7-10-2008

BAYESIAN INFERENCE FOR SMOKING CESSATION WITH A LATENT CURE STATE

Sheng Luo
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, sluo@jhsph.edu

Ciprian M. Crainiceanu
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Thomas A. Louis
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Nilanjan Chatterjee
Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health

Bayesian Inference for Smoking Cessation with a Latent Cure State

Sheng Luo,[1,2,*] Ciprian M. Crainiceanu,[1] Thomas A. Louis,[1] and Nilanjan Chatterjee[2]

[1]Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St., Baltimore, Maryland 21205, USA

[2]Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, Maryland 20852, USA

[*]*email*: sluo@jhsph.edu

We present a Bayesian approach to modeling dynamic smoking addiction behavior processes when cure is not directly observed due to censoring. Subject-specific probabilities model the stochastic transitions among three behavioral states: smoking, transient quitting, and permanent quitting (absorbent state). A multivariate normal distribution for random effects is used to account for the potential correlation among the subject-specific transition probabilities. Inference is conducted using a Bayesian framework via Markov Chain Monte Carlo simulation. This framework provides various measures of subject-specific predictions, which are useful for policy making, intervention development, and evaluation. Simulations are used to validate our Bayesian methodology, and assess its frequentist properties. Our methods are motivated by, and applied to the Alpha-Tocopherol, Beta-Carotene (ATBC) Lung Cancer Prevention study, a large (29,133 individuals) longitudinal cohort study of smokers from Finland.

Key Words: Cure Model; MCMC, Mixed-effects Model; Prediction; Recurrent Events; Smoking Cessation.

1 Introduction

Cigarette smoking continues to be the leading cause of premature morbidity and mortality in the United States (Samet, 1992; McBride, 1992; CDC, 1997). Intervention efforts to encourage and assist smokers to quit are an important component of the public health campaign against this epidemic (Novotny *et al.* , 1992). The slow progress in reducing the prevalence of smoking in recent years is, in part, attributable to the high relapse rate (Cui *et al.* , 2006). Hunt *et al.* (1971) estimated relapse rates following treatment for smoking cessation at approximately 80%, while Glasgow & Lichtenstein (1987) found that between 50% and 75% of smokers who quit following treatment relapse within one year. Piasecki *et al.* (2002) conjectured that the poor treatment success rates reflects a lack of understanding of the dynamic nature of addiction and relapse processes. The goal of this article is to design, implement, and evaluate the statistical framework of the dynamic process of smoking cessation.

This problem was originally addressed by Luo *et al.* (2008) in an application to the Alpha-Tocopherol, Beta-Carotene (ATBC) Lung Cancer Prevention study. Here, we provide a brief description of the ATBC dataset as well as a summary of Luo *et al.* (2008) modeling approach. The ATBC study is a large (29, 133 individuals) longitudinal cohort study. The individuals were followed for 5 to 8 years with three follow-up visits per year (i.e., every four months). At each visit, each individual was queried about health and smoking status since the last visit. Smoking status was defined by the following question (translated from the Finnish): "Have you smoked since your last visit?" Individuals were allowed to indicate that during the previous four months, they (1) had not smoked at all, (2) had smoked but had stopped at some time during the interval, or (3) had smoked continuously. For the individuals who answered (2), the quit time, and duration of cessation were unknown. We do not distinguish between (2) and (3), treating them as "smokers since last visit." Individuals who answered (1) are treated as nonsmokers since their last visit. Therefore, the smoking status at each visit is either smoking or nonsmoking.

The smoking patterns alternate between smoking and nonsmoking states with sojourn time in each state varying within and between individuals. The smoking status is unknown after censoring. To describe the full stochastic nature of the smoking addiction pattern, Luo *et al.* (2008) proposed a discrete-time mixed-effects model with three states: smoking, transient cessation (temporarily smoking-free but relapse later), and permanent cessation (lifelong smoking-free), which is a latent state because of censoring. Rather than dichotomizing each individual as quitter or non-quitter as is the common practice in epidemiology, Luo *et al.* (2008) incorporated a "cure" component, and estimated the cure probability defined as the probability of permanent cessation given a quit attempt. Random subject-specific transition probabilities among these three states were used to account for the between-subject heterogeneity. Luo *et al.* (2008) provided a computationally fast fitting algorithm using an innovative combination of geometric-like distributions of waiting times between addiction states and Beta distributions of subject-specific random effects. This combination resulted in a closed-form marginal likelihood which, though complicated-looking, is easy to maximize using standard optimization software.

While in this article the stochastic smoking patterns are addressed by the same discrete-time mixed-effects model with three states as in Luo *et al.* (2008), we use a different modeling and inferential framework for random effects to address *subject-specific predictions* as well as potential correlation among random effects corresponding to the transitions among three states. We replace the independent Beta distributions of random effects by a multivariate normal distribution with non-zero off-diagonal elements in the variance-covariance matrix. Modeling and inference are naturally conducted using a Bayesian framework via Markov Chain Monte Carlo (MCMC) simulation. This framework provides the joint distribution of vectors of subject-specific measures of the addiction behavior, while incorporating the information in the data according to the rules of probability. In addition, there is an important computational difference between the current article and Luo *et al.* (2008). While we tried to obtain the subject-specific predictions with the approach in Luo

et al. (2008) using Bayesian inference based on MCMC simulations, we were unable to obtain good convergence and mixing properties. The multivariate normal distribution assumption used in this manuscript allows us to circumvent this problem and to account for potential correlation among random effects. Both these model characteristics are needed for satisfactory inference.

The model enhancements are important, but challenging to implement. Indeed, various measures of subject-specific predictions, and their associated variability are useful in assisting policy making, treatment, and intervention assessment. For example, to maximize the use of the limited resources for smoking control, it may be helpful to categorize the individuals in the ATBC study into multiple groups, e.g., high, and low propensity for quitting. Various intervention programs could be designed accordingly to meet the smoking cessation needs of different groups. For the motivated individuals with high propensity for quitting, consistent counseling may be effective to keep them smoke-free, while more aggressive treatments could be necessary for the "hard-core" smokers with low propensity for quitting. Moreover, one could evaluate, and identify the smoking patterns (e.g., the number, and length of quit attempts), which can greatly increase the propensity for quitting. This provides valuable guidance for the design, evaluation, and implementation of smoking-control strategies.

This article presents a Bayesian modeling approach to estimate the model parameters, and various measures of subject-level predictions. Although not as computationally efficient as the modeling framework in Luo *et al.* (2008), our algorithm remains feasible using modern computing platforms and software. For reproducibility of our results, we post our code, simulated data, and results at
`www.biostat.jhsph.edu/~ccrainic/webpage/programs/smoking/MCMC_Luo_smoking.zip`

In Section 2, we summarize the modeling framework in Luo *et al.* (2008). In Section 3, we introduce the Bayesian model and inference. In Section 4, we discuss the subject-specific predictions and evaluations using our modeling framework. A simulation study is used to illustrate the methodology in Section 5. The proposed methods are applied to the ATBC study dataset in Section 6. Finally, the article is concluded with discussions in Section 7.

