Research Article

Statistics in Medicine

Received 26 September 2012, Accepted 12 June 2013 Published online 2 August 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5906

Joint analysis of stochastic processes with application to smoking patterns and insomnia

Sheng Luo[*†]

This article proposes a joint modeling framework for longitudinal insomnia measurements and a stochastic smoking cessation process in the presence of a latent permanent quitting state (i.e., 'cure'). We use a generalized linear mixed-effects model and a stochastic mixed-effects model for the longitudinal measurements of insomnia symptom and for the smoking cessation process, respectively. We link these two models together via the latent random effects. We develop a Bayesian framework and Markov Chain Monte Carlo algorithm to obtain the parameter estimates. We formulate and compute the likelihood functions involving time-dependent covariates. We explore the within-subject correlation between insomnia and smoking processes. We apply the proposed methodology to simulation studies and the motivating dataset, that is, the Alpha-Tocopherol, Beta-Carotene Lung Cancer Prevention study, a large longitudinal cohort study of smokers from Finland. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: cure model; MCMC; mixed-effects model; joint modeling; recurrent events; Bayes

1. Introduction

Insomnia is the most commonly reported sleep problem, which affects millions of individuals worldwide, giving rise to emotional distress, daytime fatigue, and loss of productivity. With the reported prevalence of insomnia ranging anywhere from 10% to 50% in the general population [1–3], the number of affected individuals could be quite large. The association between cigarette smoking and insomnia has been reported [4–7]. First of all, the stimulant effects of nicotine in cigarette contribute to insomnia. Conversely, if smoking cessation is initiated, insomnia is one of the common cigarette withdrawal symptoms. In addition, insomnia could play a role in the motivation to smoke. The clear understanding of the relationship between cigarette smoking and insomnia has important clinical and public health implications. If smoking is causally related to insomnia, smoking cessation interventions have the potential to significantly reduce the occurrence of insomnia and the associated decrement in functioning [5]. The objectives of this article are to characterize the feedback of insomnia upon smoking while accounting for other covariates [8] and to give insight into the potential correlation between the probability of having insomnia and the smoking transition probabilities.

This article is motivated by the Alpha-Tocopherol, Beta-Carotene (ATBC) Lung Cancer Prevention study, a large longitudinal study with 26, 215 current smokers sponsored by National Cancer Institute. Each individual was followed 5 to 8 years and had a clinic visit every 4 months. At each visit, each individual was asked about their smoking status and health status since the last visit. Specifically, smoking status and insomnia status were defined by the questions 'Have you smoked since your last visit?' and 'Have you had the symptom or trouble of insomnia since your last visit?', respectively. The details of this study can be found in ATBC Study Group [9]. The smoking patterns alternate between smoking and nonsmoking states with sojourn time in each state differs within and across individuals. The presence of long trailing nonsmoking intervals before censoring in some individuals indicates the potential

Division of Biostatistics, University of Texas School of Public Health, 1200 Pressler St, Houston, Texas 77030, U.S.A.
**Correspondence to: Sheng Luo, Division of Biostatistics, University of Texas School of Public Health, 1200 Pressler St, Houston, Texas 77030, U.S.A.*
†E-mail: sheng.t.luo@uth.tmc.edu

existence of permanent quitting. To fully model the stochastic nature of the complex smoking patterns, Luo *et al.* [10] proposed a discrete-time mixed-effects model with three states: smoking, transient cessation (temporarily nonsmoking with subsequent relapse), and permanent cessation (lifelong smoke-free, latent state due to censoring). Random subject-specific transition probabilities among these three states were used to account for the between-subject variability. Luo *et al.* [10] developed a computationally fast method of maximizing the marginal likelihood obtained by integrating over the Beta distribution of the transition probabilities among three states. Luo *et al.* [11] used a different modeling framework to provide subject-specific prediction and correlation among the transition probabilities that cannot be obtained in Luo *et al.* [10].

While the previous works [10, 11] provided important model development and inference and presented interesting scientific findings, the transition probabilities among smoking states were assumed time-independent by including only the baseline covariates, and the modeling frameworks did not account for the dynamic correlation structure of the smoking and insomnia processes. This article proposes a modeling framework for the joint analysis of the longitudinal insomnia process and the stochastic smoking cessation process with a latent cured state (permanent quitting). We use a generalized linear mixed-effects model for the insomnia process and a stochastic mixed-effects model for the smoking process. The correlation between these two processes is modeled via latent random effects. We develop a Bayesian framework and Markov Chain Monte Carlo (MCMC) simulations for parameter estimation. The inclusion of time-dependent covariates allows the smoking transition probabilities, and the probability of insomnia varies at different visits and hence extends the functionality of the models proposed in the previous works [10, 11]. This model enhancement is important and useful in assisting policy making and intervention assessment. For example, we expect the smoking transition probabilities to change after an effective smoking cessation program. We can evaluate the effects of this program via the parameter corresponding to the indicator variable of attending the program. In addition, we can characterize the feedback of one process upon another by including the response of one process in modeling another process while accounting for other covariates. We have posted the R codes to simulate and analyze data at the Web Supplement[‡].

We organize the rest of the article as follows. Section 2 describes the joint model and the Bayesian inference procedure. Section 3 includes simulation studies to evaluate the performance of the joint model under various inter-process correlations. We apply the joint model to the ATBC study dataset in Section 4. Section 5 provides some concluding remarks.

2. The joint modeling framework

2.1. Exploring the correlation between two response variables

We can display the smoking and insomnia patterns in time plots as Figure 1, in which S and N denote smoking and nonsmoking intervals, respectively, and D and $\overline{D}$ denote insomnia and non-insomnia, respectively. We define a quit attempt as the nonsmoking interval immediately after smoking intervals, for example, the first, third, and sixth nonsmoking intervals in Figure 1. The second nonsmoking interval is not a new quit attempt because it does not follow a smoking interval. Similarly, we define a relapse

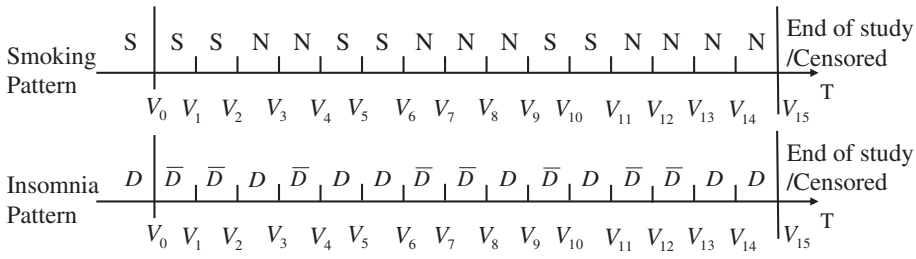

Figure 1. The smoking and insomnia patterns of one individual with S and N denoting smoking and nonsmoking, respectively, and D and $\overline{D}$ denoting insomnia and non-insomnia, respectively. The symbols before V_0 denote the baseline smoking and insomnia statuses.

