in Table 2. The reported results corresponding to the two different versions of AIC, BIC, ICL or WAIC differ sometimes, but not to a noteworthy level, and therefore we do not differentiate the summary according to the versions used.

In contrast, DIC, LPML and WAIC tend to choose too many components for both the mean and the variance in all scenarios. In particular, the numbers of mixture components selected by WAIC and LPML tend to agree regardless of the fitted model structure used in the primary outcome model. More interestingly to note is that, when fitting with a joint LC model, both WAIC and LPML tend to select the numbers of components for both the mean and the variance that lead to a higher AUC value. When fitting with a joint MSRE model, WAIC and LPML still tend to select models with more mixture components for both the mean and the variance, but the model based on such a selection does not have a higher value of the AUC. In fact, the AUC values vary little under different fitted models.

Table 1. Number of times each specified fitted model is selected by using the 10 criteria given in the first column of each sub-table. Observations are generated as described in Section 4.1 under the joint LC model with $K_D = K_C = 2$ and the (T_D, T_C) mixture structure for the mean and the variance profiles, respectively, where $T_D, T_C \in \{$separated, overlapped$\}$. Scenarios (a)–(d) specify the data-generating mechanism. The fitted models consist of both LC and MSRE structures with $K_D = 1, 2, 3$ and $K_C = 1, 2, 3$. The mixing proportions are 50-50. The corresponding values of AUC were reported at the end of the table for each scenario. The "true AUC" is the AUC obtained when predictions are generated using the correct outcome model (true parameters and random effects/latent classes)

	fitting LC model (K_D, K_C)									**fitting MSRE model** (K_D, K_C)								
	(1,1)	(1,2)	(1,3)	(2,1)	**(2,2)**	(2,3)	(3,1)	(3,2)	(3,3)	(1,1)	(1,2)	(1,3)	(2,1)	**(2,2)**	(2,3)	(3,1)	(3,2)	(3,3)
(a) true model = LC; (separated, separated) mixtures; 50:50 proportion																		
No. of times selected from 100 simulations																		
DIC	0	0	0	0	26	0	0	73	1	0	0	0	0	36	1	0	62	1
LPML	0	0	0	0	8	47	0	5	40	0	11	11	0	15	33	0	7	23
AIC_1	0	0	0	0	90	1	0	9	0	0	0	0	0	90	1	0	8	1
AIC_2	0	0	0	0	88	1	0	11	0	0	0	0	0	85	1	0	13	1
BIC_1	0	0	0	0	99	1	0	0	0	0	0	0	0	98	1	0	1	0
BIC_2	0	0	0	0	99	1	0	0	0	0	0	0	0	97	1	0	2	0
ICL_1	0	0	0	1	98	1	0	0	0	0	0	0	0	98	1	0	1	0
ICL_2	0	0	0	1	98	1	0	0	0	0	0	0	0	98	1	0	1	0
$WAIC_1$	0	0	0	0	0	49	0	0	51	0	14	51	0	4	26	0	0	5
$WAIC_2$	0	0	0	0	1	68	0	1	30	0	31	24	0	16	22	0	2	5
Area under the ROC curves																		
true AUC=0.81	0.5	0.54	0.66	0.79	0.81	0.85	0.79	0.82	0.86	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
(b) true model = LC; (overlapped, overlapped) mixtures; 50:50 proportion																		
No. of times selected from 100 simulations																		
DIC	56	0	0	2	0	0	42	0	0	27	0	0	17	0	0	56	0	0
LPML	0	0	1	3	13	38	2	15	28	9	14	22	10	13	12	5	8	7
AIC_1	76	0	0	19	0	0	5	0	0	80	0	0	17	0	0	3	0	0
AIC_2	70	0	0	21	0	0	9	0	0	72	0	0	23	0	0	5	0	0
BIC_1	94	0	0	6	0	0	0	0	0	100	0	0	0	0	0	0	0	0
BIC_2	92	0	0	8	0	0	0	0	0	99	0	0	1	0	0	0	0	0
ICL_1	100	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
ICL_2	100	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
$WAIC_1$	0	0	0	0	2	60	1	0	37	32	19	27	9	8	1	0	1	3
$WAIC_2$	0	0	2	0	2	65	1	0	30	16	25	36	4	8	5	1	0	5
Area under the ROC curves																		
true AUC=0.81	0.5	0.63	0.72	0.78	0.85	0.88	0.78	0.85	0.88	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77
(c) true model = LC; (overlapped, separated) mixtures; 50:50 proportion																		
No. of times selected from 100 simulations																		
DIC	0	57	0	0	3	2	0	37	1	0	41	1	0	12	1	0	45	0
LPML	0	0	0	0	13	43	0	11	33	0	9	28	0	16	19	0	13	15
AIC_1	0	80	0	0	17	1	0	2	0	0	81	1	0	14	1	0	3	0
AIC_2	0	72	0	0	24	1	0	3	0	0	76	1	0	19	1	0	3	0
BIC_1	0	97	1	0	2	0	0	0	0	0	98	2	0	0	0	0	0	0
BIC_2	0	94	0	0	5	1	0	0	0	0	98	2	0	0	0	0	0	0
ICL_1	0	99	1	0	0	0	0	0	0	0	99	1	0	0	0	0	0	0
ICL_2	0	99	1	0	0	0	0	0	0	0	99	1	0	0	0	0	0	0
$WAIC_1$	0	0	1	0	2	66	0	0	31	0	22	45	0	0	20	0	3	10
$WAIC_2$	0	0	0	0	2	65	0	1	32	0	39	19	0	16	7	0	14	5
Area under the ROC curves																		
true AUC=0.82	0.5	0.55	0.66	0.80	0.83	0.87	0.80	0.83	0.87	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78
(d) true model = LC; (separated, overlapped) mixtures; 50:50 proportion																		
No. of times selected from 100 simulations																		
DIC	0	0	0	11	0	0	88	1	0	1	0	0	16	0	0	82	1	0
LPML	0	0	0	0	17	37	2	11	33	6	8	2	13	14	16	8	16	17
AIC_1	0	0	0	85	0	0	15	0	0	1	0	0	77	1	0	21	0	0
AIC_2	0	0	0	73	1	0	26	0	0	1	0	0	71	0	0	27	1	0
BIC_1	0	0	0	100	0	0	0	0	0	1	0	0	98	0	0	1	0	0
BIC_2	0	0	0	100	0	0	0	0	0	1	0	0	98	0	0	1	0	0
ICL_1	0	0	0	100	0	0	0	0	0	2	0	0	98	0	0	0	0	0
ICL_2	0	0	0	100	0	0	0	0	0	1	0	0	98	0	0	1	0	0
$WAIC_1$	0	0	0	0	0	59	0	0	41	31	16	28	6	7	9	2	1	0
$WAIC_2$	0	0	0	0	0	74	0	0	26	13	17	29	6	10	17	1	2	5
Area under the ROC curves																		
true AUC=0.82	0.5	0.65	0.73	0.79	0.85	0.88	0.79	0.85	0.88	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81

Table 2. As in Table 1 but the observations are generated under the joint MSRE model

	fitting LC model (K_D, K_C)									**fitting MSRE model** (K_D, K_C)								
	(1,1)	(1,2)	(1,3)	(2,1)	**(2,2)**	(2,3)	(3,1)	(3,2)	(3,3)	(1,1)	(1,2)	(1,3)	(2,1)	**(2,2)**	(2,3)	(3,1)	(3,2)	(3,3)
(a) true model = MSRE; (separated, separated) mixtures; 50:50 proportion																		
No. of times selected from 100 simulations																		
DIC	0	1	0	0	25	4	0	65	5	0	0	0	0	37	1	0	60	2
LPML	0	3	14	0	8	47	0	3	25	0	6	7	0	21	32	0	11	23
AIC_1	0	1	0	0	57	5	0	34	3	0	0	0	0	88	3	0	9	0
AIC_2	0	1	0	0	55	5	0	36	3	0	0	0	0	81	3	0	16	0
BIC_1	0	3	0	0	73	4	0	19	1	0	0	0	0	94	3	0	3	0
BIC_2	0	2	0	0	72	4	0	20	2	0	0	0	0	94	3	0	3	0
ICL_1	0	5	0	0	74	2	0	19	0	0	1	0	0	97	1	0	1	0
ICL_2	0	5	0	0	71	3	0	21	0	0	1	0	0	96	1	0	2	0
$WAIC_1$	0	0	18	0	7	45	0	1	29	0	24	45	0	8	18	0	1	4
$WAIC_2$	0	0	31	0	8	48	0	1	12	0	37	21	0	19	13	0	4	6
Area under the ROC curves																		
true AUC=0.82	0.5	0.64	0.72	0.52	0.70	0.80	0.54	0.70	0.77	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
(b) true model = MSRE; (overlapped, overlapped) mixtures; 50:50 proportion																		
No. of times selected from 100 simulations																		
DIC	15	0	0	11	3	0	56	13	2	23	0	0	17	0	0	60	0	0
LPML	0	1	2	0	32	30	0	20	15	13	24	17	8	7	11	5	6	9
AIC_1	74	0	0	15	8	1	2	0	0	84	0	0	13	0	0	3	0	0
AIC_2	63	0	0	18	13	1	5	0	0	73	0	0	21	0	0	6	0	0
BIC_1	98	0	0	0	2	0	0	0	0	100	0	0	0	0	0	0	0	0
BIC_2	95	0	0	1	4	0	0	0	0	99	0	0	1	0	0	0	0	0
ICL_1	99	0	0	0	1	0	0	0	0	100	0	0	0	0	0	0	0	0
ICL_2	99	0	0	0	1	0	0	0	0	100	0	0	0	0	0	0	0	0
$WAIC_1$	0	1	3	0	50	18	0	20	8	23	32	26	3	3	8	3	0	2
$WAIC_2$	0	1	4	0	39	27	0	17	12	7	30	38	1	5	14	1	1	3
Area under the ROC curves																		
true AUC=0.85	0.5	0.74	0.79	0.52	0.91	0.92	0.53	0.92	0.93	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
(c) true model = MSRE; (overlapped, separated) mixtures; 50:50 proportion																		
No. of times selected from 100 simulations																		
DIC	0	65	3	0	3	2	0	18	9	0	38	2	0	16	2	0	42	0
LPML	0	0	3	0	10	44	0	8	35	0	19	31	0	7	21	0	8	14
AIC_1	0	83	3	0	7	6	0	0	1	0	81	4	0	10	0	0	5	0
AIC_2	0	77	3	0	10	4	0	1	5	0	79	4	0	12	0	0	5	0
BIC_1	0	95	2	0	0	3	0	0	0	0	97	3	0	0	0	0	0	0
BIC_2	0	94	2	0	0	4	0	0	0	0	96	4	0	0	0	0	0	0
ICL_1	0	98	1	0	0	1	0	0	0	0	98	2	0	0	0	0	0	0
ICL_2	0	98	1	0	0	1	0	0	0	0	97	3	0	0	0	0	0	0
$WAIC_1$	0	0	0	0	0	54	0	3	43	0	26	54	0	3	9	0	2	6
$WAIC_2$	0	0	0	0	2	55	0	4	39	0	42	27	0	9	10	0	6	6
Area under the ROC curves																		
true AUC=0.82	0.5	0.63	0.71	0.52	0.83	0.88	0.53	0.84	0.89	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
(d) true model = MSRE; (separated,overlapped) mixtures; 50:50 proportion																		
No. of times selected from 100 simulations																		
DIC	0	0	0	16	0	0	83	1	0	0	0	0	19	1	0	80	0	0
LPML	0	2	8	0	30	32	0	9	19	5	8	10	12	18	21	7	7	12
AIC_1	0	0	0	79	0	0	21	0	0	0	0	0	78	1	0	21	0	0
AIC_2	0	0	0	71	0	0	29	0	0	0	0	0	68	1	0	31	0	0
BIC_1	2	0	0	94	1	0	3	0	0	1	0	0	91	1	0	7	0	0
BIC_2	1	0	0	94	1	0	4	0	0	1	0	0	91	1	0	7	0	0
ICL_1	2	0	0	95	0	0	3	0	0	2	0	0	91	0	0	7	0	0
ICL_2	2	0	0	95	0	0	3	0	0	2	0	0	91	0	0	7	0	0
$WAIC_1$	0	4	10	0	37	22	0	7	20	13	21	41	8	9	5	0	3	0
$WAIC_2$	0	2	19	0	33	23	0	6	17	7	17	36	9	12	14	0	3	2
Area under the ROC curves																		
true AUC=0.85	0.5	0.75	0.80	0.53	0.85	0.88	0.54	0.81	0.86	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85