2 A Stochastic Mixed Model for Addiction Behavior

To illustrate the complexity of the dataset, Figure 1a displays the smoking patterns of four individuals in the ATBC study. The follow-up visit numbers are shown on the x axis and the individuals' IDs are on the y axis. Within the interval between two consecutive visits (i.e., 4 months), the individuals either smoked (indicated by a shaded area) or did not smoke (indicated by a unshaded area). Some individuals experienced smoking and nonsmoking periods in an alternating fashion (e.g., individuals 2, 3, and 4), while others never made quit attempts (e.g., individual 1). Although the smoking patterns are unknown after censoring, the long trailing nonsmoking intervals of some individuals (e.g., individuals 2 and 3) suggest the existence of a potential "cured" subpopulation (i.e., individuals who successfully quit smoking). This type of data arises frequently in medical studies such as infectious diseases (e.g., ear infection, (Eerola *et al.* , 2003); Pnc bacteria carriage, (Auranen *et al.* , 2000); Hib infection, (Auranen, 2000)), chronic diseases (e.g., epilepsy, (Cowling *et al.* , 2006); soft tissue sarcoma, (Huang *et al.* , 2006)), and substance addiction, where patients make transitions among several disease states or between the presence or absence of symptoms. After the administration of various treatments, some patients are cured, and no longer experience disease states or symptoms.

Luo *et al.* (2008) modeled the data using a 3-state discrete-time stochastic mixed-effects model with subject-specific transition probabilities denoted by P_{ij}, with $j = 1, 2, 3$, as illustrated in Figure 1b. This model distinguishes the transient from the permanent quitting state because the processes that describe transient and permanent quitting are likely to be different, and have different policy making implications. When individual i is in the smoking state, quit attempts are made at the beginning of each 4 month interval with probability P_{i1}. Once a quit attempt is made, the individual may become a permanent quitter with probability P_{i3} at the visit following the quit attempt. With probability $1 - P_{i3}$, the individual enters the transient quitting state, from which he has probability P_{i2} to relapse back to the smoking state in the current interval. Conditional on the random rates P_{ij}, the transition to the next state is determined only by the current and the previous states. A

quit attempt is defined as the nonsmoking interval immediately after the smoking intervals. The quit attempt is a gateway either to permanent or to transient quitting and is not a state in the proposed stochastic process.

This modeling structure can be described 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. The second type (Type II) of geometric process models the number of nonsmoking intervals before next relapse (a relapse is defined as the smoking interval immediately after the nonsmoking intervals), conditional on being in a transient quitting state. The likelihood for individual i is constructed by multiplying the likelihood contribution of both types of processes

$$\begin{aligned} L_i &= P_{i1}^{K_{i1}}(1-P_{i1})^{S_{i1}}P_{i2}^{K_{i2}}(1-P_{i2})^{S_{i2}+N_{ik_3}}(1-P_{i3})^{K_{i2}+1} \\ &\quad + P_{i1}^{K_{i1}}(1-P_{i1})^{S_{i1}}P_{i2}^{K_{i2}}(1-P_{i2})^{S_{i2}}P_{i3}(1-P_{i3})^{K_{i2}}, \end{aligned} \tag{1}$$

where K_{i1} is the number of quit attempts for individual i, K_{i2} is the number of of relapses (unsuccessful quit attempts), S_{i1} is the total number of smoking intervals excluding the relapsing intervals, S_{i2} is the total number of nonsmoking intervals (excluding the quit attempts) in the Type II geometric process with observed relapses, N_{ik_3} is the number of trailing nonsmoking intervals (i.e., the nonsmoking intervals between the final quit attempt and censoring if the last observed interval is neither smoking nor a quit attempt, and $N_{ik_3} = 0$ otherwise). For a detailed derivation of the likelihood formulation in (1), please see Section 2.2 in Luo et al. (2008).

The likelihood in (1) is a sum of products of binomial-like distributions with the transition probabilities P_{ij} being the "success" probabilities. By assuming that P_{ij} have Beta distributions, and are independent given the covariates, the closed-form of the marginal likelihood can be obtained by integrating out P_{ij}. The stochastic model and the likelihood formulation in this article are similar to Luo *et al.* (2008), but the random effects are modeled using the multivariate normal distribution. This distribution conveniently accounts for between-subject heterogeneity and within-subject corre-

lation in the transition probabilities. For individual i ($i = 1, \ldots, m$, where m is the total number of individuals), let $\boldsymbol{y}_i$ denote the outcome variable vector. Corresponding to transition probability P_{ij}, let $\mathbf{X_{ij}}$ denote a $p \times 1$ vector of predictors. Let $\boldsymbol{\beta}_j$ be a $p \times 1$ vector of fixed effects regression coefficients and let u_{ij} be the subject-specific random effects for transition probability P_{ij}. To model the transition probability vector $\mathbf{P}_i = (P_{i1}, P_{i2}, P_{i3})$ for individual i , we let

$$g_j(P_{ij}; u_{ij}) = \boldsymbol{X}_{ij}\boldsymbol{\beta}_j + u_{ij} \quad \text{for } j = 1, 2, 3, \tag{2}$$

where $g_j(\cdot)$ are some link functions. For example, we let $g_1(\cdot)$ and $g_2(\cdot)$ be the complementary log-log link function and let $g_3(\cdot)$ be the logit link function. We use the complementary log-log link to make the transition probabilities between smoking and transient quitting states analogous to hazard functions in discrete-time proportional hazards model (Kalbfleisch & Prentice, 2002). Note that $\boldsymbol{\beta}_j$ may be the same or different for different subscripts j and denote by $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \boldsymbol{\beta}_3')'$.

The trivariate random effects vectors $\boldsymbol{u}_i = (u_{i1}, u_{i2}, u_{i3})'$ are assumed to be independent and identically distributed with normal probability density function $h(\boldsymbol{u}_i; \boldsymbol{\Sigma})$, i.e., $\boldsymbol{u}_i|\boldsymbol{\Sigma} \sim \mathrm{N}_3(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{u}_i = (u_{i1}, u_{i2}, u_{i3})^t$ and $\boldsymbol{\Sigma}$ is an unknown 3×3 covariance matrix with the (i, j)th entry denoted by σ_{ij}. The non-zero off-diagonal elements in $\boldsymbol{\Sigma}$ can account for the within-individual dependence among random transition probabilities. With this structure of random effects, the marginal likelihood for individual i is $L_i(\boldsymbol{\Phi}; \boldsymbol{y}_i) = \int L_i(P_{ij}|\boldsymbol{u}_i; \boldsymbol{\beta}_j) h(\boldsymbol{u}_i; \boldsymbol{\Sigma}) d\boldsymbol{u}_i$, where $\boldsymbol{\Phi} = (\boldsymbol{\beta}, \boldsymbol{\Sigma})$. This integral cannot be evaluated analytically as in Luo *et al.* (2008). To avoid this problem, we use Bayesian inference based on MCMC posterior simulations.