[‡]*Supporting information may be found in the online version of this article.*

Table I. The log odds ratios and the p-values under different time lag values.

Lag	log OR	p
−4	−0.049	0.025
−3	−0.093	$1.26e-05$
−2	−0.138	$2.91e-11$
−1	−0.178	$1.50e-18$
0	−0.163	$1.42e-16$
1	−0.181	$1.83e-18$
2	−0.115	$8.22e-08$
3	−0.069	0.002
4	−0.052	0.003

OR, odds ratio.

to smoking as the smoking interval immediately after nonsmoking intervals, for example, the third and fifth smoking intervals in Figure 1.

Next, we explore the correlation between two time-varying variables smoking and insomnia. Let $y_{i1,t}$ (1 if smoke, 0 otherwise) and $y_{i2,t}$ (1 if insomnia, 0 otherwise) be the smoking and insomnia statuses of individual i ($i = 1, \ldots, m$, m is the total number of individuals) at visit t ($t = 0, \ldots, v_i$, where 0 is baseline visit and v_i is individual i's total number of follow-up visits), respectively. Let $\mathbf{y}_i$ denote individual i's outcome variable vector including both smoking and insomnia processes across all visits. We compute the correlation at each time lag k by using logarithm of odds ratio (OR) defined as $OR(k) = n_{00}^{(k)} n_{11}^{(k)} / (n_{01}^{(k)} n_{10}^{(k)})$, where $n_{ab}^{(k)} = \sum_{i=1}^{m} n_{iab}^{(k)}$ with $a, b = 0$ or 1, $n_{iab}^{(k)}$ is the total number of occurrences of $y_{i1,t-k} = a$ and $y_{i2,t} = b$ of individual i for $t = 0, \ldots, v_i + k$ if $k < 0$ and for $t = k, \ldots, v_i$ if $k \geqslant 0$. For example, the individual displayed in Figure 1 has $v_i = 15$, $n_{i00}^{(-1)} = 5$, $n_{i01}^{(-1)} = 4$, $n_{i10}^{(-1)} = 3$, $n_{i11}^{(-1)} = 3$, for time lag −1, and $n_{i00}^{(1)} = 4$, $n_{i01}^{(1)} = 4$, $n_{i10}^{(1)} = 4$, $n_{i11}^{(1)} = 3$, for time lag 1.

Table I displays the log OR and the p-values under different time lags k computed from 2849 individuals in the ATBC dataset who made at least one quit attempt and had at least one interval with insomnia symptom. It suggests that the correlation peaks at small lags, that is, −1, 0, and 1, and decreases as the time lag increases. The negative sign in log OR at negative lags indicates that insomnia at the previous visits is associated with nonsmoking at the current visit, while the negative sign at positive lags indicates that smoking at the previous visits is associated with non-insomnia at the current visit. Smoking and insomnia are strongly correlated under small lags as indicated by the extremely small p-values; for example, $p = 1.50e-18$ at lag −1 and $p = 1.83e-18$ at lag 1. Therefore, it is essential to consider the association between smoking and insomnia.

2.2. *The joint model*

This section first illustrates a three-state discrete-time stochastic process with $P_{ij,t}$, $j = 1, 2, 3$, denoting individual i's transition probabilities at visit t, as in Figure 2. This process distinguishes transient quitting state (temporarily nonsmoking with subsequent relapse) from permanent quitting state (lifelong smoke-free, latent state due to censoring) because the processes describing them are different and the identification and quantification of the risk factors associated with permanent quitting are more relevant to smoking cessation and public health. Because all individuals in the ATBC study were smokers at baseline, let the stochastic process start from the smoking state. When individual i is in the smoking state, he makes quit attempts at visit t with probability $P_{i1,t}$. Conditional on making a quit attempt at visit t, the individual may become a permanent quitter with probability $P_{i3,t+1}$ at visit $t+1$. With probability $1 - P_{i3,t+1}$, the individual enters the transient quitting state at visit $t+1$, from which he has probability $P_{i2,t+1}$ to relapse back to the smoking state at visit $t+1$. For example, the individual in Figure 1 makes a quit attempt at visit 2 ($t = 2$) with probability $P_{i1,2}$. With probability $1 - P_{i3,3}$, he enters transient quitting state at visit 3, from which he sustains at visit 3 with probability $1 - P_{i2,3}$, and relapses back to smoking at visit 4 with probability $P_{i2,4}$. Conditional on the transition probability $P_{ij,t}$, the transition to the next state is determined only by the current state and the previous state.

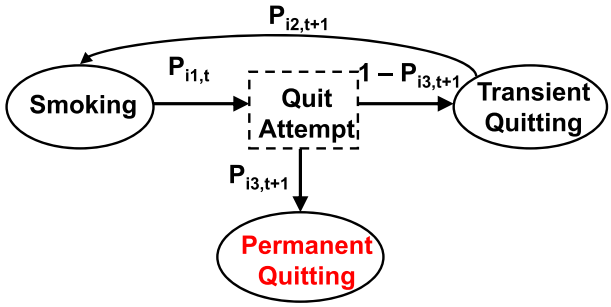

Figure 2. Transition among three states.

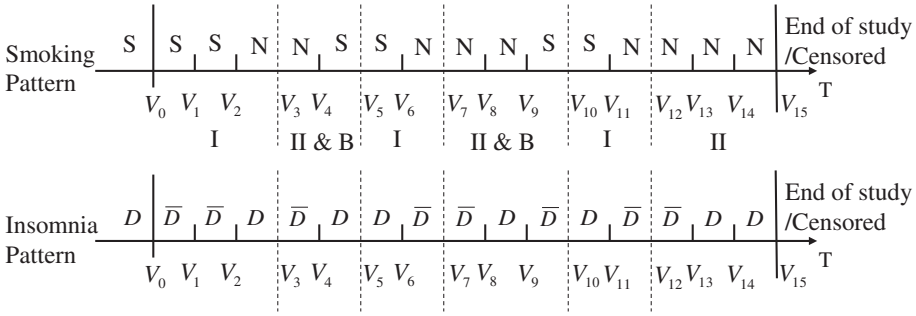

Figure 3. The partitioned smoking and insomnia patterns with I and II denoting type I and II geometric processes, respectively, and B denoting unsuccessful quit attempt.

We can describe this modeling structure by using two types of geometric processes corresponding to the sojourn time distributions in the smoking and nonsmoking states. The first type (Type I) of geometric process describes the number of smoking intervals before the next quit attempt. After a quit attempt is made, the individual becomes permanent quitter with probability $P_{i3,t+1}$. The second type (Type II) of geometric process models the number of nonsmoking intervals before the next relapse, conditional on being in a transient quitting state. Figure 3 displays the partition of the stochastic smoking pattern and the longitudinal insomnia pattern of the individual in Figure 1. Visits 1 to 3 are modeled as a Type I geometric processes (denoted by I). The individual has an unsuccessful quit attempt (denoted by B) at visit 3 and enters the transient quitting state, which lasts until visit 5 (denoted by II). Conditional on having a relapse at visit 4, the individual transitions again into a Type I process at visit 5. The modeling continues using the same rules.