As suggested by one referee, we also report the results from fitting the longitudinal submodel (5) alone in Table 3. In this case, the influence from the outcome model is no longer present; the performance of these criteria tend to behave similarly as for MSRE models.

As we can see from the AUC values given in Tables 1, even correctly assuming the LC structure can lead to either (i) lower or (ii) higher AUC values than the AUC values by the true models. We believe that (i) is due to the difficulty in separating the mixture components of the means that

Table 3. As in Table 1 but the observations are generated under the longitudinal submodel alone

	fitting longitudinal submodel (K_D, K_C)								
	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
	(separated, separated) mixtures; 50:50 proportion								
	No. of times selected from 100 simulations								
DIC	0	0	0	0	34	0	0	64	2
LPML	0	4	26	0	17	21	0	8	24
AIC_1	0	0	0	0	87	0	0	11	2
AIC_2	0	0	0	0	77	0	0	21	2
BIC_1	0	0	0	0	97	2	0	1	0
BIC_2	0	0	0	0	97	2	0	1	0
ICL_BIC_1	0	0	0	1	96	2	0	1	0
ICL_BIC_2	0	0	0	1	96	2	0	1	0
$WAIC_1$	0	25	51	0	3	12	0	1	8
$WAIC_2$	0	38	25	0	16	8	0	7	6
	(overlapped, overlapped) mixtures; 50:50 proportion								
	No. of times selected from 100 simulations								
DIC	23	0	0	6	0	0	71	0	0
LPML	14	16	14	15	8	10	7	10	6
AIC_1	74	0	0	19	0	0	7	0	0
AIC_2	69	0	0	19	0	0	12	0	0
BIC_1	100	0	0	0	0	0	0	0	0
BIC_2	100	0	0	0	0	0	0	0	0
ICL_BIC_1	100	0	0	0	0	0	0	0	0
ICL_BIC_2	100	0	0	0	0	0	0	0	0
$WAIC_1$	17	35	29	7	7	2	0	2	1
$WAIC_2$	11	30	40	2	7	6	0	3	1
	(overlapped, separated) mixtures; 50:50 proportion								
	No. of times selected from 100 simulations								
DIC	0	36	2	0	19	0	0	42	1
LPML	0	11	26	0	13	24	0	6	20
AIC_1	0	77	3	0	13	0	0	7	0
AIC_2	0	73	3	0	14	0	0	10	0
BIC_1	0	98	2	0	0	0	0	0	0
BIC_2	0	96	2	0	2	0	0	0	0
ICL_BIC_1	0	100	0	0	0	0	0	0	0
ICL_BIC_2	0	99	0	0	1	0	0	0	0
$WAIC_1$	0	15	57	0	1	15	0	0	12
$WAIC_2$	0	31	34	0	11	13	0	8	3
	(separated, overlapped) mixtures; 50:50 proportion								
	No. of times selected from 100 simulations								
DIC	0	0	0	31	0	0	68	1	0
LPML	6	5	7	21	14	14	14	11	8
AIC_1	0	0	0	79	1	0	20	0	0
AIC_2	0	0	0	76	1	0	23	0	0
BIC_1	0	0	0	98	1	0	1	0	0
BIC_2	0	0	0	98	1	0	1	0	0
ICL_BIC_1	0	0	0	99	0	0	1	0	0
ICL_BIC_2	0	0	0	99	0	0	1	0	0
$WAIC_1$	12	34	26	11	5	7	2	2	1
$WAIC_2$	5	26	29	7	10	15	2	3	3

leads, in turn, to a higher potential for misclassification. An extreme case is when almost all subjects are assigned to one mixture component and the prediction of the outcome is solely determined by the variance profile, which results in worse prediction ability than the true model. On the other hand, (ii) is likely due to the outcome-informed artifact, where subjects are assigned to spurious mixture components to generate predictions of the outcome that are more accurate than those obtained with the true model, as discussed in Jiang et al. (2015). Meanwhile, fitting joint MSRE models when LC is the true structure leads only to a slight loss of the predictive power relative to the true model, when the mixture components within the mean profile overlap; otherwise the AUC values obtained by the MSRE model are similar to the true model.

On the other hand, as shown in Tables 1, when MSRE is the true data-generating mechanism, the AUC values obtained by fitting the joint MSRE models are always close to the AUC values by the true models in all scenarios we consider, indicating that the predictive ability under such scenarios is not affected much by any potential misclassification due to the overlapping mixture components. However, when MSRE is the true structure, fitting LC models would lead to either an increase in predictive power, indicated by extremely high AUC values due to artificially recognized new components, or loss of predictive power with low AUC values, originated from replacing a set of continuous variables (i.e., MSRE) by a discretized version (i.e., LC) in the primary outcome model. Because of these mentioned potentials when fitting joint LC and MSRE models, all criteria suggest that it is difficult to differentiate the LC and MSRE models under the scenarios with overlapping components.

4.2.2 $K_D = K_C = 1$

As shown in Table 4, under the true LC model and for all assumed structures, the modified AIC, BIC and ICL all perform very well when distinct mixture components do not exist for both the mean and the variance. Slightly different from BIC and ICL, the AIC favors more complex structures at times. DIC, LPML and WAIC also have the tendency to select models with too many mixture components, where LPML and WAIC tend to select the number of components that lead to higher AUC values. This once again suggests that WAIC and LPML tend to select models with high predictive accuracy regardless of true association structure.

Under this null structure, we clearly see that the joint LC model has the ability to create additional outcome-informed components that lead to misleadingly high AUC values. The highest spurious AUC value is 0.97 as shown in Table 4 when the true model is MSRE, giving the impression of almost perfect prediction when there is no mixture at all in either the mean or the variance profiles. In this case, all criteria under the assumed LC model tend to choose more numbers of components than that of the data-generating scheme, reaching a better goodness-of-fit.

5. VALIDATION OF THE MODELS SELECTED BY DIFFERENT CRITERIA

It is well known that using the predictive model built on the same data set where the prediction is conducted would lead to optimistically biased prediction evaluation. In this section, we conduct evaluations of different model-selection criteria using newly generated independent samples to obtain a better assessment of their predictive performance. With Tables 1 to 4 showing that the key over-fitting phenomenon reflected by AUC is preserved in the simplest data-generation scheme of $K_D = K_C = 1$, we focus on this setup and again allow the fitted models to have 1 or 2 components. We choose this particular scenario as an extreme case of completely overlapped mixture component to amplify the effects of potential outcome-informed artifacts by fitting LC models, since the true LC model is essentially a null model with AUC = 0.5. For this simplest scenario, we observe that the joint LC models with $K_D > 1$ or $K_C > 1$ could lead to exceedingly high AUC values relative to the true AUC, and that such joint LC models are frequently favored by LPML and WAIC. On the positive side, when the observations are generated from the MSRE model but fitted with LC ones, the models selected by LPML, WAIC and DIC do result in better predictive performance on the validation samples. The obtained predictions also allow us to see whether such high predictive accuracy is real or an artifact.

For each data set $\{(\boldsymbol{y}_i^{(s)}, o_i^{(s)})\}_{i=1}^n$, $s = 1, \cdots, 100$, generated from the given true joint LC and MSRE models, we generate an additional new validation data set $\{(\tilde{\boldsymbol{y}}_i^{(s)}, \tilde{o}_i^{(s)})\}_{i=1}^{\tilde{n}}$, $\tilde{n} = 50$. We use $\mathrm{H}_a^{(s)}$ to denote the model selected by model selection criteria $a \in \{\mathrm{DIC}, \mathrm{LPML}, \mathrm{AIC}_1, \mathrm{BIC}_1, \mathrm{ICL}_1, \mathrm{WAIC}_1\}$ for the data set $\{(\boldsymbol{y}_i^{(s)}, o_i^{(s)})\}_{i=1}^n$ and then each $\mathrm{H}_a^{(s)}$ has unique estimated values of K_D and K_C.