3 Bayesian Inference

In this section we describe our Bayesian framework. Recall that the model parameters are $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ describing the mean and correlation structure of the transition probabilities, respectively. We used the following priors $\beta_{jk} \sim \mathrm{N}(0, 100)$, where $j = 1, 2, 3$, $k = 1 \ldots p$, and p varies with the model. For $\boldsymbol{\Sigma}$, we use an approach suggested by Moller & Syversveen (1998), which is based on the Cholesky

decomposition. Let $\mathbf{\Sigma} = \Omega\Omega'$, where Ω is a matrix with zero entries above the main diagonal and let $\omega_{i,j}$ be the (i,j)th entry for $i \leq j$. Consider a latent random effects vector $\boldsymbol{z}_i = (z_{i1}, z_{i2}, z_{i3})'$ with independent $N(0,1)$ components. Then $\boldsymbol{u}_i = \Omega\boldsymbol{z}_i$ has mean zero, and variance $\mathbf{\Sigma}$. This corresponds to the following linear reparameterization of the random effects $\boldsymbol{u}_i$:

$$u_{i1} = \omega_{11}z_{i1}; \quad u_{i2} = \omega_{12}z_{i1} + \omega_{22}z_{i2}; \quad u_{i3} = \omega_{13}z_{i1} + \omega_{23}z_{i2} + \omega_{33}z_{i3}. \tag{3}$$

Note that the entries of the matrix $\mathbf{\Sigma}$ are computed as $\sigma_{jk} = \sum_{l=1}^{j\wedge k} \omega_{lj}\omega_{lk}$, $1 \leq j,k \leq 3$, where $j \wedge k = \min(j,k)$. Non-negativity constraints on ω_{11}, ω_{22}, and ω_{33} are imposed by assuming Uniform$(0,10)$ prior distributions. The prior distributions for ω_{12}, ω_{13}, and ω_{23} are N$(0,100)$ to allow for potential negative correlation in $\mathbf{\Sigma}$. For notational convenience, let $\boldsymbol{\sigma} = (\sigma_{11}, \sigma_{12}, \sigma_{13}, \sigma_{22}, \sigma_{23}, \sigma_{33})$, $\boldsymbol{\omega} = (\omega_{11}, \omega_{12}, \omega_{13}, \omega_{22}, \omega_{23}, \omega_{33})$, $\boldsymbol{z}_i = (z_{i1}, z_{i2}, z_{i3})$, and $\boldsymbol{\rho} = (\rho_{12}, \rho_{13}, \rho_{23})$ denotes the pairwise correlation coefficients among the components in random effects $\boldsymbol{u}_i$. The joint distribution of the data and parameters is

$$P(\boldsymbol{\beta}, \mathbf{\Sigma}) = \prod_{i=1}^{m}\left[L_i(\boldsymbol{y}_i; \mathbf{P})\left\{\prod_{\mathbf{j=1}}^{\mathbf{3}} \mathbf{p}(\mathbf{P_{ij}}; \boldsymbol{\beta}_\mathbf{j}, \boldsymbol{\omega}, \boldsymbol{z}_\mathbf{i})\mathbf{P}(\boldsymbol{z}_\mathbf{i})\right\}\right]\mathbf{P}(\boldsymbol{\beta})\mathbf{P}(\boldsymbol{\omega}). \tag{4}$$

and the full conditionals are detailed in Web Appendix A.

We can substitute (2) and (3) into the above full conditional distributions to get the functions in terms of $\boldsymbol{\beta}_j$, $\boldsymbol{\omega}$, and $\boldsymbol{z}$. The parameters are updated in the following order $(\boldsymbol{\beta}_1, \omega_{11})$, $(\boldsymbol{\beta}_2, \omega_{12}, \omega_{22})$, $(\boldsymbol{\beta}_3, \omega_{13}, \omega_{23}, \omega_{33})$, and (z_{i1}, z_{i2}, z_{i3}). These full conditionals do not have an explicit form and are simulated using the single-component Metropolis-Hastings (M-H) algorithm (Metropolis *et al.* , 1953; Hastings, 1970; Li, 1988) with a normal proposal distribution centered at the current value and a small variance. Each parameter or block of parameters is updated in turn by conditioning on all other parameters (Geman & Geman, 1984; Gelfand & Smith, 1990). The posterior distributions of $\boldsymbol{\sigma}$ and $\boldsymbol{\rho}$ are computed from the posterior samples of $\boldsymbol{\omega}$ and the posterior distributions of the subject-specific transition probabilities P_{ij} are computed from the posterior samples of $\boldsymbol{\beta}$ and $\boldsymbol{u}_i$.

4 Subject-Specific Predictions and Evaluations

Our model and Bayesian inferential machinery provide straightforward subject-specific prediction calculation even in very complex, but policy relevant, contexts. For example, given the study data and model, one might be interested in predicting "who is a permanent quitter two years after censoring". Note that the individual who was smoking at censoring has probability zero to be a permanent quitter at censoring, but nonnegative probability of being a permanent quitter two years after censoring. In Section 4.1, we show how to calculate the probability of permanent quitting two years after censoring (we call it 2-year quitting probability and denote it by P_i). In Section 4.2, we describe the predictive properties of the decision making process based on P_i.

4.1 Subject-Specific Predictions

In this section, we show how to derive the 2-year quitting probability, P_i, which is defined as the probability that the ith individual who was followed in the study for n_i years becomes a permanent quitter by the end of year $n_i + 2$. For example, for an individual who was followed for five years in the ATBC study, the 2-year quitting probability is the probability that he becomes a permanent quitter by the end of year 7.

To compute P_i, the unobserved 2-year (24 months) period that follows the observed smoking pattern is partitioned into 6 four-month intervals to emulate the design of the ATBC study. To ensure that permanent quitting happens by the end of the 2-year period, the last two intervals (the fifth and the sixth) must be nonsmoking. This is because (i) if permanent quitting occurs at or before the fifth interval, the last two intervals are nonsmoking (e.g., the smoking patterns 1 to 8 in Web Figure 1); (ii) if the quit attempt occurs at the fifth interval and permanent quitting occurs at the last interval, the last two intervals are still nonsmoking (e.g., the smoking patterns 9 to 16 in Web Figure 1). Denoted by **SP** are the 16 possible smoking patterns for the first four intervals. Web Figure 1 displays **SP** indicating that permanent quitting happens at or before the second interval for pattern 1, at the third interval for pattern 2, at the fourth interval for patterns 3 and 4, at the fifth

interval for patterns 5 to 8, and at the sixth interval for patterns 9 to 16, respectively.