We construct the likelihood of the smoking pattern for individual i (denoted by L_{i1}) by multiplying the likelihood contribution of both types of processes. For example, the likelihood of the smoking pattern for the individual in Figure 3 is

$$\begin{aligned} L_{i1} &= (1-P_{i1,0})(1-P_{i1,1})P_{i1,2}\cdot(1-P_{i3,3})(1-P_{i2,3})P_{i2,4}\cdot(1-P_{i1,5})P_{i1,6} \\ &\cdot (1-P_{i3,7})(1-P_{i2,7})(1-P_{i2,8})P_{i2,9}\cdot(1-P_{i1,10})P_{i1,11} \\ &\cdot \{(1-P_{i3,12})(1-P_{i2,12})(1-P_{i2,13})(1-P_{i2,14})+P_{i3,12}\}. \end{aligned}$$

The term $P_{i3,12}$ at the end accounts for the probability of being a permanent quitter at visit 12.

Let $P_{i4,t}$ be the probability of individual i having insomnia at visit t (referred to as the insomnia probability). Under conditional independence assumption (conditional on the random effect, u_{i4}, P_{i4,t_1}, and P_{i4,t_2} are independent for $t_1 \neq t_2$), we obtain the likelihood of the insomnia pattern for individual i (denoted by L_{i2}) by multiplying the insomnia probabilities at all visits. For example, the likelihood of

the insomnia pattern for the individual in Figure 3 is

$$\begin{aligned} L_{i2} &= (1-P_{i4,0})(1-P_{i4,1})P_{i4,2}\cdot(1-P_{i4,3})P_{i4,4}\cdot P_{i4,5}(1-P_{i4,6}) \\ &\cdot (1-P_{i4,7})P_{i4,8}(1-P_{i4,9})\cdot P_{i4,10}(1-P_{i4,11}) \\ &\cdot (1-P_{i4,12})P_{i4,13}P_{i4,14}. \end{aligned}$$

For notational ease, the probability vector is denoted by $\boldsymbol{P}_i = (\boldsymbol{P}'_{i1}, \boldsymbol{P}'_{i2}, \boldsymbol{P}'_{i3}, \boldsymbol{P}'_{i4})'$, where $\boldsymbol{P}_{ij} = (P_{ij,1}, \ldots, P_{ij,v_i})'$. The joint model for the smoking and insomnia processes has two sub-models.

$$\begin{aligned} &g_j(P_{ij,t}|\mathbf{x}_{ij,t}, y_{i2,t-1}, u_{ij}) = \mathbf{x}_{ij,t}\boldsymbol{\beta}_{j0} + \beta_{j1}y_{i2,t-1} + u_{ij} \quad \text{for } j = 1,2,3; \\ &g_4(P_{i4,t}|\mathbf{x}_{i4,t}, y_{i1,t-1}, u_{i4}) = \mathbf{x}_{i4,t}\boldsymbol{\beta}_{40} + \beta_{41}y_{i1,t-1} + u_{i4}, \end{aligned} \tag{1}$$

where the vectors $\mathbf{x}_{ij,t}$ and $\mathbf{x}_{i4,t}$ are covariate vectors that may include time-dependent covariates and can share part of or all the covariates, $y_{i1,t-1}$ and $y_{i2,t-1}$ are the smoking and insomnia statuses at visit $t-1$, respectively, u_{ij} and u_{i4} are random effects, and $g(\cdot)$ are link functions. Let $g_1(\cdot)$ and $g_2(\cdot)$ be the complementary log–log link function and $g_3(\cdot)$ and $g_4(\cdot)$ be the logit link function. We use the complementary log–log link function to make the transition probabilities between smoking and transient quitting states analogous to hazard functions in a discrete-time proportional hazards model [12].

To model the feedback effect, we include a single lagged covariate with the lag value being one [13]. Specifically, β_{j1} denotes the feedback effect of the insomnia symptom at visit $t-1$ ($y_{i2,t-1}$) on the smoking transition probability $P_{ij,t}$ at visit t conditional on the covariate vector $\boldsymbol{x}_{ij,t}$ and the random effect u_{ij}. Similarly, β_{41} represents the feedback effect of smoking at visit $t-1$ ($y_{i1,t-1}$) on the insomnia probability at visit t conditional on the covariate vector $\boldsymbol{x}_{i4,t}$ and the random effect u_{i4}. For notational ease, the coefficient vector is denoted by $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \boldsymbol{\beta}'_3, \boldsymbol{\beta}'_4)'$, where $\boldsymbol{\beta}'_j = (\boldsymbol{\beta}'_{j0}, \beta_{j1})$ for $j = 1,2,3$, and $\boldsymbol{\beta}'_4 = (\boldsymbol{\beta}'_{40}, \beta_{41})$. For individual i, let $\mathbf{x}_i$ denote the covariate information, and let the multivariate random effect vector be $\mathbf{u}_i = (u_{i1}, u_{i2}, u_{i3}, u_{i4})'$.

We link the two sub-models in (1) via the random effect vector $\mathbf{u}_i$, which is assumed to be independent and identically distributed with normal probability density function $\mathbf{u}_i|\boldsymbol{\Sigma} \sim \mathrm{N}_4(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a 4×4 covariance matrix with the (l, m)th entry denoted by σ_{lm}. As pointed out by Molenberghs and Verbeke [14, Chap. 25.2], a special case of this model specification for random effects is the shared parameter model, which assumes the same set of random effects for both smoking and insomnia outcomes in this context. While the shared parameter model has relatively lower dimension of the random effects distribution when compared to the previous model, it is based on much stronger assumptions about the association between outcomes, which is difficult to validate in this application.

The joint modeling framework has accounted for three sources of correlation, that is, intra-process correlation (measurements from the same process at different visits), inter-process correlation (measurements from different processes at the same visit), and cross-process correlation (measurement from different processes at different visits). The intra-process correlation is modeled by the process-specific random effect $\mathbf{u}_i$. The inter-process correlation is modeled by the association between u_{i1}, u_{i2}, u_{i3}, and u_{i4} through three covariance parameters σ_{41}, σ_{42}, and σ_{43}. If the covariance parameters are significantly different from zero, it indicates the existence of the inter-process correlation. Finally, the cross-process correlation is modeled by the single lagged covariates $y_{i1,t-1}$ and $y_{i2,t-1}$, as well as three covariance parameters σ_{41}, σ_{42}, and σ_{43}.