When introducing our target prediction, we drop the superscript (s) in the notation for simplicity. We split the model parameter vector $\boldsymbol{\phi}$ into two components: 1) $\boldsymbol{\phi}_{long} = (\{\boldsymbol{\beta}_d\}_{d=1}^{K_D}, \{\boldsymbol{\Sigma}_d\}_{d=1}^{K_D}, \{\pi_d^D\}_{d=1}^{K_D}, \{\mu_c\}_{c=1}^{K_C}, \tau^2, \{\pi_c^C\}_{c=1}^{K_C})^T$, including all the population level parameters in the longitudinal submodel, and 2) $\boldsymbol{\eta}$, the vector of coefficients in the primary outcome model. We let $\tilde{\boldsymbol{Z}} = (\tilde{\boldsymbol{D}}, \tilde{\boldsymbol{C}}, \tilde{\boldsymbol{b}}, \tilde{\boldsymbol{\sigma}})^T$ include all individual level latent variables for the new validation sample and, compatibly, $\boldsymbol{Z} = (\boldsymbol{D}, \boldsymbol{C}, \boldsymbol{b}, \boldsymbol{\sigma})^T$ includes all such latent variables for the original sample set used to obtain the fitted model. The prediction of the primary outcome for new validation sample is then based on the following quantity,

$$
\begin{aligned}
(10)\quad & p\left(\tilde{\boldsymbol{o}} | \tilde{\boldsymbol{y}}, \boldsymbol{y}, \boldsymbol{o}, \mathrm{H}_a\right) \\
&= \int p(\tilde{\boldsymbol{o}} | \tilde{\boldsymbol{Z}}, \boldsymbol{\eta}, \mathrm{H}_a) \\
&\quad \times p(\tilde{\boldsymbol{Z}}, \boldsymbol{Z}, \boldsymbol{\phi}_{long}, \boldsymbol{\eta} | \tilde{\boldsymbol{y}}, \boldsymbol{y}, \boldsymbol{o}, \mathrm{H}_a) d\boldsymbol{\phi}_{long} d\boldsymbol{\eta} d\tilde{\boldsymbol{Z}} d\boldsymbol{Z} \\
&\simeq \frac{1}{M} \sum_{m=1}^{M} p\left(\tilde{\boldsymbol{o}} | \tilde{\boldsymbol{Z}}^{(m)}, \boldsymbol{\eta}^{(m)}, \mathrm{H}_a\right),
\end{aligned}
$$

Table 4. Number of times each specified fitted model is selected by using the 10 criteria given in the first column of each sub-table. Observations are generated as described in Section 4.1 under the joint LC model for scenario (a), and MSRE model for scenario (b), with $K_D = K_C = 1$. The fitted models consist of both LC and MSRE structures with $K_D = 1, 2$ and $K_C = 1, 2$. The corresponding values of AUC are reported at the end of the table for each scenario

	fitting LC model (K_D, K_C)				**fitting MSRE model** (K_D, K_C)			
	(1,1)	(1,2)	(2,1)	(2,2)	(1,1)	(1,2)	(2,1)	(2,2)
				(a) True model = LC				
				No. of times selected				
DIC	9	0	91	0	10	0	90	0
LPML	0	58	0	42	30	23	22	25
AIC_1	92	0	8	0	81	0	19	0
AIC_2	76	0	24	0	72	0	28	0
BIC_1	100	0	0	0	100	0	0	0
BIC_2	100	0	0	0	100	0	0	0
ICL_1	100	0	0	0	100	0	0	0
ICL_2	100	0	0	0	100	0	0	0
$WAIC_1$	0	46	0	54	43	41	9	7
$WAIC_2$	0	56	0	44	27	49	10	14
				Area under the ROC curves				
true AUC=0.5	0.5	0.68	0.51	0.69	0.57	0.57	0.57	0.57
				(b) True model = MSRE				
				No. of times selected				
DIC	7	0	92	1	14	0	86	0
LPML	0	0	50	50	31	28	19	22
AIC_1	15	0	84	1	84	0	16	0
AIC_2	10	0	89	1	73	0	27	0
BIC_1	42	0	58	0	100	0	0	0
BIC_2	34	0	66	0	100	0	0	0
ICL_1	69	0	31	0	100	0	0	0
ICL_2	59	0	41	0	100	0	0	0
$WAIC_1$	0	1	66	33	40	33	11	16
$WAIC_2$	0	1	58	41	35	42	6	17
				Area under the ROC curves				
true AUC=0.88	0.5	0.69	0.95	0.97	0.88	0.88	0.88	0.88

where $\tilde{\boldsymbol{Z}}^{(m)}, \boldsymbol{Z}^{(m)}, \boldsymbol{\phi}_{long}^{(m)}$ and $\boldsymbol{\eta}^{(m)}$, $m = 1, \cdots, M$ are drawn from the posterior distribution $p(\tilde{\boldsymbol{Z}}, \boldsymbol{Z}, \boldsymbol{\phi}_{long}, \boldsymbol{\eta}|\tilde{\boldsymbol{y}}, \boldsymbol{y}, \boldsymbol{o}, \mathrm{H}_a)$. Details of the MCMC sampling algorithm are given in the Appendix A.4. Further, $\tilde{p}_i^{(m)} := p(\tilde{o}_i = 1|\boldsymbol{\eta}^{(m)}, \tilde{\boldsymbol{Z}}_i^{(m)})$ can be obtained from $\Phi((\boldsymbol{\eta}^{(m)})^T \boldsymbol{Z}_i^{(m)})$ as described in (3).

Then, for each validation data set, we focus on two performance measures: 1) the posterior predictive mean squared error (PMSE): $\mathrm{PMSE} = M^{-1} \sum_{m=1}^{M} \mathrm{PMSE}^{(m)}$, where $\mathrm{PMSE}^{(m)} = \tilde{n}^{-1} \sum_{i=1}^{\tilde{n}} (\tilde{o}_i^{\mathrm{pred},(m)} - \tilde{o}_i)^2$ with $\tilde{o}_i^{\mathrm{pred},(m)}$ as a draw of a Bernoulli random variable with success probability $\tilde{p}_i^{(m)}$; 2) the area under the ROC curve $\mathrm{AUC} = M^{-1} \sum_{m=1}^{M} \mathrm{AUC}^{(m)}$ (i.e., test AUC), where $\mathrm{AUC}^{(m)}$ is obtained based on $\tilde{p}_i^{(m)}$, $i = 1, \cdots, \tilde{n}$, using the approach as described in Section 4. We calculate the posterior PMSE and test AUC for the validation sample $\{(\tilde{\boldsymbol{y}}_i^{(s)}, \tilde{o}_i^{(s)})\}_{i=1}^{\tilde{n}}$ fitted by model $\mathrm{H}_a^{(s)}$, $s = 1, \cdots, 100$, and report the posterior mean PMSE and AUC as well as the 95% credible intervals based on 100 simulations. As a comparison, we also summarize the training AUC for the sample $\{(\boldsymbol{y}_i^{(s)}, o_i^{(s)})\}_{i=1}^{\tilde{n}}$ that is reported in Table 4 for each criteria. We then repeat this procedure for the two simulation scenarios as described in the previous section: when the data are generated from the joint LC and the joint MSRE model, respectively.

Table 5 shows the PMSE and test AUC for the new validation test samples, along with the AUC for the training samples, based on the selected models by DIC, LPML, modified AIC_1, BIC_1 and ICL_1, as well as $WAIC_1$, respec-

Table 5. Values of AUC for independent validation sample ($\tilde{n} = 50$) based on the models selected by DIC, LPML, AIC, BIC, ICL and WAIC. Both the independent validation sample and the original sample to build the models are generated as described in Section 4.1 under the joint LC model, for scenario (a), and MSRE model, for scenario (b), with $K_D = K_C = 1$

	fitting joint LC			**fitting joint MSRE**		
	PMSE (95% CI)	test AUC (95% CI)	training AUC (95% CI)	PMSE (95% CI)	test AUC (95% CI)	training AUC (95% CI)
	(a) true model = LC (true test AUC=0.50)					
DIC	0.50 (0.48,0.51)	0.50 (0.49,0.52)	0.51 (0.50,0.53)	0.50 (0.47,0.52)	0.50 (0.40,0.59)	0.57 (0.52,0.63)
LPML	0.50 (0.48,0.51)	0.50 (0.49,0.52)	0.68 (0.64,0.74)	0.50 (0.47,0.52)	0.50 (0.40,0.59)	0.57 (0.52,0.62)
AIC_1	0.50 (0.48,0.51)	0.50 (0.50,0.50)	0.50 (0.50,0.53)	0.50 (0.47,0.52)	0.50 (0.40,0.59)	0.57 (0.52,0.62)
BIC_1	0.50 (0.48,0.51)	0.50 (0.50,0.50)	0.50 (0.50,0.50)	0.50 (0.47,0.52)	0.50 (0.40,0.59)	0.57 (0.52,0.63)
ICL_1	0.50 (0.48,0.51)	0.50 (0.50,0.50)	0.50 (0.50,0.50)	0.50 (0.47,0.52)	0.50 (0.40,0.59)	0.57 (0.52,0.63)
$WAIC_1$	0.50 (0.48,0.51)	0.50 (0.48,0.52)	0.69 (0.65,0.74)	0.50 (0.47,0.52)	0.50 (0.40,0.59)	0.57 (0.52,0.62)
	(b) true model = MSRE (true test AUC=0.88)					
DIC	0.31 (0.24,0.44)	0.68 (0.50,0.76)	0.93 (0.50,0.98)	0.28 (0.22,0.35)	0.86 (0.76,0.93)	0.88 (0.85,0.93)
LPML	0.30 (0.24,0.36)	0.70 (0.63,0.76)	0.96 (0.93,0.98)	0.28 (0.22,0.34)	0.86 (0.76,0.93)	0.88 (0.85,0.93)
AIC_1	0.32 (0.24,0.47)	0.66 (0.50,0.76)	0.89 (0.50,0.98)	0.28 (0.22,0.35)	0.86 (0.76,0.93)	0.88 (0.85,0.93)
BIC_1	0.36 (0.24,0.49)	0.61 (0.50,0.75)	0.77 (0.50,0.98)	0.28 (0.22,0.34)	0.86 (0.76,0.93)	0.88 (0.85,0.93)
ICL_1	0.40 (0.25,0.50)	0.56 (0.50,0.75)	0.65 (0.50,0.98)	0.28 (0.22,0.34)	0.86 (0.76,0.93)	0.88 (0.85,0.93)
$WAIC_1$	0.30 (0.24,0.37)	0.69 (0.62,0.76)	0.96 (0.93,0.98)	0.28 (0.22,0.34)	0.86 (0.76,0.93)	0.88 (0.85,0.93)

tively. When the data are generated from the LC models with $K_D = K_C = 1$, the true AUC is always 0.5. Fitting both LC and MSRE models leads to comparable predictive performance on the test sample, with the values of PMSE and test AUC varying little among different model selection criteria, and the estimates of training AUC all centering around the true value. When fitting joint MSRE models, the values of PMSE and test AUC vary little among models selected via different criteria, with the 95% credible intervals of training AUC always covering the true AUC value for the test sample, regardless of the true structure in the primary outcome model. In particular, when the data are generated from LC models, fitting MSRE models leads to comparable predictive performance on the validation sample in comparison to those obtained by LC models. We note that this is due to our choice in setting the LC and MSRE models to be comparable. As expected, the values of test AUC are slightly smaller than those of the training AUC.