For individual i, the probability that smoking pattern j occurs is

$$P_i(\text{sp}_j|P_{i1}, P_{i2}, P_{i3}) = P_{i1}^{K_{i1j}}(1-P_{i1})^{S_{i1j}}P_{i2}^{K_{i2j}}(1-P_{i2})^{S_{i2j}}P_{i3}(1-P_{i3})^{K_{i2j}}, \tag{5}$$

where sp_j denotes the smoking pattern j. The 2-year quitting probability for individual i is

$$P_i(\text{quit in 2-year}) = \sum_{\text{sp}_j \in \mathbf{SP}} P_i(\text{sp}_j|P_{i1}, P_{i2}, P_{i3}). \tag{6}$$

The probability formulation in (5) is slightly different from (1) because it does not contain N_{ik_3}. This is because permanent quitting occurs in every smoking pattern and there is no need to account for the probability of observing the same pattern if permanent quitting does not occur. Note that for pattern j, K_{i1j}, S_{i1j}, K_{i2j}, S_{i2j}, N_{ik_3j}, and $P_i(\text{sp}_j|P_{i1}, P_{i2}, P_{i3})$ are different for the individual who smoked at censoring and for the one who did not smoke at censoring. For example, for pattern 1, we have $K_{i11} = 1$, $S_{i11} = 0$, $K_{i21} = 0$, $S_{i21} = 0$, $N_{ik_31} = 5$, and $P_i(\text{sp}_j|P_{i1}, P_{i2}, P_{i3}) = P_{i1}P_{i3}$ for the individual who smoked at censoring, but $K_{i11} = 0$, $S_{i11} = 0$, $K_{i21} = 0$, $S_{i21} = 0$, $N_{ik_31} = 6$, and $P_i(\text{sp}_j|P_{i1}, P_{i2}, P_{i3}) = P_{i3}$ for the individual who did not smoke at censoring. When P_{i1} is small, $P_{i1}P_{i3} \ll P_{i3}$. This explains why the individual who smoked at censoring has a much smaller 2-year quitting probability P_i than the individual who smoked at censoring.

4.2 Decision-Making Evaluation

The 2-year quitting probabilities calculated in the previous section are very important predictive measures and could be used in a decision-making framework. One way to formalize such a framework could be to categorize the ATBC individuals into two groups, e.g., permanent quitters and non-permanent quitters. A reasonable decision rule could be to fix a particular probability threshold (i.e., p_0), and predict that individuals are permanent quitters if $P_i > p_0$ and are non-permanent quitters if $P_i \leq p_0$. To study the properties of this classification procedure we investigate the effect of various thresholds on its sensitivity and specificity. Sensitivity is defined as $\text{Sens}(Q, p_0) = \frac{1}{Q}\sum_{i\in Q} I\{P_i > p_0\}$, where Q is the set of the true permanent quitters. Sensitivity is the frequency with which the

procedure correctly identifies the permanent quitters (true positive) using the probability threshold p_0. Similarly, the specificity is defined as $\text{Spec}(Q, p_0) = \frac{1}{|M\backslash Q|} \sum_{i \in M\backslash Q} I\{P_i \leq p_0\}$, where M is set of all individuals, and $M\backslash Q$ denotes the set of non-permanent quitters, and $|M\backslash Q|$ is the cardinality of the set $M\backslash Q$. Specificity is the frequency with which the procedure correctly identifies the individuals who are non-permanent quitters (true negative) using the probability threshold p_0. The threshold p_0 could be anything between 0 and 1, but some insight into reasonable values can be obtained using simulations, as described in Section 5.

5 Simulation Study

In this section, we evaluate the performance of our methodology using simulations. To start with, we consider data generating processes that are straightforward to explain, but complex enough to capture the main features of the data. We consider the case when all processes depend only on one binary covariate X_i, e.g., the baseline insomnia symptom. Because the prevalence of insomnia at baseline is around 20% in the ATBC study, the covariate X_i is simulated independently from a Bernoulli distribution with success probability .2 for $i = 1, \ldots, m$, where $m = 10,000$ and for $N = 100$ simulated datasets. After the covariate for individual i is generated, the smoking pattern is generated using the following algorithm.

1. Simulate the number of follow-up visits independently from a $N(14.7, 5.8^2)$ distribution (this is an approximation of the empirical distribution of number of visits in the ATBC study).

2. Simulate independently $\boldsymbol{u}_i \sim N(0, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \begin{pmatrix} .09 & -.01 & -.12 \\ -.01 & .16 & .05 \\ -.12 & .05 & .25 \end{pmatrix}.$$

We make this choice of $\boldsymbol{\Sigma}$ to approximate the results in the ATBC study.

3. Simulate P_{ij} using (2) with $\boldsymbol{\beta}_1 = (.186, -1.217)'$, $\boldsymbol{\beta}_2 = (-1.031, 1.217)'$ and $\boldsymbol{\beta}_3 = (.405, -2.603)'$.

4. Conditional on smoking in the last interval, simulate the number of smoking intervals before

the next quit attempt via the Type I geometric process with success probability P_{i1}.

5. With probability P_{i3}, the individual becomes a permanent quitter, all the remaining intervals are nonsmoking and the simulation for individual i is finished.

6. With probability $1 - P_{i3}$, the individual enters a transient quitting state. The number of nonsmoking intervals before the next relapse is simulated from the Type II geometric process with succes probability P_{i2}.

7. Repeat until the smoking pattern is generated for each individual.

Using the Bayesian methodology described in Section 3, we obtain the joint distributions of all model parameters given the data. For each simulated dataset, we run five parallel chains using initial values that are over-dispersed. For each of the five chains, we run $100,000$ simulations. The first $20,000$ simulations of each chain are discarded, and inference is based on the remaining $80,000$ simulations from each chain. The MCMC convergence and mixing properties are assessed by visual inspection of the chain histories of many parameters of interest. Web Figure 2, 3, and 4 display the histories of 12 parameters of interest from three randomly selected chains for one of the simulated datasets. These plots indicate reasonable convergence and mixing properties, even though, for clarity, we only display every 500th simulation. Similar good chain properties have been noted in all other examples presented in this article.

Simulation results are reported in Web Table 1. The row labeled "EST" provides the average of the posterior means from 100 simulated datasets. The row labeled "SE" provides square root of the average of the variances. The nominal 95% credible intervals of parameters (e.g., σ_{11}) are obtained from the 2.5% and 97.5% percentiles of the posterior distributions of the parameters (denoted by $\hat{\sigma}_{i11}^{2.5}$ and $\hat{\sigma}_{i11}^{97.5}$ for simulated dataset i). The coverage probabilities of these intervals (displayed in the row labeled "Coverage probability") are calculated as $\sum_{i=1}^{N} I(\widehat{\sigma}_{i11}^{2.5} \leq \sigma_{11} \leq \widehat{\sigma}_{i11}^{97.5})/N$, where $I(\cdot)$ denotes the indicator function. Results in Web Table 1 indicate that bias is negligible and the credible interval coverage probabilities are reasonably close to the nominal level of 95 percent.

To predict who is a permanent quitter two years after censoring, we use the threshold method described in Section 4 and calculate the sensitivity and specificity functions. To gain some insight into how the variability of the estimated P_{ij} change the results, we substitute into (5) and (6) either the true P_{ij} generated in Step 3 of the simulation algorithm or the estimated P_{ij} from MCMC samples.