It is assumed that the smoking process is independent of the insomnia process, conditional on the covariates and the random effect vector $\mathbf{u}_i$. The observed likelihood conditional on $\mathbf{u}_i$ for individual i is $L(\boldsymbol{\Phi}; \mathbf{u}_i, \mathbf{y}_i) = L_{i1}L_{i2}$, where the parameter vector of interests $\boldsymbol{\Phi} = \{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$. The marginal likelihood is $L(\boldsymbol{\Phi}; \mathbf{y}_i) = \int L(\boldsymbol{\Phi}; \mathbf{u}_i, \mathbf{y}_i)h(\mathbf{u}_i; \boldsymbol{\Sigma})d\mathbf{u}_i$, where $h(\mathbf{u}_i; \boldsymbol{\Sigma})$ is $\mathrm{N}_4(\mathbf{0}, \boldsymbol{\Sigma})$. Because this integral cannot be evaluated analytically, we can obtain the samples of the parameter vector $\boldsymbol{\Phi}$ by using the Bayesian inference framework via MCMC simulations introduced in Section 2.3.

2.3. *Bayesian inference*

This section proposes a Bayesian approach for the model inference. We use noninformative priors for the parameter vectors. Each component in the coefficient vector $\boldsymbol{\beta}$ is independently assigned normal $\mathrm{N}(0, 100)$ prior distribution. For the ease of sampling for $\boldsymbol{\Sigma}$, we use an approach based on the Cholesky decomposition [15]. Let $\boldsymbol{\Sigma} = \boldsymbol{\Omega}\boldsymbol{\Omega}'$, where $\boldsymbol{\Omega}$ is a lower triangular matrix with ω_{lm} being the (l,m)th entry for $1 \leqslant m \leqslant l \leqslant 4$ and zero entries above the main diagonal. Consider a latent vector $\mathbf{z}_i = (z_{i1}, \ldots, z_{i4})'$ with $N(0,1)$ independent components. The linear reparameterization of $\mathbf{u}_i = \boldsymbol{\Omega}\mathbf{z}_i$

(with element being $u_{ij} = \sum_{l=1}^{j} \omega_{jl} z_{il}$, e.g., $u_{i2} = \omega_{21} z_{i1} + \omega_{22} z_{i2}$) has mean zero and variance $\boldsymbol{\Sigma}$, whose entries are $\sigma_{jk} = \sum_{l=1}^{j \wedge k} \omega_{jl} \omega_{kl}$, $1 \leqslant j,k \leqslant 4$, where $j \wedge k = \min(j,k)$. We impose uniform(0, 10) prior distribution on ω_{ll} to ensure non-negativity and N(0, 100) prior distribution on ω_{lm} when $l \neq m$ to allow for possible negative correlation. For notational ease, let vectors $\boldsymbol{\sigma}$ and $\boldsymbol{\omega}$ denote the entries in the lower triangular part of the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$, respectively, and let vector $\boldsymbol{\rho} = (\rho_{21}, \rho_{31}, \rho_{32}, \rho_{41}, \rho_{42}, \rho_{43})$ denote the pairwise correlation coefficients among the components of the random effects vector $\mathbf{u}_i$.

The joint distribution of the data and parameters is

$$P(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_{i=1}^{m} \left[L_{i1} L_{i2} \left\{ \prod_{j=1}^{4} p(\boldsymbol{P}_{ij}; \boldsymbol{\beta}_j, \boldsymbol{\omega}, \mathbf{z}_i) P(\mathbf{z}_i) \right\} \right] P(\boldsymbol{\beta}) P(\boldsymbol{\omega}), \tag{2}$$

where $P(\boldsymbol{\beta})$, and $P(\boldsymbol{\omega})$ are the prior distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$, respectively. The full conditional distributions are derived and the parameters are sampled component-wise using a random walk Metropolis–Hastings algorithm in the following order $(\boldsymbol{\beta}_1, \omega_{11})$, $(\boldsymbol{\beta}_2, \omega_{21}, \omega_{22})$, $(\boldsymbol{\beta}_3, \omega_{31}, \omega_{32}, \omega_{33})$, $(\boldsymbol{\beta}_4, \omega_{41}, \omega_{42}, \omega_{43}, \omega_{44})$, and $\mathbf{z}_i$. We compute the posterior distributions of $\boldsymbol{\sigma}$ and $\boldsymbol{\rho}$ from the posterior samples of $\boldsymbol{\omega}$. For statistical inference, we compute the posterior means, standard deviations, and 95% equal-tail credible intervals (i.e., the intervals from 2.5 and 97.5 percentiles of the posterior distributions).

To assess the convergence of the MCMC chains, we use the trace plots and view the absence of apparent trend in the plots as evidence of convergence. In addition, we run multiple chains with overdispersed initial values and compute the Gelman–Rubin scale reduction statistics $\widehat{R}$ ensure $\widehat{R}$ of all parameters are smaller than 1.1 [16]. We assess the length of the burn-in by trace plots and autocorrelation for each parameter.

3. Simulation studies

In this section, we conduct two simulation studies to compare the performance of the proposed joint model and a separate model, that is, separately fitting a three-state stochastic process model for the smoking pattern and a generalized linear mixed model (GLMM) for the longitudinal insomnia process. In the first simulation study, there is no inter-process correlation (i.e., $\sigma_{41}, \sigma_{42}, \sigma_{43} = 0$), while in the second simulation study, there exists large inter-process correlation. In both simulation studies, we generate 500 datasets with sample size $m = 10,000$ and with data structure similar to the ATBC dataset. We consider the case where the smoking transition probabilities only depend on the insomnia status at the last visit and the insomnia probability only depends on the smoking status at the last visit. We generate no missing data. We generate the smoking and insomnia processes by using the following algorithm.

(1) For individual i, simulate the total visit number from a normal distribution with mean 14.2 and standard deviation 6.3 because it resembles the distribution of the number of follow-up visits in the ATBC study. Round the total visit number to the closest integer if it is larger than 1 and round it to 1 if it is smaller than 1.

(2) Simulate the random effects vector $\mathbf{u}_i$ from multivariate normal distribution with mean 0 and covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.09 & -0.01 & -0.12 & 0 \\ -0.01 & 0.16 & 0.05 & 0 \\ -0.12 & 0.05 & 0.25 & 0 \\ 0 & 0 & 0 & 0.36 \end{pmatrix}.$$

for the first simulation study. The correlation coefficients among the components of $\mathbf{u}_i$ are $(\rho_{21}, \rho_{31}, \rho_{32}, \rho_{41}, \rho_{42}, \rho_{43}) = (-0.083, -0.8, 0.25, 0, 0, 0)$. In the second simulation study, let $(\sigma_{41}, \sigma_{42}, \sigma_{43}) = (-0.05, -0.04, 0.05)$, which gives inter-process correlation coefficients $(\rho_{41}, \rho_{42}, \rho_{43}) = (-0.28, -0.17, 0.17)$.

(3) Simulate the baseline insomnia status from a Bernoulli distribution with probability 0.2 because the prevalence of baseline insomnia symptom is around 20%. We compute the probability $\boldsymbol{P}_{ij}$ for $j = 1, 2, 3$ and $\boldsymbol{P}_{i4}$ at the first visit from model (1) with $\boldsymbol{\beta}_1 = (0.186, -1.217)'$, $\boldsymbol{\beta}_2 = (-1.031, 1.217)'$, $\boldsymbol{\beta}_3 = (0.405, -2.603)'$, and $\boldsymbol{\beta}_4 = (-2, 1)'$. Let $y_{i1,0} = 1$ because every individual is a smoker at baseline in the ATBC study.