In contrast, when we study LC fitting, the values of training/test AUC and PMSE differ for different criteria and are affected by the true data generation mechanism. When fitting the LC generated data, DIC, LPML and WAIC frequently select the models that better classify the outcome with higher test AUC values than those chosen by other criteria. However, the values of PMSE and test AUC do not vary much by different criteria, with the training AUC centering around the true value 0.5, indicating that the better predictive performance of the models selected by LPML and WAIC is likely due to overfitting. When the data are generated from the MSRE models, the LC models selected by DIC, LPML and WAIC lead to a higher number and potentially outcome-informed mixture components, and consequently to the optimistically-biased training AUC relative to the test AUC. The joint LC models chosen by LPML, WAIC and DIC still lead to higher test AUC and lower MPSE, indicating somewhat more accurate prediction for the validation sample than the models chosen by other criteria. The modified AIC, BIC and ICL, which tend to select the correct numbers of mixture components frequently, perform as expected for the training versus test samples, suggesting the validity of the compatibly chosen models. Since now the data is generated from the MSRE models, fitting the LC models causes inferior predictive performance on the validation sample in comparison to those obtained by the MSRE models. However, only focusing on the predictive performance on the training sample leads to an impression that LC models tend to classify the outcome much better.

6. PENN OVARIAN AGING STUDY REVISITED

In this section, we use the knowledge obtained from the simulation study to guide us to identify the plausible models in the Penn Ovarian Aging Study with the purpose of linking the longitudinal trajectory of Follicle Stimulating Hormone (FSH) and the occurrence of severe hot flashes during the study period. In our analysis, a total of 4,244 FSH values

Table 6. Model comparison statistics from different joint models for the analysis of Penn Ovarian Aging Study data. The ten selection criteria are given in the first column of the table. The top and bottom panels correspond to the scenarios of fitting with the joint LC and MSRE models, respectively. Best fit model is given by boldface

	joint model (K_D, K_C)								
	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
					fitting joint LC model				
DIC	6930.1	**6857.9**	7024.4	6924.5	6860.5	7020.9	6973.7	6899.9	7076.3
LPML	−3794.1	−3779.6	−3772.9	−3779.8	−3763.1	−3757.3	−3780.3	−3761.5	**−3747.3**
AIC_1	6937.5	**6871.7**	7039.6	6939.7	6878.6	7047.2	6997.4	6930.9	7114.5
AIC_2	6928.9	**6860.9**	7026.9	6924.9	6861.3	7027.5	6977.1	6907.9	7088.7
BIC_1	6486.5	**6431.3**	6609.7	6513.3	6466.1	6648.7	6595.4	6546.4	6747.5
BIC_2	6477.9	**6420.5**	6596.9	6498.5	6448.8	6629.0	6575.1	6523.4	6721.8
ICL_1	6486.5	**6459.5**	6771.9	6641.7	6612.3	6902.5	6775.8	6732.4	7056.7
ICL_2	6477.9	**6448.4**	6759.6	6627.3	6594.9	6882.8	6757.0	6710.0	7032.2
$WAIC_1$	7078.8	7025.2	6986.4	7054.3	6997.1	6948.9	7045.9	6964.9	**6938.7**
$WAIC_2$	7343.8	7305.0	7279.1	7322.6	7287.9	7258.8	7322.0	7268.3	**7251.3**
AUC	0.55	0.65	0.78	0.59	0.67	0.83	0.69	0.79	0.86
					fitting joint MSRE model				
DIC	6907.9	**6852.9**	7044.0	6910.8	6862.9	7060.6	6991.4	6938.4	7128.3
LPML	−3788.2	−3784.0	−3776.9	−3773.2	−3768.8	**−3764.4**	−3778.3	−3769.9	−3768.3
AIC_1	6923.1	**6870.2**	7063.5	6932.5	6886.3	7086.3	7019.7	6968.4	7160.9
AIC_2	6912.3	**6857.5**	7048.9	6916.2	6867.6	7066.0	6997.9	6944.4	7135.5
BIC_1	6489.6	**6443.7**	6644.0	6520.0	6480.8	6687.8	6628.2	6583.9	6783.4
BIC_2	6478.8	**6431.0**	6629.4	6503.7	6462.2	6667.5	6606.4	6559.9	6758.0
ICL_1	6489.6	**6474.3**	6865.7	6653.2	6645.8	7048.7	6844.1	6828.6	7221.7
ICL_2	6478.8	**6461.3**	6852.0	6637.2	6627.1	7029.9	6824.9	6806.6	7199.6
$WAIC_1$	7062.6	7032.6	7033.7	7043.9	7012.4	**7011.6**	7046.1	7015.6	7015.8
$WAIC_2$	7328.6	7309.6	7310.5	7309.1	7293.3	**7291.5**	7314.5	7301.5	7298.9
AUC	0.68	0.69	0.69	0.67	0.68	0.68	0.67	0.68	0.68

were observed for the final sample of 245 women, with a minimum of 3 and a maximum of 26 observations per woman. Of the 245 women without severe hot flash symptoms at baseline, 118 (48.2%) had experienced the outcome of interest, an indicator variable for experiencing severe hot flashes at least once during the study. We fit both joint LC and MSRE models, as described in Section 2, to the FSH trajectories and severe hot flash outcome data, adjusting for baseline log(BMI) and smoking indicator in the primary outcome model (2).

Table 6 reports the model selection statistics for the joint LC and MSRE model for the analysis of Penn Ovarian Aging Study, with K_D and $K_C \in \{1, 2, 3\}$, respectively. For both joint LC and MSRE models, DIC, the modified AIC, BIC and ICL choose $K_D = 1$ and $K_C = 2$, while LPML and WAIC prefer models with more mixture components. Under the joint MSRE model, the AUC in the primary outcome model vary little. The AUC for the joint LC model is elevated in models with more latent classes, and WAIC and LPML tend to favor such LC models, likely due to their higher AUC values. This overall finding is not surprising and reflects some typical behaviors of these criteria as we have observed in the simulation study. In particular, the outcomes in our simulation study suggest that the notable difference between the AUC under the joint LC model chosen by WAIC, with $K_D = K_C = 3$, and the AUC under the MSRE model with any values of K_D and K_C implies potential over-fitting of the larger model. As the second best choice for both LC and MSRE models, the model with $K_D = K_C = 2$ is favored by DIC, the modified AIC and BIC. The true model is likely to be a model of $K_C = 2$ and $K_D = 1$ or, alternatively, $K_D = 2$ but with the two components closely overlapping each other. However, as already indicated by AUC values, assuming $K_D = 2$ instead of $K_D = 1$ when $K_C = 2$ had very little impact on the predictive power for both joint LC and MSRE models; Jiang et al. (2015) also reported that the effect of mean profile was not significantly associated with the risk of severe hot flash in the primary outcome model using both models. Therefore, $K_D = 1$ and $K_C = 2$ is the most parsimonious choice for both joint LC and MSRE models.

Finally, in terms of choosing between LC and MSRE models assuming $K_D = 1$ and $K_C = 2$, DIC, the modified AIC, BIC and ICL do not choose the same model although these statistics from the two models are very similar, indicating similar fit to the FSH trajectories and severe hot

flash outcome. The LC and MSRE models also share similar overall predictive performance; with ΔAUC being .04 and the corresponding credible set covering 0, $(-0.04, 0.11)$, as reported in Jiang et al. (2015). The advantages of studying this data set using both modeling approaches with different evaluation criteria lie on a higher level of confidence that suitable models are used and that the results are not heavily determined by the assumed model.

7. DISCUSSION

In this article, we studied several commonly-used model selection criteria in terms of choosing the numbers of mixture components in a joint modeling context, when both correctly and incorrectly assuming the association structure to link the longitudinal submodel and the primary outcome model. These criteria are all built upon Bayesian principles in the sense that they are evaluated over the entire posterior distribution rather than conditional on single point estimates. In particular, DIC and the modified AIC, BIC, ICL are based on deviance, while LPML is based on leave-one-out cross validated predictive density, which is shown by Watanabe (2010) to be asymptotically equivalent to WAIC.

In terms of choosing the numbers of mixture components, the performance of the modified AIC, BIC and ICL appear to be more reliable and predictable than other criteria when fitting joint LC and MSRE models with correctly assumed structure in the primary outcome model in the sense that, when the mixture components are easily separated, they frequently identify the correct numbers of mixture components while when the mixture components are fairly overlapping and hence difficult to separate, they frequently choose one instead of multiple mixture components. On the contrary, the numbers of mixture components chosen by DIC, LPML and WAIC are often more than the truth for the purpose of reaching improved prediction. In particular, WAIC and LPML tend to select the same models with higher AUC values relative to the models selected by other criteria.

