Published in final edited form as:

Biometrics. 2018 June ; 74(2): 636–644. doi:10.1111/biom.12792.

Bayesian variable selection for multistate Markov models with interval-censored data in an ecological momentary assessment study of smoking cessation

Matthew D. Koslovsky[1,*], **Michael D. Swartz**[1], **Wenyaw Chan**[1], **Luis Leon-Novelo**[1], **Anna V. Wilkinson**[2], **Darla E. Kendzor**[3], and **Michael S. Businelle**[3]

[1]Department of Biostatistics & Data Science, UTHealth, Houston, TX, U.S.A

[2]Department of Epidemiology, UTHealth, Austin, TX, U.S.A

[3]Department of Family and Preventive Medicine, The University of Oklahoma Health Sciences Center, Oklahoma City, OK, U.S.A

Summary

The application of sophisticated analytical methods to intensive longitudinal data, collected with ecological momentary assessments (EMA), has helped researchers better understand smoking behaviors after a quit attempt. Unfortunately, the wealth of information captured with EMAs is typically underutilized in practice. Thus, novel methods are needed to extract this information in exploratory research studies. One of the main objectives of intensive longitudinal data analysis is identifying relations between risk factors and outcomes of interest. Our goal is to develop and apply expectation maximization variable selection for Bayesian multistate Markov models with interval-censored data to generate new insights into the relation between potential risk factors and transitions between smoking states. Through simulation, we demonstrate the effectiveness of our method in identifying associated risk factors and its ability to outperform the LASSO in a special case. Additionally, we use the expectation conditional-maximization algorithm to simplify estimation, a deterministic annealing variant to reduce the algorithm's dependence on starting values, and Louis's method to estimate unknown parameter uncertainty. We then apply our method to intensive longitudinal data collected with EMA to identify risk factors associated with transitions between smoking states after a quit attempt in a cohort of socioeconomically disadvantaged smokers who were interested in quitting.

Keywords

Bayesian multistate models; continuous-time Markov process; ecological momentary assessment; EMVS; tobacco cessation

* mkoslovsky12@gmail.com.

6. Supplementary Materials

Web Appendix A & B, referenced in Section 2.3, Web Appendix C, referenced in Section 2.4, Web Appendices D-F, referenced in Section 3, and the simulation code as well as a working tutorial for our method are available with this paper at the Biometrics website on Wiley Online Library.

1. Introduction

Ecological momentary assessment (EMA) is a sampling method that allows researchers to collect a rich stream of repeated assessment data which can help determine the psychological and environmental factors associated with an individual's behavioral change, in their natural environment (Shiffman et al., 1997). EMAs capture an individual's experiences close to their occurrence at a high temporal resolution using various assessment tools, such as smart phone apps. Consequently, a larger number of moments are observed than in traditional longitudinal studies, which may provide a more accurate depiction of an individual's behavior over time. EMA data are referred to as intensive longitudinal data (Walls and Schafer, 2005). One of the main objectives of intensive longitudinal data analysis is to identify or re-affirm complex relations between risk factors and behavioral outcomes over time (Walls and Schafer, 2005).

In both intensive and traditional longitudinal studies, researchers often monitor individuals as they transition through discrete behavioral states, such as smoking status. In practice, assessments rely on compliance. As a result, assessments are sometimes missing, unequally spaced, and the exact time of transition between states is unknown (i.e., transition times are interval-censored). For traditional longitudinal studies, this type of data structure is commonly analyzed using multistate, continuous-time Markov models (MSMs) (Kay, 1986; Kalbfleisch and Lawless, 1985; Jones et al., 2006; Marshall and Jones, 1995; Saint-Pierre et al., 2003; Pan et al., 2007; Ma et al., 2015). MSMs can offer insights into behavioral processes (Saint-Pierre et al., 2003). By including covariates in these models, researchers are able to assess which risk factors are associated with an individual's transition between behavioral states (Saint-Pierre et al., 2003). For instance, in an exploratory study monitoring an individual's smoking behaviors after a planned quit date, a two-state Markov model could help identify which risk factors are associated with transitioning from a non-smoking to smoking state or from a smoking to non-smoking state. Even though MSMs are a versatile and convenient approach to analyzing traditional longitudinal data (Farewell and Tom, 2014) and are an available tool for identifying complex relations between potential risk factors and behavioral outcomes over time, they have yet to be applied to intensive longitudinal data.

The main objective of this study is to identify risk factors associated with transition between discrete smoking states using a MSM for intensive longitudinal data with interval-censoring. Currently, MSMs for interval-censored data lack an efficient, practical approach for selecting risk factors associated with transition rates. For traditional longitudinal data analyses, variable selection in MSMs has been conducted using goodness-of-fit tests and comparison methods (Marshall and Jones, 1995; Saint-Pierre et al., 2003; Pan et al., 2007; Jones et al., 2006; Aguirre-Hernández and Farewell, 2002; Farewell and Tom, 2014). This approach is suitable when the number of potential covariates is relatively small. However, for a process with k possible transitions and p potential risk factors, there are 2^{kp} possible models to compare. Additionally as k and p increase, MSMs' likelihood functions become complicated to compute and parameter estimates become unstable during estimation (Saint-Pierre et al., 2003). While larger sample sizes associated with intensive longitudinal data help mitigate parameter instability compared with traditional longitudinal data, multiple comparison methods remain impractical for variable selection in large model spaces and

inflate type I error. Model spaces are reducible by intuitively constraining regression coefficients (Marshall and Jones, 1995), however this approach is suggested for more confirmatory research settings testing hypotheses about a behavioral process as opposed to exploratory research settings when the process being modeled is less understood. Thus, the question remains as to which covariates to select for the model.