Figure 2 displays the means and the 2.5% and 97.5% percentiles for sensitivity and specificity at each threshold p_0. The means and the percentiles are obtained using the 100 simulated datasets. Figure 2 clearly shows the trade-off between sensitivity and specificity. When the threshold p_0 is less than .5, the sensitivity and specificity using the estimated P_{ij} (solid line with dash-dot lines for the percentiles) and using the true P_{ij} (solid line with dashed lines for the percentiles) agree very closely, e.g., the sensitivity remains at about .9 while the specificity plateaus at about .7. When p_0 varies between .5 and .8, the results using the estimated P_{ij} deviate markedly from the ones using the true P_{ij}, and show larger variability. These larger deviation and variability are partially due to higher statistical variability of the estimated P_{ij}. Moreover, low sensitivity is traded off for high specificity in this range of thresholds, p_0. When $p_0 > .8$, the sensitivity gradually reaches zero and the specificity gradually reaches one. When a threshold $p_0 \in [.3, .5]$ is selected, one could obtain roughly .9 average sensitivity and .7 average specificity using the estimated P_{ij}. In this range of thresholds p_0, the sensitivity and specificity results using the estimated P_{ij} and using the true P_{ij} line up almost perfectly. This figure provides insight on the range of thresholds when the classification and decision-making are of scientific interest.

6 Application to the ATBC Study

6.1 Parameter estimation and interpretation

In this section, we apply the proposed methodology to the ATBC dataset. For all results in this section we use five parallel chains with overdispersed initial values with respect to the posterior, and run each chain for $150,000$ simulations. The first $50,000$ simulations are discarded, and the parameter estimates are based on the remaining $100,000$ simulations from each chain.

First, we fit a simplified model with only one binary covariate, presence of the insomnia symptom at baseline. Web Table 2 provides the posterior means, standard deviations and 95% credible intervals for some of the parameters of interest. A negative sign for the insomnia effect indicates a smaller probability of having a certain event. For example, the individuals with insomnia are less likely to make a quit attempt than those without. After the quit attempt is made, the estimated odds ratio of permanent cessation is .748 (i.e., $\exp(-.29)$; 95% CI: $[.571, .970]$) comparing the individuals with insomnia to those without. These results are consistent with those in Luo *et al.* (2008) both in direction and magnitude. While expected given the large sample size, it is reassuring that the different structure of random effects does not have a more serious impact on our marginal inferences. Web Table 2 also shows a high negative correlation between u_{i1} and u_{i3}, i.e., $\rho_{13} = -.93$. We provide more insight into this at the end of this section.

Second, we fit a richer model with the following eight covariates: age, years of smoking, cigarettes per day, alcohol consumption (g/day), inhalation (yes/no), and factor 1, 2, and 3 obtained from a factor analysis on the 16 baseline symptoms. The 16 baseline symptoms are: anxiety, depression, poor memory, difficulty concentrating, fatigue, poor appetite, insomnia, headache, back ache, walking pain in knees, joint ache, muscle ache, walking pain in hips, leg cramps, nocturnal restless legs, and cutaneous itching. The covariates age, years of smoking, cigarettes per day, alcohol consumption are centered and standardized. For interpretability of results, note that factors 1 and 2 are heavily loaded on psychological and chronic medical conditions symptoms, respectively. Factor 3 is heavily loaded on insomnia and walking pain, but it only explains 6.6% of the total variance. The history plots of the chains for the model parameters are omitted because of space limit, but the mixing property of the chains are comparable to the ones in the simplified model.

The rows labeled P_{i1} in Table 1 display the results of modeling the probability of making quit attempts. A negative sign of a parameter β indicates a smaller probability of having an event, i.e., making a quit attempt. Therefore, we conclude that older individuals have higher probability of

making quit attempts, while years of smoking, cigarette, and alcohol consumption are negatively associated with the probability of making quit attempts. The rows labeled P_{i2} in Table 1 show the results of modeling the probability of relapsing for the transient quitters, conditional on making a quit attempt and being in a transient quitting state. We conclude that the individuals who smoked more cigarettes per day take longer to relapse when they are in a transient quitting state. This is unexpected, but consistent with results in Luo *et al.* (2008). Finally, the rows labeled P_{i3} in Table 1 provide the results of modeling the probability of being a permanent quitter, conditional on making a quit attempt. We conclude that the odds ratio of permanent cessation for an increase of 8.4 years of smoking history (i.e., one standard deviation) is 1.160 (i.e., $\exp(.148)$; 95% CI: $[1.036, 1.309]$), holding other covariates fixed. In addition, individuals with psychological symptoms (factor 1) have significantly smaller probability of quitting permanently. The odds ratio of permanent quitting for one unit increase in factor 1 is .890 (i.e., $\exp(-.115)$; 95% CI: $[.787, .998]$), holding other covariates fixed. The results in Table 1 are consistent with Table 6 in Luo *et al.* (2008) with respect to the direction, size, and significance of covariates, e.g., age, year of smoking, cigarette, and alcohol consumption in modeling P_{i1}, cigarette consumption in modeling P_{i2}, and factor 1 in modeling P_{i3}. However, our modeling results show a significant positive association between years of smoking and probability of permanent quitting, while Luo *et al.* (2008) reports an insignificant negative association.

Web Table 2 and and Table 1 display high negative correlation between P_{i1} and P_{i3} (ρ_{13}), and relatively high positive correlation between P_{i2} and P_{i3} (ρ_{23}). We now provide some insight into why these correlations may occur. Consider first $\widehat{\rho}_{13}$. Note that there are $1,974$ (6.8%) long-term sustainers, i.e., individuals who did not smoke for at least 10 consecutive visits (40 months) and sustained until censoring. These long-term sustainers, in our model, are most likely to be permanent quitters, and contribute the most to estimating the parameters of P_{i3}. Among them, 1899 (96.2%) made only one quit attempt, indicating why high P_{i3} (long trailing nonsmoking intervals) might be so highly associated with small P_{i1} (few quit attempts). Consider next $\widehat{\rho}_{23}$. Note that there are

1, 188 (4.1%) relapsers, i.e., individuals who had at least one quit attempt but did not have a trailing nonsmoking interval. These relapsers, in our model, are most likely to have a small P_{i3}. We count the number of nonsmoking intervals before an observable relapse (take average if multiple relapses occurred) for every relapser. Relapsers had an average smoke-free interval of 2.6 visits (10.4 months) before next relapse. This relatively long smoke-free interval before relapsing indicates small P_{i2}. Therefore, the association of small P_{i2} and small P_{i3} might lead to a high correlation coefficient ρ_{23}.

6.2 Subject-Specific Predictions in the ATBC Study

In this section, we provide more insight into our model's ability to provide subject-specific estimates and predictions in the ATBC study conditional on the observed covariates and smoking patterns.