For joint MSRE models, assuming different numbers of mixture components is not crucial in deciding the predictive performance as assessed by AUC values; however, for joint LC models, the predictive performance is closely related to the assumed numbers of mixture components. In particular, when the mixture components are difficult to separate and the true structure is MSRE, joint LC models tend to have a high chance of artificially creating spurious mixture components to enhance AUC values for the sample that is used to derive the model, giving the impression of much better prediction power by LC models than MSRE models. This phenomenon could cause some of our considered criteria to frequently choose incorrect numbers of mixture components and favor specific LC structure.

When this happens, our simulation studies suggest that new independent sample validation can be helpful to confirm whether the chosen models are suitable for the desired purposes. We find that the test AUC values for the validation sample based on the models chosen by WAIC and LPML also tend to be higher than the test AUC values based on the models chosen by other criteria. One needs to be cautious, however, if the training AUC values are much larger than the test AUC values for LC models. In this case, it is likely that the high predictive values of the LC model are inflated due to the outcome-informed artifact issue previously discussed, and are incorrect.

Based on our experience in the simulation study and the data example, we suggest fitting both LC and MSRE models and comparing the two sets of results with the same numbers of mixture components. When the outcome-informed artifact is present, the inference about the mixture components from the two models usually do not match and the AUC value obtained by LC model is often much higher than that by MSRE model. Otherwise, the two models assuming the same numbers of mixture components tend to lead to similar inference results, including similar scientific interpretation in the primary outcome and similar predictive power assessed by AUC. In addition to comparing the results by different model selection criteria, these rules can be helpful to guide us to choose the best suitable models we can in practice.

APPENDIX A

A.1 Posterior computations

For the probit primary outcome model, we use the Albert and Chib (1993) data augmentation method. Let W_i denote the underlying latent variable, then for LC model $W_i \sim \mathrm{N}(\boldsymbol{Q}_i'\boldsymbol{\gamma}, 1)$ and for MSRE model $W_i \sim \mathrm{N}(\boldsymbol{Q}_i'\boldsymbol{\theta}, 1)$. $o_i = 1$ is equivalent to $W_i > 0$ and $o_i = 0$ is equivalent to $W_i < 0$.

For LC model, when given $C_{ic} = 1$ and $\boldsymbol{D}_i$ as well as other covariates, we denote $\boldsymbol{Q}_i'\boldsymbol{\theta}$ by $\theta_{c,\boldsymbol{D}_i}$; when given $D_{id} = 1$ and $\boldsymbol{C}_i$ as well as other covariates, we denote $\boldsymbol{Q}_i'\boldsymbol{\theta}$ by $\theta_{\boldsymbol{C}_i,d}$.

(1) update for longitudinal submodel

- **update $\boldsymbol{D}_i$**: for $i = 1, ..., n$, the full conditional is given by $[\boldsymbol{D}_i|\cdot] \sim \text{Multinomial}(1, \tilde{\pi}_1^D, ..., \tilde{\pi}_{K_D}^D)$, where

  (11)

  LC model:

$$\tilde{\pi}_d^D = \frac{\pi_d^D \mathrm{N}_r(\boldsymbol{b}_i; \boldsymbol{\beta}_d, \boldsymbol{\Sigma}_d)\mathrm{N}(W_i; \theta_{\boldsymbol{C}_i,d}, 1)}{\sum_{d=1}^{K_D} \pi_d^D \mathrm{N}_r(\boldsymbol{b}_i; \boldsymbol{\beta}_d, \boldsymbol{\Sigma}_d)\mathrm{N}(W_i; \theta_{\boldsymbol{C}_i,d}, 1)}$$

  MSRE model:

$$\tilde{\pi}_d^D = \frac{\pi_d^D \mathrm{N}_r(\boldsymbol{b}_i; \boldsymbol{\beta}_d, \boldsymbol{\Sigma}_d)}{\sum_{d=1}^{K_D} \pi_d^D \mathrm{N}_r(\boldsymbol{b}_i; \boldsymbol{\beta}_d, \boldsymbol{\Sigma}_d)}$$

- **update $\boldsymbol{C}_i$**: for $i = 1, ..., n$, the full conditional is given by $[\boldsymbol{C}_i|\cdot] \sim \text{Multinomial}(1, \tilde{\pi}_1^C, ..., \tilde{\pi}_{K_C}^C)$, where

(12)

LC model:

$$\tilde{\pi}_c^C = \frac{\pi_c^C \mathrm{LN}(\sigma_i^2; \mu_c, \tau^2)\mathrm{N}(W_i; \theta_{c,\boldsymbol{D}_i}, 1)}{\sum_{c=1}^{K_C} \pi_c^C \mathrm{LN}(\sigma_i^2; \mu_c, \tau^2)\mathrm{N}(W_i; \theta_{c,\boldsymbol{D}_i}, 1)}$$

MSRE model:

$$\tilde{\pi}_c^C = \frac{\pi_c^C \mathrm{LN}(\sigma_i^2; \mu_c, \tau^2)}{\sum_{c=1}^{K_C} \pi_c^C \mathrm{LN}(\sigma_i^2; \mu_c, \tau^2)}$$

- **update** $\boldsymbol{\beta}_d$: for $d = 1, ..., K_D$, assuming the prior $\boldsymbol{\beta}_d \overset{ind}{\sim} \mathrm{MVN}(\boldsymbol{\nu}, \boldsymbol{V})$, then its full conditional is given by $[\boldsymbol{\beta}_d|.] \sim \mathrm{MVN}(\tilde{\boldsymbol{\nu}}_d, \tilde{\boldsymbol{V}}_d)$ where

$$\tilde{\boldsymbol{V}}_d = \left\{ \boldsymbol{V}^{-1} + \boldsymbol{\Sigma}_d^{-1} \sum_{i=1}^{n} \mathrm{I}(D_{id} = 1) \right\}^{-1}$$

$$\tilde{\boldsymbol{\nu}}_d = \tilde{\boldsymbol{V}}_d \left\{ \boldsymbol{V}^{-1}\boldsymbol{\nu} + \boldsymbol{\Sigma}_d^{-1} \sum_{i=1}^{n} \mathrm{I}(D_{id} = 1)\boldsymbol{b}_i \right\}$$

- **update** $\boldsymbol{\Sigma}_d$: for $d = 1, ..., K_D$, assuming the prior $\boldsymbol{\Sigma}_d \overset{ind}{\sim} \text{Inv-Wishart}(m, \boldsymbol{\Lambda})$, then its full conditional is given by $[\boldsymbol{\Sigma}_d|\cdot] \sim \text{Inv-Wishart}(\tilde{m}_d, \tilde{\boldsymbol{\Lambda}}_d)$ where

$$\tilde{m}_d = m + \sum_{i=1}^{n} \mathrm{I}(D_{id} = 1)$$

$$\tilde{\boldsymbol{\Lambda}}_d = \left\{ \Lambda^{-1} + \sum_{i=1}^{n} \mathrm{I}(D_{id} = 1) \left(\boldsymbol{b}_i - \boldsymbol{\beta}_d\right) \left(\boldsymbol{b}_i - \boldsymbol{\beta}_d\right)' \right\}^{-1}$$

- **update** $\{\pi_d^D\}_d$: assuming the prior $\{\pi_d^D\}_d \sim \text{Dirichlet}(e_1^D, ..., e_{K_D}^D)$ then its full conditional is

$$\left[\{\pi_d^D\}_d|\cdot\right] \sim \text{Dirichlet}(\{e_d^D + \sum_{i=1}^{n} \mathrm{I}(D_{id} = 1)\}_d).$$

- **update** μ_c: for $c = 1, \cdots, K_C$, assuming the prior $\mu_c \overset{ind}{\sim} \mathrm{N}(a, b)$, then its full conditional is given by $[\mu_c|\cdot] \sim \mathrm{N}(\tilde{a}, \tilde{b})$ where

$$\tilde{a} = \frac{\sum_{i=1}^{n} \mathrm{I}(C_{ic} = 1)\log\sigma_i^2/\tau^2 + a/b}{1/b + \sum_{i=1}^{n} \mathrm{I}(C_{ic} = 1)/\tau^2}$$

$$\tilde{b} = \left\{ 1/b + \sum_{i=1}^{n} \mathrm{I}(C_{ic} = 1)/\tau^2 \right\}^{-1}$$

- **update** τ^2: assuming $\tau^2 \sim \mathrm{IG}(v, e)$, then the full conditional posterior distribution is $\left[\tau^2|\cdot\right] \sim$

$$\mathrm{IG}\left\{ v + \frac{n}{2}, e + \sum_{i=1}^{n}\sum_{c=1}^{K_C} \frac{1}{2}\mathrm{I}(C_{ic} = 1)\left(\log\sigma_i^2 - \mu_c\right)^2 \right\}.$$

- **update** $\{\pi_c^C\}_c$: assuming the prior $\{\pi_c^C\}_c \sim \text{Dirichlet}(e_1^C, ..., e_{K_C}^C)$ then its full conditional is

$$\left[\{\pi_c^C\}_c|\cdot\right] \sim \text{Dirichlet}(\{e_c^C + \sum_{i=1}^{n} \mathrm{I}(C_{ic} = 1)\}_c).$$

- **update** $\boldsymbol{b}_i$, for $i = 1, ..., n$, its full conditional is $\boldsymbol{b}_i$ $[\boldsymbol{b}_i|.] \sim \mathrm{MVN}(\tilde{\boldsymbol{\beta}}_i, \tilde{\boldsymbol{\Sigma}}_i)$, where
LC model

$$\tilde{\boldsymbol{\Sigma}}_{id} = \left(\boldsymbol{\Sigma}_{I(D_{id}=1)}^{-1} + \frac{1}{\sigma_i^2} \sum_{j=1}^{n_i} \boldsymbol{t}_{ij}\boldsymbol{t}_{ij}' \right)^{-1}$$

$$\tilde{\boldsymbol{\beta}}_i = \tilde{\boldsymbol{\Sigma}}_{id} \left(\Sigma_{D_i}^{-1}\boldsymbol{\beta}_{D_i} + \frac{1}{\sigma_i^2} \sum_{j=1}^{n_i} y_{ij}\boldsymbol{t}_{ij} \right)$$