While variable selection methods for MSMs with exact transition times are available (Reulen and Kneib, 2016, 2015), no variable selection methods have been developed for MSMs with interval-censored data. Expectation maximization variable selection (EMVS), a deterministic Bayesian variable selection method inspired by stochastic search variable selection (Ročková and George, 2014; George and McCulloch, 1993), is a promising method for MSMs because it is efficient at identifying associated covariates and is capable of accommodating various outcome data structures (Ročková and George, 2014; Koslovsky et al., 2016; Zhao and Lian, 2016; McDermott et al., 2016). Since this method performs selection on all covariates simultaneously, it does not face issues of multiple comparisons, which increases modeling efficiency and controls type error rates (Gelman et al., 2014). In contrast to stochastic search variable selection, where inference is drawn from the fully sampled posterior distribution using Markov Chain Monte Carlo, EMVS simply estimates the posterior modes with the expectation maximization (EM) algorithm (Dempster et al., 1977). It is known that the EM algorithm is sensitive to starting values. By adding a deterministic annealing variant, its dependency on initial values is reduced (Ueda and Nakano, 1998). As a result, EMVS outperforms stochastic search variable selection in a fraction of the time (Ročková and George, 2014). However, efficiency gains come at a price, as EMVS lacks any intrinsically defined procedures to estimate unknown parameter variances. While there are several methods available for estimating variances when applying the EM algorithm (McLachlan and Krishnan, 2007), previous EMVS research has ignored variance estimation, only focusing on its performance at selecting associated covariates. As a result, researchers are unable to measure uncertainty in the final model when using EMVS. This limits the practicality of the method, since unbiased model interpretation relies on accurately accounting for model uncertainty (Chatfield, 2006).

In this paper, we take advantage of EMVS's validated performance in selecting covariates, efficiency, and flexibility to various data structures, by developing it for MSMs to identify relations between risk factors and smoking behaviors using interval-censored, intensive longitudinal data collected using EMA to investigate smoking cessation attempts by a cohort of 146 socioeconomically disadvantaged individuals who were interested in quitting (Kendzor et al., 2015). At each EMA, individuals responded to a set of core items regarding their cognitions, affect, behaviors, environment, as well as their smoking status (non-smoking or smoking) since the last assessment. Thus, we chose a two-state Markov model to analyze transitions between smoking states. Additionally, we provide closed-form expressions for the asymptotic variance estimates of the model's unknown parameters which incorporates parameter estimation as well as variable selection uncertainty to facilitate unbiased interpretation of the model and increase the usefulness of this method in practice. The main focus of our application is to demonstrate how the proposed method could be used to identify which risk factors, from a pool of EMA items and baseline measures, are associated with smoking transition rates after the scheduled quit attempt. This insight may

help public health researchers design effective, real-time smoking cessation interventions. By targeting individuals at high risk moments, these interventions could help decrease smoking lapse and ultimately prevent relapse.

The remaining sections of this paper are organized as follows. In Section 2, we develop EMVS with a deterministic annealing variant for a two-state, continuous-time Markov model and provide closed-form expressions for the asymptotic variance estimates of unknown parameters. In Section 3, we conduct simulation studies to assess the performance of our method. In Section 4, we use EMVS for MSMs to identify risk factors associated with transitioning between smoking states in a cohort of socioeconomically disadvantaged smokers in a smoking cessation trial. In Section 5, we provide a discussion of our method's development.

2. Methods

2.1 Model Formulation

We demonstrate how EMVS can be developed for a Bayesian MSM with interval-censored data to identify risk factors related to transitions between smoking states. We illustrate our method on a two-state, continuous-time Markov model, which coincides with our application. Let $Y_i(t_{ij})$ represent the smoking state of an individual, $i = 1, 2, \ldots m$, at a given assessment time, t_{ij}. At each time point, we observe individuals in one of two discrete states, defined as non-smoking (N) or smoking (S). Let $j = 1, \ldots, n_i$ represent the potentially unbalanced number of recurrent assessments for each individual, i. Under the assumption of a homogeneous Markov process, the transition rate matrix, $\boldsymbol{Q}$, for a two-state model is defined as (Cox and Miller, 1977):

$$\boldsymbol{Q} = \begin{array}{cc} & \begin{array}{cc} Next\ State & \\ N & S \end{array} \\ \begin{array}{c} Current \\ State \end{array} & \begin{array}{c} N \\ S \end{array} \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} \end{array},$$