(4) Conditional on smoking at visit $t-1$, simulate the insomnia status at visit t from a Bernoulli distribution with probability $P_{i4,t}$, and simulate the smoking status at visit t from a Bernoulli distribution with probability $P_{i1,t}$.
(5) Conditional on making a quit attempt at visit t, simulate the quitting status as follows.
   (a) With probability $P_{i3,t+1}$, the individual becomes a permanent quitter, and all the remaining visits are nonsmoking. Simulate the insomnia status at the remaining visits with probability $P_{i4,t+1}$.
   (b) With probability $1-P_{i3,t+1}$, the individual becomes a transient quitter. The smoking and insomnia statuses at visit $t+1$ are simulated with probability $P_{i2,t+1}$ and $P_{i4,t+1}$, respectively.
(6) Compute $P_{ij,t+1}$ for $j=1,2,3$ and $P_{i4,t+1}$ at visit $t+1$ conditional on the smoking and insomnia statuses at visit t.
(7) Repeat Steps 4, 5, and 6 until a smoking pattern and an insomnia pattern are generated for each individual.

We apply the Bayesian framework in Section 2.3 to obtain samples from the posterior distributions of the parameters of interest. For each dataset in both simulation studies, we run three parallel chains with overdispersed initial values. We run each chain for 50,000 iterations, discard the first 20,000 iterations as a burn-in, and use the next 30,000 samples to calculate the joint posterior distribution of the parameters of interest.

We compare the results of the separate model and the joint model of the first simulation study with no inter-process correlation in Table II. In this table, we label the average of the posterior means minus the true values as bias, the square root of the average of the variances as SE, the standard deviation of the posterior means as SD, the coverage probabilities of 95% equal-tail credible intervals (CI) as CP, and the square root of the average of the squares of the bias as root mean square error (RMSE). The results suggest that two methods generate comparable results; that is, the bias is negligible, SE is close to SD, the credible interval coverage probabilities are reasonably close to 95%, and RMSE is comparable. The estimates of σ_{41}, σ_{42}, and σ_{43} from the joint model are correctly close to zero although the standard errors of σ_{42} and σ_{43} are slightly underestimated, which leads to conservative credible intervals and the coverage probability being smaller than the nominal value.

Table III displays the results of the second simulation study with large inter-process correlation. The results from the joint model indicate that the estimates of all parameters, including the inter-process

Table II. Bias, standard error (SE), standard deviation (SD), and coverage probabilities (CP) of 95% credible intervals, for the separate model and the joint model, when there is no inter-process correlation.

	Separate model					Joint model				
Parameter	Bias	SE	SD	CP	RMSE	Bias	SE	SD	CP	RMSE
$\beta_{10}=0.186$	−0.001	0.010	0.010	0.958	0.011	−0.003	0.010	0.011	0.940	0.011
$\beta_{11}=-1.217$	0.000	0.024	0.023	0.970	0.023	0.001	0.024	0.023	0.968	0.023
$\beta_{20}=-1.031$	0.001	0.026	0.025	0.946	0.025	−0.002	0.027	0.028	0.928	0.028
$\beta_{21}=1.217$	−0.002	0.024	0.026	0.920	0.026	−0.002	0.026	0.026	0.944	0.026
$\beta_{30}=0.405$	0.006	0.026	0.025	0.976	0.025	−0.001	0.025	0.024	0.970	0.024
$\beta_{31}=-2.603$	−0.013	0.063	0.069	0.922	0.070	0.006	0.068	0.070	0.914	0.070
$\beta_{40}=-2.000$	−0.001	0.012	0.013	0.930	0.013	−0.001	0.012	0.013	0.922	0.013
$\beta_{41}=1.000$	0.000	0.016	0.015	0.960	0.015	0.000	0.017	0.016	0.938	0.016
$\sigma_{11}=0.090$	0.000	0.010	0.009	0.940	0.009	−0.002	0.009	0.010	0.916	0.011
$\sigma_{21}=-0.010$	0.000	0.011	0.012	0.906	0.012	0.000	0.011	0.013	0.910	0.013
$\sigma_{22}=0.160$	0.003	0.023	0.022	0.944	0.023	−0.002	0.024	0.026	0.924	0.026
$\sigma_{31}=-0.120$	0.001	0.015	0.016	0.924	0.016	−0.004	0.016	0.017	0.924	0.018
$\sigma_{32}=0.050$	0.002	0.033	0.031	0.948	0.031	−0.006	0.034	0.036	0.922	0.036
$\sigma_{33}=0.250$	0.024	0.063	0.067	0.944	0.071	−0.014	0.062	0.063	0.928	0.065
$\sigma_{44}=0.360$	0.002	0.014	0.013	0.960	0.013	0.002	0.014	0.013	0.950	0.013
$\sigma_{41}=0.000$						−0.001	0.008	0.008	0.914	0.008
$\sigma_{42}=0.000$						−0.001	0.013	0.015	0.880	0.015
$\sigma_{43}=0.000$						0.001	0.016	0.019	0.880	0.019

RMSE, root mean square error.

Table III. Bias, standard error (SE), standard deviation (SD), and coverage probabilities (CP) of 95% credible intervals, for the separate model and the joint model, when there is sizeable inter-process correlation.

	Separate model					Joint model				
Parameter	Bias	SE	SD	CP	RMSE	Bias	SE	SD	CP	RMSE
$\beta_{10} = 0.186$	−0.002	0.010	0.009	0.948	0.010	−0.001	0.010	0.010	0.958	0.010
$\beta_{11} = -1.217$	**−0.030**	0.024	0.025	**0.766**	0.039	−0.006	0.025	0.026	0.926	0.027
$\beta_{20} = -1.031$	−0.002	0.026	0.024	0.950	0.024	−0.006	0.031	0.033	0.922	0.034
$\beta_{21} = 1.217$	**−0.028**	0.024	0.022	**0.848**	0.036	−0.004	0.026	0.027	0.926	0.027
$\beta_{30} = 0.405$	−0.007	0.026	0.026	0.948	0.027	−0.004	0.025	0.026	0.934	0.027
$\beta_{31} = -2.603$	**0.032**	0.063	0.063	**0.886**	0.070	0.002	0.064	0.066	0.938	0.066
$\beta_{40} = -2.000$	−0.001	0.011	0.012	0.930	0.012	0.000	0.012	0.012	0.922	0.012
$\beta_{41} = 1.000$	0.006	0.016	0.016	0.910	0.017	0.000	0.017	0.017	0.952	0.017
$\sigma_{11} = 0.090$	−0.002	0.009	0.010	0.924	0.010	−0.003	0.010	0.011	0.922	0.011
$\sigma_{21} = -0.010$	−0.003	0.011	0.011	0.908	0.012	0.002	0.020	0.022	0.932	0.023
$\sigma_{22} = 0.160$	0.005	0.023	0.024	0.950	0.024	−0.009	0.040	0.044	0.942	0.044
$\sigma_{31} = -0.120$	0.002	0.015	0.016	0.904	0.016	0.003	0.017	0.019	0.924	0.019
$\sigma_{32} = 0.050$	−0.003	0.032	0.035	0.920	0.035	−0.007	0.044	0.045	0.924	0.046
$\sigma_{33} = 0.250$	0.010	0.060	0.059	0.960	0.059	−0.013	0.061	0.063	0.944	0.064
$\sigma_{44} = 0.360$	0.001	0.014	0.015	0.910	0.015	0.002	0.014	0.015	0.916	0.015
$\sigma_{41} = -0.050$						0.001	0.008	0.010	0.930	0.010
$\sigma_{42} = -0.040$						0.006	0.016	0.019	0.918	0.020
$\sigma_{43} = 0.050$						0.003	0.017	0.019	0.944	0.019