MSRE model

$$\tilde{\boldsymbol{\beta}}_i = \tilde{\boldsymbol{\Sigma}}_{id} \left\{ \boldsymbol{\Sigma}_{D_i}^{-1}\boldsymbol{\beta}_{D_i} + \frac{1}{\sigma_i^2} \sum_{j=1}^{n_i} y_{ij}\boldsymbol{t}_{ij} + (\boldsymbol{Q}_i^T\boldsymbol{\eta} - \tilde{g}(\boldsymbol{\eta}, \sigma_i^2))\boldsymbol{g}(\boldsymbol{\eta}, \sigma_i^2) \right\}$$

$$\tilde{\boldsymbol{\Sigma}}_{id} = \left\{ \boldsymbol{\Sigma}_{D_i}^{-1} + \frac{1}{\sigma_i^2} \sum_{j=1}^{n_i} \boldsymbol{t}_{ij}\boldsymbol{t}_{ij}^T + \boldsymbol{g}(\boldsymbol{\eta}, \sigma_i^2)\boldsymbol{g}(\boldsymbol{\eta}, \sigma_i^2)^T \right\}^{-1}$$

where, $\boldsymbol{t}_{ij}$ is a vector of functional forms of the observation time t_{ij} for the longitudinal measurement y_{ij} such that $y_{ij} \sim \mathrm{N}\{\mu(\boldsymbol{b}_i; t_{ij}), \sigma_i^2\}$ with $\mu(\boldsymbol{b}_i; t_{ij}) = \boldsymbol{b}_i^T\boldsymbol{t}_{ij}$. $\boldsymbol{g}(\boldsymbol{\eta}, \sigma_i^2)$ is a vector such that $\boldsymbol{Q}_i^T\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\eta}, \sigma_i^2)'\boldsymbol{b}_i + \tilde{g}(\boldsymbol{\eta}, \sigma_i^2)$ in MSRE model.

- **update** the variances σ_i^2, for $i = 1, ..., n$
LC model

$$\pi(\sigma_i^2|.) \propto \prod_{c=1}^{K_C} \mathrm{LN}\left(\sigma_i^2; \mu_c, \tau^2\right)^{\mathrm{I}(C_{ic}=1)} \times \prod_{j=1}^{n_i} \mathrm{N}\left\{y_{ij}; \mu(\boldsymbol{b}_i; t_{ij}), \sigma_i^2\right\}$$

MSRE model

$$\pi(\sigma_i^2|.) \propto \prod_{c=1}^{K_C} \mathrm{LN}\left(\sigma_i^2; \mu_c, \tau^2\right)^{\mathrm{I}(C_{ic}=1)} \times \prod_{j=1}^{n_i} \mathrm{N}\left\{y_{ij}; \mu(\boldsymbol{b}_i; t_{ij}), \sigma_i^2\right\} \mathrm{N}(W_i; \boldsymbol{Q}_i^T\boldsymbol{\eta}, 1)$$

(2) update for primary outcome model:

- **update** W_i, $i = 1, ..., n$

$$[W_i|o_i = 1, \cdot] \sim \mathrm{N}(\boldsymbol{Q}_i^T\boldsymbol{\eta}, 1)I_{(0,\infty)}(\cdot)$$
$$[W_i|o_i = 0, \cdot] \sim \mathrm{N}(\boldsymbol{Q}_i^T\boldsymbol{\eta}, 1)I_{(-\infty,0)}(\cdot)$$

- **update $\boldsymbol{\eta}$**: assuming the prior for $\boldsymbol{\eta} \sim \mathrm{MVN}(\boldsymbol{\nu}_{\boldsymbol{\eta}}, \boldsymbol{V}_{\boldsymbol{\eta}})$, then the full is given by $[\boldsymbol{\eta}|.] \sim \mathrm{MVN}(\tilde{\boldsymbol{\nu}}_{\boldsymbol{\eta}}, \tilde{\boldsymbol{V}}_{\boldsymbol{\eta}})$ where

$$\tilde{\boldsymbol{V}}_{\boldsymbol{\eta}} = \left(\boldsymbol{V}_{\boldsymbol{\eta}}^{-1} + \sum_{i=1}^{n} \boldsymbol{Q}_i\boldsymbol{Q}_i'\right)^{-1}$$
$$\tilde{\boldsymbol{\nu}}_{\boldsymbol{\eta}} = \tilde{\boldsymbol{V}}_{\boldsymbol{\eta}}\left(\boldsymbol{V}_{\boldsymbol{\eta}}^{-1}\boldsymbol{\nu}_{\boldsymbol{\eta}} + \sum_{i=1}^{n} W_i\boldsymbol{Q}_i\right)$$

  $\boldsymbol{Q}_i$ is the i^{th} row of the design matrix in the outcome model given $\boldsymbol{D}_i$ and $\boldsymbol{C}_i$ for LC model or $\boldsymbol{b}_i$ and σ_i^2 for MSRE model as well as other covariates.

A.2 Computation of LPML

For our considered joint models, we have

$$\begin{aligned}
\mathrm{CPO}_i^{-1} &= \frac{f(\boldsymbol{y}_{(-i)}, \boldsymbol{o}_{(-i)}|\boldsymbol{v})}{f(\boldsymbol{y}, \boldsymbol{o}|\boldsymbol{v})} \\
&= \int \frac{f(\boldsymbol{y}_{(-i)}, \boldsymbol{o}_{(-i)}|\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{b}, \boldsymbol{\sigma}, \boldsymbol{\phi}, \boldsymbol{v}) f(\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{b}, \boldsymbol{\sigma}) \pi(\boldsymbol{\phi})}{f(\boldsymbol{y}, \boldsymbol{o}|\boldsymbol{v})} d\boldsymbol{b} d\boldsymbol{\sigma} d\boldsymbol{C} d\boldsymbol{D} d\boldsymbol{\phi} \\
&= \int \frac{f(\boldsymbol{y}, \boldsymbol{o}|\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{b}, \boldsymbol{\sigma}, \boldsymbol{\phi}, \boldsymbol{v}) f(\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{b}, \boldsymbol{\sigma}) \pi(\boldsymbol{\phi})}{f(\boldsymbol{y}, \boldsymbol{o}|\boldsymbol{v}) f(\boldsymbol{y}_i, o_i|\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{b}, \boldsymbol{\sigma}, \boldsymbol{\phi}, \boldsymbol{v})} d\boldsymbol{b} d\boldsymbol{\sigma} d\boldsymbol{C} d\boldsymbol{D} d\boldsymbol{\phi} \\
&= \int \frac{f(\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{b}, \boldsymbol{\sigma}, \boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{o}, \boldsymbol{v})}{f(\boldsymbol{y}_i, o_i|\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{b}, \boldsymbol{\sigma}, \boldsymbol{\phi}, \boldsymbol{v})} d\boldsymbol{b} d\boldsymbol{\sigma} d\boldsymbol{C} d\boldsymbol{D} d\boldsymbol{\phi} \\
&= \int \frac{f(\boldsymbol{C}, \boldsymbol{D}, \boldsymbol{b}, \boldsymbol{\sigma}, \boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{o}, \boldsymbol{v})}{f(\boldsymbol{y}_i|\boldsymbol{b}_i, \sigma_i, \boldsymbol{\phi}, \boldsymbol{v}_i) f(o_i|C_i, D_i, \boldsymbol{b}_i, \sigma_i, \boldsymbol{\phi}, \boldsymbol{v}_i)} d\boldsymbol{b} d\boldsymbol{\sigma} d\boldsymbol{C} d\boldsymbol{D} d\boldsymbol{\phi}
\end{aligned}$$

where $\boldsymbol{v} = (\boldsymbol{v}_1, ..., \boldsymbol{v}_n)^T$ includes all observed variables including obervation time t_{ij}, $i = 1, ..., n_i$, $j = 1, ..., n_i$ and baseline covariates of interest. $\pi(\boldsymbol{\phi})$ denotes the joint prior for $\boldsymbol{\phi}$. Using the MCMC posterior draws, we estimate CPO_i^{-1} by

$$\frac{1}{S}\sum_{s=1}^{S} f^{-1}(\boldsymbol{y}_i|\boldsymbol{b}_i^{(s)}, \sigma_i^{(s)}, \boldsymbol{\phi}^{(s)}, \boldsymbol{v}_i) f^{-1}(o_i|C_i^{(s)}, D_i^{(s)}, \boldsymbol{b}_i^{(s)}, \sigma_i^{(s)}, \boldsymbol{\phi}^{(s)}, \boldsymbol{v}_i)$$

where S is the number of MCMC posterior draws and $\boldsymbol{\phi}^{(s)}$ is the vector of the posterior draws of all model parameters at the s^{th} iteration. We have,

$$f(\boldsymbol{y}_i|\boldsymbol{b}_i, \sigma_i^2, \boldsymbol{\phi}, \boldsymbol{v}_i) = \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{\{y_{ij} - f(\boldsymbol{b}_i; t_{ij})\}^2}{2\sigma_i^2}\right]$$
$$f(o_i|C_i, D_i, \boldsymbol{\phi}, \boldsymbol{v}_i) = \Phi(\boldsymbol{Q}_i^T\boldsymbol{\theta})^{o_i}\left[1 - \Phi(\boldsymbol{Q}_i^T\boldsymbol{\theta})\right]^{1-o_i} \text{ for LC}$$
$$f(o_i|\boldsymbol{b}_i, \sigma_i, \boldsymbol{\phi}, \boldsymbol{v}_i) = \Phi(\boldsymbol{Q}_i^T\boldsymbol{\gamma})^{o_i}\left[1 - \Phi(\boldsymbol{Q}_i^T\boldsymbol{\gamma})\right]^{1-o_i} \text{ for MSRE}$$

We retain every 5^{th} of the 100,000 posterior draws after the chains converge and divide these posterior draws into 20 blocks of length 1,000 draws. To obtain stable LPML measures, we calculate the CPO's and LPML based on each of the 20 blocks of draws and then report the median LPML. We found this approach would lead to relatively stable LPML results in our simulations.

A.3 Computation of DIC

Let $\boldsymbol{\phi}$ denote the vector of all model parameters and $\boldsymbol{Z}_i$ the latent variables $(\boldsymbol{D}_i, \boldsymbol{C}_i, \boldsymbol{b}_i, \sigma_i^2, W_i)'$ for the ith subject. The data $\boldsymbol{x}_i'$, $i = 1, ..., n$ correspond to the longitudinal data $(y_{i1}, ..., y_{in_i})'$ and the outcome o_i.