Figure 3 displays the smoking patterns of seven individuals in the ATBC study, who had 20 visits before censoring and no baseline insomnia symptoms. These individuals had different numbers of quit attempts and various sojourn time distributions in the smoking and nonsmoking states. Table 2 presents the number of nonsmoking intervals (in the column labeled "NS") and quit attempts (in the column labeled "QA") for all seven individuals. Using the results of the simplified model, we calculate the subject-specific posterior means of the transition probabilities P_{ij} for these seven individuals. In addition, we report the subject-specific 2-year quitting probability, P_i, as illustrated in Section 4. These estimates are displayed in columns 4 to 7 in Table 2. For reference, the last row of Table 2 presents the population means of the numbers of nonsmoking intervals and quit attempts and P_{ij} and P_i of all individuals in the ATBC study.

Table 2 reveals how the smoking patterns change the subject-specific probabilities among the individuals with identical numbers of visits and covariates. For examples, more quit attempts correspond to a higher P_{i1} (e.g., .014 in individual 1 vs. .123 in individual 7). Among the individuals with the same number of quit attempts, we conclude that (1). earlier quit attempts correspond to increased P_{i1} (e.g., .026 in individual 3 vs. .055 in individual 4; and .066 in individual 5 vs. .095 in individual 6); (2). the existence of a trailing nonsmoking interval corresponds to greatly increased

P_{i3} (e.g., .783 in individual 3 vs. .424 in individual 4; and .443 in individual 5 vs. .260 in individual 6); (3). the existence of a trailing nonsmoking interval also corresponds to increased 2-year quitting probability P_i (e.g., .796 in individual 3 vs. .105 in individual 4). As a last point, among individuals with the same length of a trailing nonsmoking interval, the ones with previous quit attempts have smaller P_{i3} and P_i than those without (e.g., individual 2 vs. 3). Intuitively, the individuals with more unsuccessful quit attempts are more likely to be transient quitters in the trailing nonsmoking interval because this interval tends to be a recurrence of the previous unsuccessful quit attempts.

7 Discussion

In this article, we introduce a computationally feasible Bayesian framework for the analysis of smoking cessation patterns with a latent cure state. This framework provides various subject-specific predictions by modeling the stochastic smoking behavior as a function of covariates and random effects. The approach expands the functionality of the framework proposed in Luo *et al.* (2008) by accounting for the correlations among subject-specific transition probabilities. It also provides additional insight into the relations among the dynamic smoking and quitting processes. We show how subject specific transition probabilities, P_{ij}, vary with smoking patterns across individuals, which provides useful prognostic information for efficient development, targeting and evaluation of interventions.

Our cure model is based on unobserved states (permanent quitting) that are identified through weak assumptions. Thus, it is reasonable to study the stability of parameter estimates to departures from the model assumptions and to potential nearly unidentified parameters Li *et al.* (2001). In particular, we evaluate the effect of using various link functions and parametric assumptions on the random effects. The size of effects and scientific interpretation under different link functions (e.g., logit and probit links) are basically unchanged. Moreover, Luo *et al.* (2008) used independent Beta distributions for the random effects and obtained essentially similar scientific results. It is important to note that, in practice, it is hard to match the flexibility of the multivariate normal

random effects assumption. For example, inducing correlation in a non-normal multivariate vector is theoretically possible, but computationally challenging. This is also the reason why we do not attempt to implement correlated nonparametric distributions of random effects.

Bayesian inference via MCMC simulations can be implemented and produces reliable and reproducible results for complex addiction behavior data (see the software posted on the link in Section 1). However, model fitting is computationally intensive. For example, it takes around 5.1 seconds to complete one sampling cycle for one of the datasets simulated in Section 5 on a PC (Dell workstation XPS Gen3, Pentium 4 3.6 Ghz dual processers, 2G RAM). It would take about 142 hours to get $100,000$ samples for a single MCMC chain. In contrast, it takes the Beta random effects methods proposed by Luo *et al.* (2008) only about 4 minutes to get the estimates. The large difference in computational time is due to the explicit function of the model parameters in Luo *et al.* (2008). Even though our implementation is slower, the models and inferences in this article produce inferential results that could not be obtained by the faster approach of Luo *et al.* (2008), e.g., subject-level predictions and residual correlation inferences.

Supplementary Materials

Web Appendix, Tables, and Figures referenced in Sections 3 to 6 are available under the Paper information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

ACKNOWLEDGMENTS

This work was supported by a predoctoral fellowship to the first author from the National Cancer Institute, National Institutes of Health. Support was provided to Thomas A. Louis by grant R01 DK061662.

References

Auranen, Kari. 2000. Back-calculating the Age-specific Incidence of Recurrent Subclinical *Haemophilus Influenzae* Type B Infection. *Statistics in Medicine*, **19**(3), 281–296.

Auranen, Kari, Arjas, Elja, Leino, Tuija, & Takala, Aino K. 2000. Transmission of Pneumococcal Carriage in Families: A Latent Markov Process Model for Binary Longitudinal Data. *Journal of the American Statistical Association*, **95**(452), 1044–1053.

CDC. 1997. Smoking-attributable mortality and years of potential life lost—United States, 1984. *MMWR*, **46**, 444–51.

Cowling, B. J., Hutton, J. L., & Shaw, J. E. H. 2006. Joint Modelling of Event Counts and Survival Times. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, **55**(1), 31–39.

Cui, Yong, Wen, Wanqing, Moriarty, Cynthia J., & Levine, Robert S. 2006. Risk factors and their effects on the dynamic process of smoking relapse among veteran smokers. *Behaviour Research and Therapy*, **44(7)**, 967–81.

Eerola, Mervi, Gasbarra, Dario, Helena Mkel, P., Linden, Henri, & Andreev, Andrei. 2003. Joint Modelling of Recurrent Infections and Antibody Response by Bayesian Data Augmentation. *Scandinavian Journal of Statistics*, **30**(4), 677–698.

Gelfand, A. E., & Smith, A. F. M. 1990. Sampling-based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85**, 398–409.

Geman, S., & Geman, D. 1984. Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Glasgow, R. E., & Lichtenstein, E. 1987. Long-term effects of behavioral smoking cessation interventions. *Behavior Therapy*, **18**, 297–324.

Hastings, W. K. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**, 97–109.

Huang, Xuelin, Cormier, Janice N., & Pisters, Peter W. T. 2006. Estimation of the Causal Effects on Survival of Two-stage Nonrandomized Treatment Sequences for Recurrent Diseases. *Biometrics*, **62**(3), 901–909.

Hunt, W. A., Barnett, L. W., & Brauch, L. G. 1971. Relapse rates in addiction programs. *Journal of Clinical Psychology*, **27**, 455–56.

Kalbfleisch, J. D., & Prentice, Ross L. 2002. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

Li, Chin-Shang, Taylor, Jeremy MG, & Sy, Judy P. 2001. Identifiability of Cure Models. *Statistics and Probability Letters*, **54**(4), 389–395.

Li, K-H. 1988. Imputation Using Markov Chains. *Journal of Statistical Computing and Simulation*, **30**, 57–79.