Note: We highlight large bias and poor CP in boldface.

correlation coefficients, have negligible bias, SE being close to SD. The coverage probabilities of 95% credible intervals are all reasonably around the nominal value. In contrast, the separate model gives biased estimates, low coverage probabilities, and larger RMSE for the insomnia effect in modeling the smoking transition probabilities (β_{11}, β_{21}, and β_{31}, shown in boldface), due to ignoring the inter-process correlation, and the consequent information loss. There is no apparent difference in the estimation of the longitudinal insomnia process comparing the separate model to the joint model.

From the simulation studies, the conclusion is that the joint model provides results comparable to the separate model when there is no inter-process correlation, while it provides more accurate estimates for the smoking process than the separate model when the inter-process correlation is large.

4. Application to the Alpha-Tocopherol, Beta-Carotene study

In this section, we apply the proposed joint model and the Bayesian inference framework to the motivating ATBC dataset. For all the results in this section, we use three parallel chains with overdispersed initial values, and we run each chain for 150, 000 iterations. We discard the first 50, 000 iterations as burn-in, and the inference is based on the remaining 100, 000 iterations. We compare the results from the separate model and from the joint model.

We fit models with the following covariates: smoking or insomnia status at the last visit, and baseline covariates including age, years of smoking, cigarettes per day, alcohol consumption (g/day), and inhalation (yes/no). Table IV shows the estimation results with a negative sign indicating a smaller probability of having a certain event. It is observed that the joint model and the separate model give different estimates (highlighted in boldface) for the insomnia effect in modeling the smoking transition probabilities, although the same set of parameters are identified for significance by both models. For example, conditional on the random effect u_{i1}, both models indicate that individuals with insomnia at the last visit have higher probability to make quit attempts than those without insomnia. In addition, conditional on u_{i2} and u_{i3}, the joint model results suggest that insomnia at the last visit is associated with higher probability of relapse and permanent quitting given the quit attempts, while the separate model results suggest the association in opposite direction. The differences between the results from the joint model and the separate model might be explained by the significant high negative inter-process correlation coefficients

Table IV. Results of fitting the separate model and the joint model with six covariates in the Alpha-Tocopherol, Beta-Carotene dataset.

		Separate model			Joint model		
Models	Parameters	Mean_{SD}	95% CI		Mean_{SD}	95% CI	
	Intercept	$-4.409_{0.026}$	−4.463	−4.359	$-4.425_{0.027}$	−4.479	−4.372
	Insomnia*	$\mathbf{0.210_{0.036}}$	**0.138**	**0.281**	$\mathbf{0.281_{0.043}}$	**0.198**	**0.362**
	Age*	$0.195_{0.017}$	0.162	0.228	$0.195_{0.016}$	0.163	0.228
$\boldsymbol{P}_{i1}$	Years smoked*	$-0.271_{0.015}$	−0.301	−0.242	$-0.274_{0.015}$	−0.303	−0.246
	Cigarette/day*	$-0.295_{0.016}$	−0.326	−0.264	$-0.295_{0.016}$	−0.329	−0.266
	Alcohol*	$-0.199_{0.018}$	−0.234	−0.162	$-0.202_{0.018}$	−0.238	−0.165
	Inhale	$0.006_{0.029}$	−0.050	0.062	$0.008_{0.030}$	−0.047	0.067
	Intercept	$-0.550_{0.214}$	−0.958	−0.121	$-0.440_{0.235}$	−0.866	0.018
	Insomnia	$\mathbf{-0.014_{0.091}}$	**−0.192**	**0.163**	$\mathbf{0.123_{0.094}}$	**−0.061**	**0.305**
	Age	$0.008_{0.058}$	−0.102	0.124	$-0.007_{0.058}$	−0.120	0.105
$\boldsymbol{P}_{i2}$	Years smoked	$-0.030_{0.050}$	−0.128	0.069	$-0.014_{0.051}$	−0.108	0.087
	Cigarette/day*	$-0.144_{0.054}$	−0.250	−0.040	$-0.140_{0.051}$	−0.239	−0.035
	Alcohol	$0.132_{0.068}$	−0.006	0.265	$0.100_{0.067}$	−0.032	0.240
	Inhale	$0.050_{0.097}$	−0.141	0.241	$0.025_{0.095}$	−0.165	0.211
	Intercept	$2.611_{0.214}$	2.214	3.048	$2.719_{0.244}$	2.284	3.224
	Insomnia	$\mathbf{-0.262_{0.146}}$	**−0.549**	**0.022**	$\mathbf{0.061_{0.189}}$	**−0.317**	**0.431**
	Age	$0.071_{0.066}$	−0.059	0.200	$0.053_{0.071}$	−0.085	0.193
$\boldsymbol{P}_{i3}$	Years smoked*	$0.132_{0.058}$	0.022	0.248	$0.150_{0.062}$	0.026	0.275
	Cigarette/day	$0.033_{0.065}$	−0.095	0.159	$0.033_{0.063}$	−0.091	0.160
	Alcohol	$0.003_{0.073}$	−0.141	0.147	$-0.026_{0.077}$	−0.176	0.133
	Inhale	$-0.024_{0.115}$	−0.251	0.204	$-0.054_{0.124}$	−0.300	0.185
	Intercept	$-3.798_{0.040}$	−3.871	−3.713	$-3.772_{0.044}$	−3.856	−3.685
	Smoking*	$-0.353_{0.028}$	−0.410	−0.300	$-0.390_{0.032}$	−0.455	−0.335
	Age*	$0.191_{0.027}$	0.131	0.252	$0.179_{0.031}$	0.123	0.251
$\boldsymbol{P}_{i4}$	Years smoked	$0.019_{0.030}$	−0.042	0.076	$0.035_{0.034}$	−0.034	0.113
	Cigarette/day*	$0.125_{0.024}$	0.087	0.180	$0.119_{0.025}$	0.073	0.176
	Alcohol*	$0.344_{0.021}$	0.304	0.393	$0.360_{0.028}$	0.306	0.408
	Inhale*	$0.125_{0.058}$	0.009	0.253	$0.132_{0.048}$	0.039	0.225
	ρ_{21}	$-0.125_{0.112}$	−0.340	0.109	$-0.148_{0.124}$	−0.380	0.111
	ρ_{31}	$-0.962_{0.022}$	−0.994	−0.909	$-0.920_{0.026}$	−0.963	−0.863
$\boldsymbol{\rho}$	ρ_{32}	$0.354_{0.141}$	0.067	0.607	$0.459_{0.135}$	0.181	0.690
	ρ_{41}				$-0.051_{0.019}$	−0.081	−0.015
	ρ_{42}				$-0.274_{0.028}$	−0.339	−0.231
	ρ_{43}				$-0.141_{0.032}$	−0.205	−0.077

Note: Entries in boldface indicate different results from the two models.
*Represents statistical significance.