For MSRE model, we divide $\boldsymbol{Z}_i$ into $\boldsymbol{Z}_{i1} = (\boldsymbol{D}_i, \boldsymbol{C}_i)$ and $\boldsymbol{Z}_{i2} = (\boldsymbol{b}_i, \sigma_i^2, W_i)$, then for the complete data log-likelihood (ignoring normalizing constants), we have

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{Z},\boldsymbol{\phi}}\left\{\log f(\boldsymbol{x}, \boldsymbol{Z} \mid \boldsymbol{\phi}) \mid \boldsymbol{x}\right\} \\
&\quad = \mathrm{E}_{\boldsymbol{Z}_2,\boldsymbol{\phi}}\left[\mathrm{E}_{\boldsymbol{Z}_1}\left\{\log f(\boldsymbol{x}, \boldsymbol{Z} \mid \boldsymbol{\phi}) \mid \boldsymbol{x}, \boldsymbol{\phi}, \boldsymbol{Z}_2\right\}\right]
\end{aligned}$$

The expectation $\mathrm{E}_{\boldsymbol{Z},\boldsymbol{\phi}}\{\log f(\boldsymbol{x}, \boldsymbol{Z} \mid \boldsymbol{\phi}) \mid \boldsymbol{x}\}$ can then be approximated from MCMC draws since

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{Z}_1}\left\{\log f(\boldsymbol{x}, \boldsymbol{Z} \mid \boldsymbol{\phi}) \mid \boldsymbol{x}, \boldsymbol{\phi}, \boldsymbol{Z}_2\right\} \\
&\quad = \sum_{i=1}^{n}\Bigg[\sum_d \tilde{\pi}_d^D \mathrm{N}(\boldsymbol{b}_i; \boldsymbol{\beta}_d, \boldsymbol{\Sigma}_d) + \sum_c \tilde{\pi}_c^C \mathrm{LN}(\sigma_i^2; \mu_c, \tau^2) \\
&\qquad + \sum_{j=1}^{n_i} \mathrm{N}(y_{ij}; \mu(\boldsymbol{b}_i; t_{ij}), \sigma_i^2) \\
&\qquad + o_i \log\Phi(\boldsymbol{Q}_i'\boldsymbol{\eta}) + (1 - o_i)\log\left\{1 - \Phi(\boldsymbol{Q}_i'\boldsymbol{\eta})\right\}\Bigg].
\end{aligned}$$

To obtain $E_{\boldsymbol{Z}}[\log f\{\boldsymbol{x}, \boldsymbol{Z} \mid \ E_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \boldsymbol{x}, \boldsymbol{Z})\} \mid \boldsymbol{x}]$, we use the same approach of averaging over the MCMC draws to integrate with respect to $\boldsymbol{Z}$, but instead of using the draws of the model parameters directly, we need to obtain their expectation conditional on the current draw of $\boldsymbol{Z}$. So

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{Z}}\left[\log f\left\{\boldsymbol{x}, \boldsymbol{Z} \mid \mathrm{E}_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \boldsymbol{x}, \boldsymbol{Z})\right\} \mid \boldsymbol{x}\right] \approx M^{-1}\sum_{m=1}^{M}\Bigg[\sum_{i=1}^{n} \\
&\sum_d I(D_{id}^{(m)} = 1)\left\{\log\hat{\pi}_d^{(m)} - \log\mathrm{N}(\boldsymbol{b}_i^{(m)}; \hat{\boldsymbol{\beta}}_d^{(m)}, \hat{\boldsymbol{\Sigma}}_d^{(m)})\right\} \\
&+ \sum_c I(C_{ic}^{(m)} = 1)\left\{\log\hat{\pi}_c^{(m)} - \log\mathrm{LN}(\sigma_i^{(m)2}; \hat{\mu}_c^{(m)}, \hat{\tau}^{(m)2})\right\} \\
&+ \sum_{j=1}^{n_i}\log\mathrm{N}(y_{ij}; \mu(\boldsymbol{b}_i^{(m)}; t_{ij}), \sigma_i^{(m)2}) \\
&\quad + o_i\log\Phi(\boldsymbol{Q}_i'\boldsymbol{\eta}^{(m)}) + (1 - o_i)\log\left\{1 - \Phi(\boldsymbol{Q}_i'\boldsymbol{\eta}^{(m)})\right\}\Bigg]
\end{aligned}$$

where $\hat{\boldsymbol{\phi}}^{(m)} = E_{\boldsymbol{\phi}}(\boldsymbol{\phi} \mid \boldsymbol{x}, \boldsymbol{Z}^{(m)})$.

Some components of $\hat{\boldsymbol{\phi}}^{(m)}$ have closed form solutions:

$$\hat{\pi}_d^{(m)} = \frac{e_d^D + \sum_{i=1}^{n} \mathrm{I}(D_{id}^{(m)} = 1)}{\sum_d e_d^D + n}$$

$$\hat{\pi}_c^{(m)} = \frac{e_c^C + \sum_{i=1}^n \mathrm{I}(C_{ic}^{(m)} = 1)}{\sum_c e_c^C + n}$$

$$\hat{\boldsymbol{\eta}}^{(m)} = \left(V_{\boldsymbol{\eta}}^{-1} + \sum_{i=1}^n \boldsymbol{Q}_i^{(m)} \boldsymbol{Q}_i^{(m)'} \right)^{-1} \times \left(V_{\boldsymbol{\eta}}^{-1} \nu_{\boldsymbol{\eta}} + \sum_{i=1}^n W_i^{(m)} \boldsymbol{Q}_i^{(m)} \right)$$

where $\boldsymbol{Q}_i^{(m)}$ is the i^{th} row of the design matrix in the probit submodel for the mth MCMC draw and e_d^D, e_c^C, $V_{\boldsymbol{\eta}}$, and $\boldsymbol{\nu}_{\boldsymbol{\eta}}$ are specified hyperprior values.

The other components of $\hat{\boldsymbol{\phi}}^{(m)}$ will have to be obtained by running small MCMC chains *for each of the main MCMC iterations* to get the marginal expectations: $\hat{\boldsymbol{\beta}}_d^{(m)} = (M^*)^{-1} \sum_{m*} \boldsymbol{\beta}_d^{(m,m^*)}$ and $\hat{\boldsymbol{\Sigma}}_d^{(m)} = (M^*)^{-1} \sum_{m*} \Sigma_d^{(m,m^*)}$, where $\boldsymbol{\beta}_d^{(m,m^*)}$ and $\boldsymbol{\Sigma}_d^{(m,m^*)}$ are obtained by alternating draws from the following distributions with known hyperparameters V, ν, m, and $\boldsymbol{\Lambda}$:

$\boldsymbol{\beta}_d^{(m,m^*)} \sim \mathbf{MVN}(\tilde{\boldsymbol{\nu}}_d^{(m,m^*)}, \tilde{\boldsymbol{V}}_d^{(m,m^*)})$, where

$$\tilde{\boldsymbol{V}}_d^{(m,m^*)} = \left\{ \boldsymbol{V}^{-1} + (\boldsymbol{\Sigma}_d^{(m,m^*)})^{-1} \sum_{i=1}^n \mathrm{I}(D_{id}^{(m)} = 1) \right\}^{-1}$$

$$\tilde{\boldsymbol{\nu}}_d^{(m,m^*)} = \tilde{\boldsymbol{V}}_d^{(m,m^*)} \left\{ \boldsymbol{V}^{-1} \boldsymbol{\nu} + (\boldsymbol{\Sigma}_d^{(m,m^*)})^{-1} \sum_{i=1}^n \mathrm{I}(D_{id}^{(m)} = 1) \boldsymbol{b}_i^{(m)} \right\}.$$

$\boldsymbol{\Sigma}_d^{(m,m^*)} \sim \mathbf{Inv\text{-}Wishart}(\tilde{m}_d^{(m)}, \tilde{\boldsymbol{\Lambda}}_d^{(m,m^*)})$, where

$$\tilde{m}_d^{(m)} = m + \sum_{i=1}^n \mathrm{I}(D_{id}^{(m)} = 1),$$

$$\tilde{\boldsymbol{\Lambda}}_d^{(m,m^*)} = \left\{ \Lambda^{-1} + \sum_{i=1}^n \mathrm{I}(D_{id}^{(m)} = 1) \left(\boldsymbol{b}_i^{(m)} - \boldsymbol{\beta}_d^{(m,m^*)} \right) \times \left(\boldsymbol{b}_i^{(m)} - \boldsymbol{\beta}_d^{(m,m^*)} \right)' \right\}^{-1}.$$

Similarly, $\hat{\mu}_c^{(m)} = (M^*)^{-1} \sum_{m*} \mu_c^{(m,m^*)}$ and $(\hat{\tau}^2)^{(m)} = (M^*)^{-1} \sum_{m*} (\tau^2)^{(m,m^*)}$, where $\mu_c^{(m,m^*)}$ and $(\tau^2)^{(m,m^*)}$ are obtained by alternating draws from the following distributions with known hyperparameters a, b, e, and f:

$\mu_c^{(m,m^*)} \sim \mathbf{N}(\tilde{a}^{(m,m^*)}, \tilde{b}^{(m,m^*)})$, where

$$\tilde{a}^{(m,m^*)} = \frac{\sum_{i=1}^n \mathrm{I}(C_{ic}^{(m)} = 1) \log(\sigma_i^2)^{(m)} / (\tau^2)^{(m,m^*)} + a/b}{1/b + \sum_{i=1}^n \mathrm{I}(C_{ic}^{(m)} = 1) / (\tau^2)^{(m,m^*)}}$$

$$\tilde{b}^{(m,m^*)} = \left\{ 1/b + \sum_{i=1}^n \mathrm{I}(C_{ic}^{(m)} = 1) / (\tau^2)^{(m,m^*)} \right\}^{-1}$$

$(\tau^2)^{(m,m^*)} \sim \mathbf{IG}(\tilde{v}, \tilde{e}^{(m,m^*)})$, where

$$\tilde{v} = v + \frac{n}{2}$$

$$\tilde{e}^{(m,m^*)} = e + \sum_{i=1}^n \sum_{c=1}^{K_C} \frac{1}{2} \mathrm{I}(C_{ic}^{(m)} = 1) \times \left\{ \log(\sigma_i^2)^{(m)} - \mu_c^{(m,m^*)} \right\}^2$$

A modest number of drawn (here we use $M^* = 250$) is found to be sufficient to obtain an accurate approximation.