Luo, S., Crainiceanu, C. M., Louis, T., & Chatterjee, N. 2008. Analysis of Smoking Cessation Patterns Using a Stochastic Mixed Effects Model with a Latent Cured State. *Journal of the American Statistical Association to apprear.*

McBride, P. E. 1992. The health consequences of smoking. Cardiovascular diseases. *The Medical Clinics of North America*, **76(2)**, 333–353.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. 1953. Equations of State Calculations by Fast Computing Machine. *Journal of Chem Phys.*, **21**, 1087–1091.

Moller, J., & Syversveen, A. R. 1998. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, **25**, 451–482.

Novotny, T. E., Romano, R. A., Davis, R. M., & Mills, S. L. 1992. The public health practice of tobacco control: lessons learned and directions for the states in the 1990s. *Annual Review of Public Health*, **13**, 287–318.

Piasecki, T. M., Fiore, M. C., McCarthy, D. E., & B., Baker. T. 2002. Have we lost our way? The need for dynamic formulations of smoking relapse proneness. *Addiction*, **97(9)**, 1093–1108.

Samet, J. M. 1992. The health benefits of smoking cessation. *The Medical Clinics of North America*, **76(2)**, 399–414.

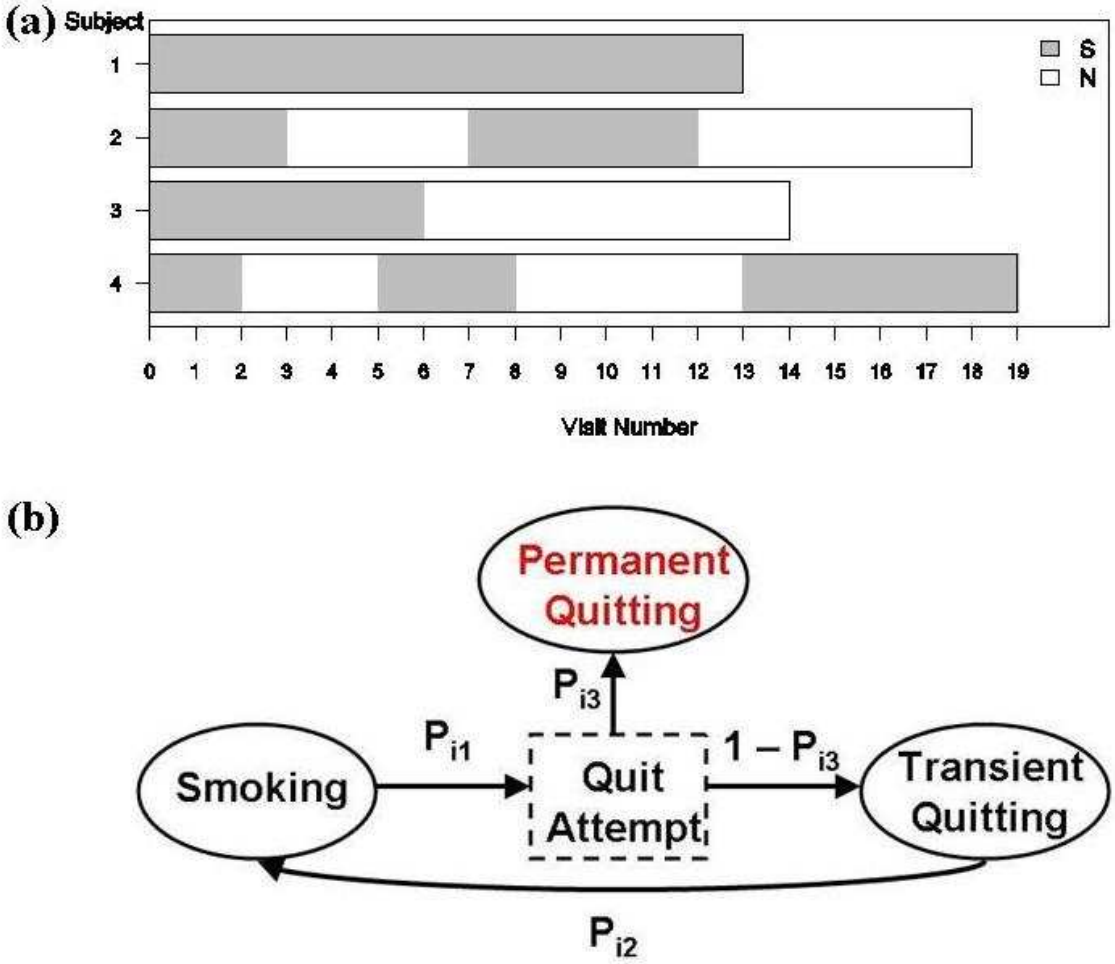

Figure 1: (a) Sample profiles of some smoking patterns from the ATBC study. Shaded regions indicate reported smoking and unshaded regions indicate reported nonsmoking. (b) Transition among three states.