($\widehat{\rho}_{41} = -0.051$, $\widehat{\rho}_{42} = -0.274$, and $\widehat{\rho}_{43} = -0.141$). With the help of jointly modeling the correlated stochastic smoking process and the longitudinal insomnia process, we expect the joint model to improve the estimation of the parameters of insomnia effects, as demonstrated in Section 3.

The joint model and the separate model produce similar results for other parameters in terms of means, standard deviations, and 95% CIs and identified similar set of significant covariates. The rows labeled $\boldsymbol{P}_{i1}$ in Table IV display the results of modeling the probability of making quit attempts at a given visit. We conclude that conditional on the random effect u_{i1}, individuals with insomnia at the last visit or older individuals are more likely to make quit attempts, while years of smoking, cigarettes per day, and alcohol consumption are negatively associated with the probability of making quit attempts. The rows labeled $\boldsymbol{P}_{i2}$ in Table IV display the results of modeling the probability of relapsing at a given visit for individuals in the transient quitting stage. It suggests that conditional on the random effect u_{i2}, individuals who smoke more cigarettes per day are less likely to relapse once they make quit attempts. This unexpected results have been identified and reported in the previous works [10, 11]. The rows labeled $\boldsymbol{P}_{i3}$ in Table IV display the results of modeling the probability of permanent quitting at a certain visit.

Conditional on the random effect u_{i3}, individuals with longer smoking history are more likely to be permanent quitter once quit attempt are made; that is, the OR of permanent quitting for an increase of 8.4 years of smoking history (i.e., one standard deviation) is 1.162 (95% CI: [1.026, 1.317]), holding other covariates fixed. The rows labeled $\boldsymbol{P}_{i4}$ in Table IV display the results of modeling the probability of insomnia at a certain visit. Conditional on the random effect u_{i4}, smoking at the last visit is negatively associated with the insomnia probability, while age, years of smoking, cigarettes per day, alcohol consumption, and inhalation show positive association.

The data analysis results suggest the existence of a feedback system. First, conditional on the random effect u_{i1}, the complement probability of making quit attempts for individuals with insomnia at the last visit is the complement probability for those without insomnia raised to the power 1.324 (95% CI: [1.219, 1.436]). Moreover, conditional on the random effect u_{i4}, the OR of having insomnia for individuals who did not smoke at the last visit is 1.477 (95% CI: [1.398, 1.576]), compared with those who smoked. Hence, insomnia increases the likelihood of making quit attempts, which further increases the risk of future insomnia in a feedback cycle. These results of the feedback system are consistent with the negative smoking and insomnia correlations displayed in Table I.

Our model identifies a high negative correlation between $\boldsymbol{P}_{i1}$ and $\boldsymbol{P}_{i3}$ (ρ_{31}), and a relative high positive correlation between $\boldsymbol{P}_{i2}$ and $\boldsymbol{P}_{i3}$ (ρ_{32}). We now provide some insight about these high correlations. Consider $\widehat{\rho}_{31}$ first. There are 1501 long-term sustainers (individuals who sustained at least 40 months until censoring) who are more likely to be permanent quitters and hence have high $\boldsymbol{P}_{i3}$. Among them, 1453 (96.8%) made only one quit attempt. The association of high $\boldsymbol{P}_{i3}$ (long trailing nonsmoking intervals) with small $\boldsymbol{P}_{i1}$ (only one quit attempt) leads to high negative ρ_{31}. Consider $\widehat{\rho}_{32}$ next. The 1115 relapsers (individuals who made at least one quit attempt but did not sustain until censoring) had an average smoke-free interval of 2.56 visits (10.2 months) before next relapse. The association of small $\boldsymbol{P}_{i3}$ (relapse frequently with not trailing nonsmoking interval) and small $\boldsymbol{P}_{i2}$ (long smoke-free interval) leads to high positive ρ_{32}.

Table IV displays strong correlation between the stochastic smoking process and the longitudinal insomnia process, for example, high negative correlation between $\boldsymbol{P}_{i1}$ and $\boldsymbol{P}_{i4}$ (ρ_{41}), between $\boldsymbol{P}_{i2}$ and $\boldsymbol{P}_{i4}$ (ρ_{42}), and between $\boldsymbol{P}_{i3}$ and $\boldsymbol{P}_{i4}$ (ρ_{43}). Here, we provide some insight into this interesting phenomenon. Let us first consider $\widehat{\rho}_{41}$. There are 6034 ever quitters (individuals who made at least one quit attempt) and 20, 181 never quitters (individuals who never made any quit attempts). In our model, the ever quitters are more likely to have larger probabilities of making quit attempts. The empirical estimate of probability of insomnia is smaller among ever quitters than among never quitters (i.e., mean: 0.131 vs. 0.144; $p < 0.001$). The association of larger probabilities of making quit attempts and smaller probabilities of insomnia indicates negative correlation of ρ_{41}. Next, we consider $\widehat{\rho}_{42}$. There are 15, 757 non-insomnia individuals (individuals who never had insomnia) and 10, 458 insomnia individuals (individuals who had insomnia at least one visit). In our model, non-insomnia individuals are more likely to have smaller probabilities of insomnia than the insomnia individuals. Among them, there are 3495 and 2539 individuals who made at least one quit attempt, respectively. The non-insomnia individuals have shorter smoke-free intervals before relapse than the insomnia individuals (i.e., 0.6 vs. 2.2 months; $p < 0.001$). The association of smaller probabilities of insomnia and higher relapse probabilities $\boldsymbol{P}_{i2}$ (shorter smoke-free intervals) indicates negative correlation of ρ_{42}. At last, we consider $\widehat{\rho}_{43}$. There are 1501 long-term sustainers and 1115 relapsers. In our model, the long-term sustainers are more likely to have higher permanent quitting probabilities than the relapsers. The empirical estimate of probability of insomnia is smaller among long-term sustainers than relapsers (i.e., mean: 0.124 vs. 0.140; $p = 0.10$). The association of higher permanent quitting probabilities with smaller probabilities of insomnia indicates negative correlation.