Similarly, we can obtain DIC for our LC model.

A.4 Computation details to predict outcome for new validation sample

In this section, we give the details to draw $\tilde{\boldsymbol{Z}}^{(m)}$, $\boldsymbol{Z}^{(m)}$, $\boldsymbol{\phi}_{\text{long}}^{(m)}$ and $\boldsymbol{\eta}^{(m)}$ from the posterior distribution $p(\tilde{\boldsymbol{Z}}, \boldsymbol{Z}, \boldsymbol{\phi}_{\text{long}}, \boldsymbol{\eta} | \tilde{\boldsymbol{y}}, \boldsymbol{y}, \boldsymbol{o}, \mathrm{H}_a)$, $m = 1, \cdots, M$ for some large M. After initializing the chain, we repeat the following steps (1) to (5) for $m = 1, \cdots, M$:

(1) update individual level latent variables $\tilde{\boldsymbol{Z}}$ for validation sample (note: this is only conditional on the longitudinal trajectories $\tilde{\boldsymbol{y}} = \{\tilde{\boldsymbol{y}}_i\}_{i=1}^{\tilde{n}}$)

- **draw** $\tilde{\boldsymbol{D}}_i^{(m)}$, for $i = 1, ..., \tilde{n}$ from Multinomial $(1, \tilde{\pi}_1^{D(m)}, ..., \tilde{\pi}_{K_D}^{D(m)})$, where

$$\tilde{\pi}_d^{D(m)} = \frac{\pi_d^{D(m)} \mathrm{N}_r(\boldsymbol{b}_i^{(m)}; \boldsymbol{\beta}_d^{(m)}, \boldsymbol{\Sigma}_d^{(m)})}{\sum_{d=1}^{K_D} \pi_d^{D(m)} \mathrm{N}_r(\boldsymbol{b}_i^{(m)}; \boldsymbol{\beta}_d^{(m)}, \boldsymbol{\Sigma}_d^{(m)})}$$

- **draw** $\tilde{\boldsymbol{b}}_i^{(m)}$, for $i = 1, ..., \tilde{n}$ from $\mathrm{MVN}(\tilde{\boldsymbol{\beta}}_i^{(m)}, \tilde{\boldsymbol{\Sigma}}_i^{(m)})$, where

$$\tilde{\boldsymbol{\Sigma}}_i = \left((\boldsymbol{\Sigma}_{\tilde{D}_i^{(m)}}^{(m)})^{-1} + \frac{1}{(\sigma_i^{(m)})^2} \sum_{j=1}^{\tilde{n}_i} \boldsymbol{t}_{ij} \boldsymbol{t}_{ij}' \right)^{-1}$$

$$\tilde{\boldsymbol{\beta}}_i = \tilde{\boldsymbol{\Sigma}}_i \left[(\boldsymbol{\Sigma}_{\tilde{D}_i^{(m)}}^{(m)})^{-1} \boldsymbol{\beta}_{\tilde{D}_i^{(m)}}^{(m)} + \frac{1}{(\sigma_i^{(m)})^2} \sum_{j=1}^{n_i} \tilde{y}_{ij} \boldsymbol{t}_{ij} \right]$$

- **draw** $\tilde{\boldsymbol{C}}_i^{(m)}$, for $i = 1, ..., \tilde{n}$ from Multinomial $(1, \tilde{\pi}_1^{C(m)}, ..., \tilde{\pi}_{K_C}^{C(m)})$, where

$$\tilde{\pi}_c^C = \frac{\pi_c^{C(m)} \mathrm{LN}(\sigma_i^{(m)2}; \mu_c^{(m)}, \tau^{(m)2})}{\sum_{c=1}^{K_C} \pi_c^{C(m)} \mathrm{LN}(\sigma_i^{(m)2}; \mu_c^{(m)}, \tau^{(m)2})}$$

- **draw** the variances $\tilde{\sigma}_i^{(m)2}$, $i = 1, ..., \tilde{n}$ from

$$\pi((\tilde{\sigma}_i^{(m)})^2|.)$$
$$\propto \prod_{c=1}^{K_C} \mathrm{LN}\left(\tilde{\sigma}_i^{(m)})^2; \mu_c^{(m)}, (\tau^{(m)})^2\right)^{\mathrm{I}(\tilde{C}_i^{(m)}=c)}$$
$$\times \prod_{j=1}^{\tilde{n}_i} \mathrm{N}\left\{y_{ij}; \mu(\tilde{\boldsymbol{b}}_i^{(m)}; t_{ij}), (\tilde{\sigma}_i^{(m)})^2\right\}$$

(2) update individual level latent variables $\boldsymbol{Z}$ for old sample (note: this is conditional on both the longitudinal trajectories $\boldsymbol{y} = \{\boldsymbol{y}_i\}_{i=1}^n$ and the outcome $\boldsymbol{o} = \{o_i\}_{i=1}^n$): they have the same full conditional posterior density as given in Appendix A.1.

(3) update population level parameter $\phi_{\mathbf{long}}$ in the longitudinal submodel: they have the same full conditional posterior density as given in Appendix A.1 except now the updating is based on both $\tilde{\boldsymbol{Z}}$ for validation sample and $\boldsymbol{Z}$ for old sample.

(4) update the parameter η in the primary outcome model (note: this is only conditional on the old sample): it has the same full conditional posterior density as given in Appendix A.1.

(5) The prediction of outcome for a new validation sample i, $i = 1, \cdots, \tilde{n}$ can be based on probability $\tilde{p}_i^{(m)} = \Phi(\boldsymbol{\eta}^{(m)T}\tilde{\boldsymbol{Q}}_i^{(m)})$, where $\tilde{\boldsymbol{Q}}_i^{(m)}$ contains $\tilde{\boldsymbol{D}}_i^{(m)}$ and $\tilde{\boldsymbol{C}}_i^{(m)}$ for LC model; and $\tilde{\boldsymbol{Q}}_i^{(m)}$ contains $\tilde{\boldsymbol{b}}_i^{(m)}$ and $\tilde{\sigma}_i^{(m)}$ for MSRE model.

Received 11 August 2014

REFERENCES

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory* (B. N. Petrov and F. Csaki, eds.) 267–281. Budapest: Akademiai Kiado. MR0483125

[2] Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association* **88** 669–679. MR1224394

[3] Biernacki, C., Celeux, G. and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **22** 719–725.

[4] Celeux, G., Forbes, F., Robert, C. P. and Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis* **1** 651–673. MR2282197

[5] Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56** 463–474. MR0254956

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

[7] 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. MR2972319

[8] 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.

[9] Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer. MR2265601

[10] Garrett, E. S. and Zeger, S. L. (2000). Latent class model diagnosis. *Biometrics* **56** 1055–1067. MR1815583

[11] Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74** 153–160. MR0529531

[12] Gelman, A., Hwang, J. and Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*. DOI: 10.1007/s11222-013-9416-2.

[13] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). *Bayesian data analysis*. CRC Press. MR2027492

[14] Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001). *Bayesian survival analysis*. New York: Springer-Verlag. MR1876598

[15] Jiang, B., Elliott, M. R., Sammel, M. D. and Wang, N. (2015). Joint modeling of cross-sectional health outcomes and longitudinal predictors via mixtures of means and variances. *Biometrics*. DOI: 10.1111/biom.12284.

[16] 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. MR2221285

[17] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley. MR1789474

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

[19] 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. MR2898840

[20] Plummer, M. (2002). Discussion of the paper by Spiegelhalter et al. *Journal of the Royal Statistical Society: Statistical Methodology* **4** 620.

[21] Plummer, M. (2006). Comment on article by Celeux et al. *Bayesian Analysis* **4** 681–686. MR2282200

[22] Proust-Lima, C. and Taylor, J. M. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biostatistics* **10** 535–549.

[23] 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*.

[24] Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67** 819–829. MR2829256

[25] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464. MR0468014

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

[27] Speigelhalter, D., Best, N., Carlin, B. and Van der Linde, A. (2003). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* **64** 583–616. MR1979380

[28] Steel, R. J. and Raftery, A. E. (2010). Performance of Bayesian model selection criteria for Gaussian mixture models. In *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger* (M.-H. Chen, P. Müller, D. Sun, K. Ye and D. K. Dey, eds.) 113–130. New York: Springer. MR2766461

[29] Taylor, J. M., Ankerst, D. P. and Andridge, R. R. (2008). Validation of biomarker-based risk prediction models. *Clinical Cancer Research* **14** 5977–5983.

[30] Taylor, J. M., Yu, M. and Sandler, H. M. (2005). Individualized predictions of disease progression following radiation therapy for prostate cancer. *Journal of Clinical Oncology* **23** 816–825.

[31] Taylor, J. M., Park, Y., Ankerst, D. P., Proust-Lima, C., Williams, S., Kestin, L., Bae, K., Pickles, T. and Sandler, H. (2013). Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics* **69(1)** 206–13. MR3058067

[32] Vehtari, A. and Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys* **6** 142–228. MR3011074

[33] Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press. MR2554932

[34] Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research* **9999** 3571–3594. MR2756194

[35] Yu, M., Taylor, J. M. G. and Sandler, H. M. (2008). Individual prediction in prostate cancer studies using a joint longitudinal survival–cure model. *Journal of the American Statistical Association* **103** 178–187. MR2420225

Bei Jiang
Department of Biostatistics
Columbia University
New York, NY 10032
USA

Division of Biostatistics
Department of Child and Adolescent Psychiatry
New York University
New York, NY 10016
USA
E-mail address: bj2332@cumc.columbia.edu

Michael R. Elliott
Department of Biostatistics
Survey Methodology Program
Institute for Social Research
University of Michigan
Ann Arbor, MI 48109
USA
E-mail address: mrelliot@umich.edu

Mary D. Sammel
Center for Clinical Epidemiology and Biostatistics
Perelman School of Medicine
University of Pennsylvania
Philadelphia, PA 19104
USA
E-mail address: msammel@mail.med.upenn.edu

Naisyin Wang
Department of Statistics
University of Michigan
Ann Arbor, MI 48109
USA
E-mail address: nwangaa@umich.edu