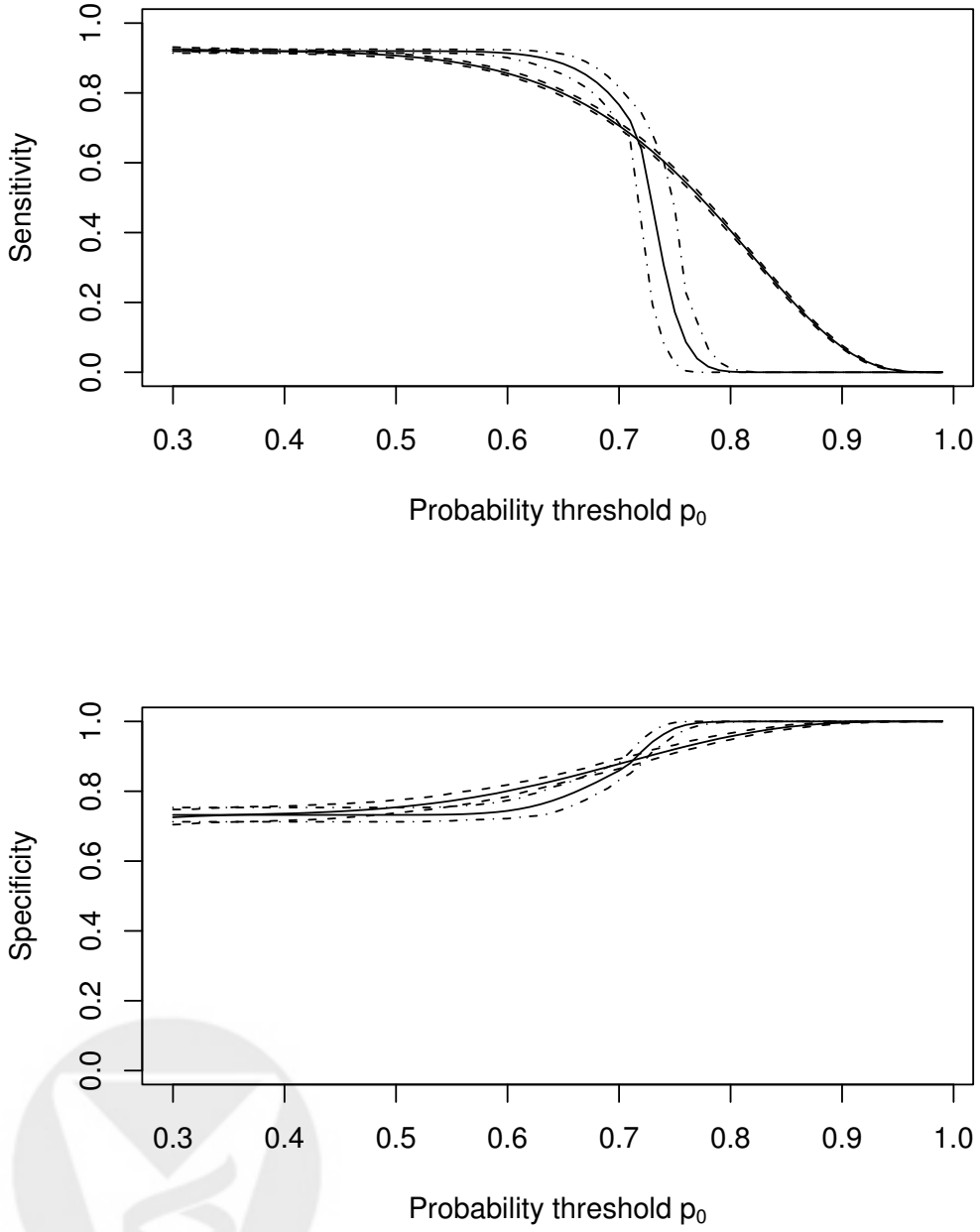

Figure 2: The means and the 2.5% and 97.5% percentiles for sensitivity and specificity from 100 simulated datasets. The results from the true P_{ij} is displayed as a solid line with dashed lines for the percentiles. The results from the estimated P_{ij} is displayed with a solid line with dash-dot lines for the percentiles.

Table 1: The posterior means (PM), standard deviations (SD)and 95% credible intervals (CI) of the parameters from (2) for 8 covariates in the ATBC dataset

Models	Parameters	PM	SD	95% CI	
				lower	upper
	Intercept	−4.366	.027	−4.419	−4.314
	age*	.208	.017	.175	.243
	Years Smoked*	−.281	.015	−.312	−.251
	cigarettes/day*	−.301	.016	−.333	−.269
P_{i1}	alcohol*	−.199	.018	−.235	−.163
	factor1	.023	.015	−.006	.052
	factor2	−.001	.014	−.029	.027
	factor3	.017	.013	−.009	.042
	inhale	.006	.029	−.052	.063
	Intercept	−.380	.237	−.817	.118
	age	−.015	.064	−.138	.111
	Years Smoked	−.010	.055	−.115	.100
	cigarettes/day*	−.152	.057	−.267	−.042
P_{i2}	alcohol	.121	.076	−.031	.267
	factor1	−.027	.054	−.133	.077
	factor2	−.055	.054	−.162	.048
	factor3	.074	.051	−.028	.173
	inhale	.032	.108	−.178	.247
	Intercept	2.505	.222	2.098	2.984
	age	.055	.067	−.079	.188
	Years Smoked *	.148	.060	.035	.269
	cigarettes/day	.032	.064	−.096	.156
P_{i3}	alcohol	−.003	.075	−.151	.141
	factor1*	−.116	.061	−.239	−.002
	factor2	−.109	.060	−.228	.005
	factor3	.074	.051	−.025	.174
	inhale	−.020	.117	−.248	.208
	σ_{11}	.884	.044	.802	.973
	σ_{12}	−.184	.116	−.427	.026
$\boldsymbol{\sigma}$	σ_{13}	−1.869	.178	−2.253	−1.555
	σ_{22}	1.139	.223	.768	1.638
	σ_{23}	1.193	.470	.415	2.282
	σ_{33}	4.675	.975	3.101	6.950
	ρ_{12}	−.180	.107	−.390	.027
$\boldsymbol{\rho}$	ρ_{13}	−.926	.034	−.982	−.851
	ρ_{23}	.504	.128	.235	.732

NOTE: * represents statistical significance.

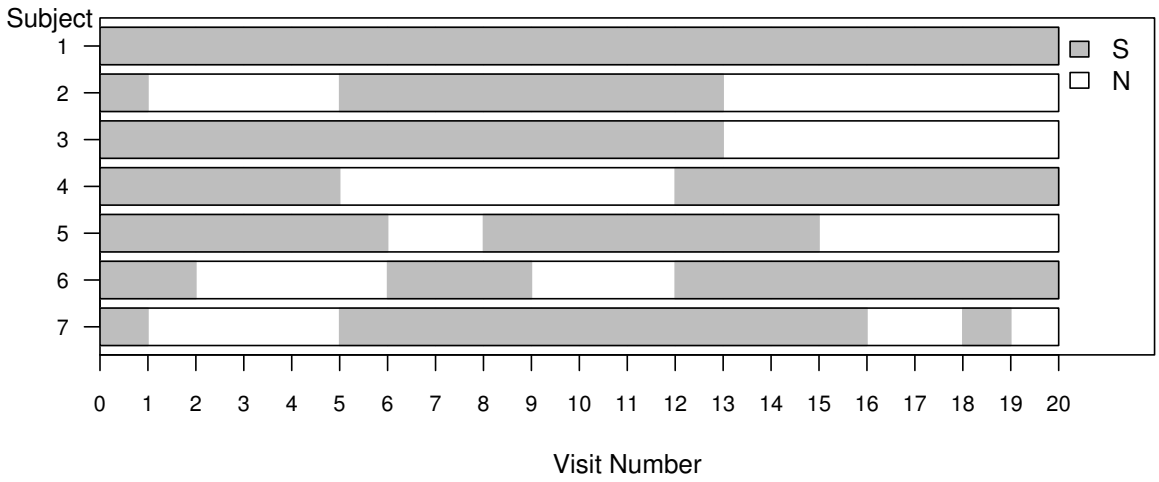

Figure 3: The smoking patterns of seven individuals in the ATBC study.

Table 2: The number of nonsmoking (NS) intervals, the quit attempts (QA), the posterior means of P_{ij}, for $j = 1, 2, 3$, and the 2-year quitting probability P_i for seven individuals displayed in Figure 3. The last row is the population means of the numbers of NS intervals and QAs and also P_{ij} and P_i of all individuals in the ATBC study

Individuals	NS	QA	P_{i1}	P_{i2}	P_{i3}	P_i
1	0	0	.014	.512	.891	.061
2	11	2	.075	.276	.356	.389
3	7	1	.026	.492	.783	.796
4	7	1	.055	.221	.424	.105
5	7	2	.066	.391	.443	.483
6	7	2	.095	.296	.260	.106
7	7	3	.123	.332	.202	.244
Population mean	1.646	.253	.021	.496	.827	.192