5. Discussion

In this article, we propose a joint model and a Bayesian approach to analyze the longitudinal insomnia process and the stochastic smoking process with a latent cure state. By combining the information from the longitudinal data, the joint model improves the accuracy of the parameter estimates compared with the separate model and provides similar precision, when strong inter-process correlation exists. On the other hand, the joint model produces comparable results to the separate model when there is no inter-process correlation. Our joint model extends the functionality of the modeling framework in Luo *et al.* [11] by including time-dependent covariates and by accounting for the correlation between the

subject-specific smoking transition probabilities and the insomnia probability. Consequently, we identify significant negative correlation between the smoking and insomnia processes. An important but previously unknown finding is the existence of a feedback system between insomnia and smoking; for example, insomnia at the last visit increases the likelihood of making quit attempts at the current visit, which further increases the risk of future insomnia in a feedback cycle. In addition, insomnia at the last visit has shown significant positive association with the probability of making quit attempts but insignificant positive association with the probabilities of relapse and permanent quitting given the quit attempts.

The proposed joint modeling framework is attractive in several respects. First, the joint model provides correction of potential biases in the separate model when the insomnia and smoking processes are strongly correlated. Second, the joint model accounts for and provides insight into the within-subject correlation between the insomnia and the smoking processes. Third, we develop a method to formulate and calculate the likelihood function involving time-dependent covariates. To the best of our knowledge, this article is the first one to propose a joint model for a stochastic process and a longitudinal outcome with time-dependent covariates. Computationally, the proposed Bayesian inference method can account for high-dimensional random effects, and it also allows incorporation of prior information.

The proposed joint model is flexible enough to address many questions of scientific interest. For example, if it is of interest to jointly model more longitudinal measurements of diseases with the smoking process, we could expand $P_{i4,t}$ in model (1) to a vector of probabilities with each component representing the probability of the presence of each disease. Additionally, we can incorporate more time-dependent covariates (e.g., the participation of a smoking cessation program or the increase of cigarette tax) into the model to estimate the effects of these covariates.

The smoking and insomnia information in the ATBC dataset is based on 4-month interval, and visit-to-visit transitions of smoking status are modeled while some recent articles on the analysis of smoking cessation data modeled the smoking transition in a more continuous manner [17, 18]. One limitation of the proposed model is that the cross-process correlation is modeled by the single lagged covariates and the covariance parameters σ_{41}, σ_{42}, and σ_{43}. It is difficult to distinguish the contribution of each source. We will address this issue in our future research. Another issue is the normality assumption of random effects in our joint model. Some researchers [19, 20] have reported that the statistical inference is generally robust to the departure from the normality assumption. It is of interest to investigate our joint model's performance when the underlying random effects distribution is symmetric non-normal or even asymmetric. Moreover, we assume the random effects covariance matrix to be homogeneous (same for all individuals). However, the covariance matrix may depend on subject-specific characteristics and is thus heterogeneous. Ignoring the heterogeneity can result in biased estimates [21, 22]. As a future direction, we would address the issue of accounting for heterogeneity in the covariance matrix in the proposed joint modeling framework.

Acknowledgements

Sheng Luo's research was supported in part by NIH/NINDS grants U01NS043127 and U01NS43128, and NIH/NCATS grant KL2 TR000370. The authors are grateful to Dr. Nilanjan Chatterjee for access to the dataset and helpful discussion and to Drs. Thomas A. Louis, Ciprian M. Crainiceanu, and Wenyaw Chan for insightful comments and suggestions. We performed the computations on the high-performance computational capabilities of the Linux cluster system at University of Texas School of Public Health (UTSPH). The author expresses appreciation to UTSPH information technology staff for their technical support of the cluster.

References

1. Bixler EO, Kales A, Soldatos CR, Kales JD, Healey S. Prevalence of sleep disorders in the Los Angeles metropolitan area. *The American Journal of Psychiatry* 1979; **136**:1257–1262.
2. Mellinger GD, Balter MB, Uhlenhuth EH. Insomnia and its treatment. Prevalence and correlates. *Archives of General Psychiatry* 1985; **42**:225–232.
3. Ford DE, Kamerow DB. Epidemiologic study of sleep disturbances and psychiatric disorders. An opportunity for prevention?. *The Journal of American Medical Association* 1989; **262**:1479–1484.
4. Prochaska JO, DiClemente CC. Stages and processes of self-change of smoking: toward an integrative model of change. *Journal of Consulting and Clinical Psychology* 1983; **31**:390–395.
5. Wetter DW, Young TB. The relation between cigarette smoking and sleep disturbance. *Preventive Medicine* 1994; **23**:328–334.

6. Phillips B, Mannino DM. Do insomnia complaints cause hypertension or cardiovascular disease?. *Journal of Clinical Sleep Medicine* 2007; **3**:489–94.
7. Hughes JR. Effects of abstinence from tobacco: valid symptoms and time course. *Nicotine & Tobacco Research* 2007; **9**:315–327.
8. Zeger SL, Liang KY. Feedback models for discrete and continuous time series. *Statistica Sinica* 1991; **1**:51–64.
9. Group AS. Incidence of cancer and mortality following α-tocopherol and β-carotene supplementation. *Journal of American Medical Association* 2003; **290**(4):476–485.
10. Luo S, Crainiceanu CM, Louis TA, Chatterjee N. Analysis of smoking cessation patterns using a stochastic mixed-effects model with a latent cured state. *Journal of the American Statistical Association* 2008; **103**:1002–13.
11. Luo S, Crainiceanu CM, Louis TA, Chatterjee N. Bayesian inference for smoking cessation with a latent cure state. *Biometrics* 2009; **65**:970–978.
12. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2002.
13. Diggle PJ, Heagerty P, Liang KY, Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press, 2002.
14. Molenberghs G, Verbeke G. *Models for Discrete Longitudinal Data*. Springer Verlag, 2005.
15. Anderson T. *An Introduction to Multivariate Statistical Analysis*, 3rd edn. John Wiley & Sons, 2003.
16. Gelman A, Carlin J, Stern H, Rubin D. *Bayesian Data Analysis*. CRC press, 2004.
17. Li Y, Wileyto EP, Heitjan DF. Modeling smoking cessation data with alternating states and a cure fraction using frailty models. *Statistics in Medicine* 2010; **29**(6):627–638.
18. Li Y, Wileyto EP, Heitjan DF. Prediction of individual long-term outcomes in smoking cessation trials using frailty models. *Biometrics* 2011; **67**:1321–1329.
19. Song X, Davidian M, Tsiatis AA. A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* 2002; **58**(4):742–753.
20. Zeng D, Cai J. Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *The Annals of Statistics* 2005; **33**(5):2132–2163.
21. Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* 2001; **88**(4):973.
22. Daniels MJ, Zhao YD. Modelling the random effects covariance matrix in longitudinal data. *Statistics in Medicine* 2003; **22**(10):1631–1647.