where λ and μ are the positive transition rates from N $\rightarrow$ S and S $\rightarrow$ N, respectively. The transition probability matrix, $\boldsymbol{P}(\delta_{ij}) = \exp(\boldsymbol{Q}\delta_{ij})$, is defined as :

$$\boldsymbol{P}(\delta_{ij}) = \begin{array}{cc} & \begin{array}{cc} Next & State \\ N & S \end{array} \\ \begin{array}{c} Current \\ State \end{array} & \begin{array}{c} N \\ S \end{array} \begin{bmatrix} P_{NN}(\delta_{ij}) & P_{NS}(\delta_{ij}) \\ P_{SN}(\delta_{ij}) & P_{SS}(\delta_{ij}) \end{bmatrix} \end{array},$$

where $\delta_{ij} = t_{i,j} - t_{i,j-1}$. This illustrates the transition probability for an increment of time, δ_{ij}, between assessments. The transition probabilities are obtainable in closed-form, where

$$P_{NS}(\delta_{ij}) = 1 - P_{NN}(\delta_{ij}) = \frac{\lambda}{\lambda + \mu}[1 - \exp(-(\lambda + \mu)\delta_{ij})] \quad (1)$$

and

$$P_{SN}(\delta_{ij}) = 1 - P_{SS}(\delta_{ij}) = \frac{\mu}{\lambda + \mu}[1 - \exp(-(\lambda + \mu)\delta_{ij})]. \quad (2)$$

Pinsky and Karlin (2010) provide a detailed proof of the transition probabilities' derivation.

Our method is focused on identifying the relation between transition rates and a set of risk factors or covariates (e.g., negative affect, cigarette availability). Thus, we introduce individual i's observed covariates at assessment j, $\boldsymbol{x}'_{ij} = (x_{ij1}, \ldots, x_{ijp})$, into the model by redefining the transition rates λ and μ in Equations (Eq.) 1 and 2 with $\lambda = \lambda^{ij} = \exp(\lambda_0 + \boldsymbol{x}'_{ij}\boldsymbol{\beta}_\lambda)$ and $\mu = \mu^{ij} = \exp(\mu_0 + \boldsymbol{x}'_{ij}\boldsymbol{\beta}_\mu)$, similar to (Jones et al., 2006). Here, $\exp(\lambda_0)$ and $\exp(\mu_0)$ represent baseline hazard rates ($\boldsymbol{x}_{ij} = 0$) for transitioning between smoking states. Each term in the two regression coefficient vectors, $\boldsymbol{\beta}'_\lambda = (\beta_{\lambda,1}, \ldots, \beta_{\lambda,p})$ and $\boldsymbol{\beta}'_\mu = (\beta_{\mu,1}, \ldots, \beta_{\mu,p})$, is interpreted as a log-hazard ratio (Cox, 1972). This formulation allows each covariate, x_{iji}, to uniquely affect both transition rates through $\beta_{\lambda,r}$ and $\beta_{\mu,r}$. Transition rates are parameterized with an exponential form since it provides a likelihood function that has a higher chance for parameter convergence (Pan et al., 2007). We assume that covariate values remain constant between consecutive assessments, but immediately at the j^{th} assessment, the covariate value changes from the value at assessment $j - 1$ (Jones et al., 2006). If the covariates remain constant over the assessment window, we can use one $p \times 1$ vector of covariates, $\boldsymbol{x}_i$, to calculate each individual's transition rates. However, we may observe a different set of covariates at each assessment. So for each individual's $n_i - 1$ observed transitions, $n_i - 1$ different $\boldsymbol{x}_{ij}$ could be used to compute their transition rates.

Since we assume that data from different individuals are independent, the likelihood function for these data is calculated as the product of each of the m individuals' $n_i - 1$ observed transition probabilities, conditioned on their respective covariates. In this analysis, we are primarily interested in the effect of each covariate on transition rates, so we treat the probability of starting out in any state as constant, similar to (Li and Chan, 2006; Saint-Pierre et al., 2003). The likelihood function is then defined as

$$L(\boldsymbol{\beta}_\lambda, \boldsymbol{\beta}_\mu, \lambda_0, \mu_0 \mid \boldsymbol{y}) = \prod_{i=1}^{m} \prod_{j=2}^{n_i} P_{y_i(t_{i,j-1}), y_i(t_{i,j})}(\delta_{ij} \mid \boldsymbol{x}_{i,j-1}). \quad (3)$$

To illustrate: say individual i logs an assessment at times $t_{i,1}$, $t_{i,2}$,… t_{i,n_i}. At time $t_{i,1}$, we observe him/her in state N, and at time $t_{i,2}$ the individual is observed in state S. Then, the contribution to the likelihood for this individual's transition from non-smoking at time $t_{i,1}$ to smoking at time $t_{i,2}$ is represented by $P_{NS}(\delta_{i2}|\boldsymbol{x}_{i,1})$.

As for any Bayesian model's formulation, the posterior distribution is proportional to the likelihood contribution of the data multiplied by the unknown parameters' prior distributions. First, we set the log baseline hazard rates, λ_0 and μ_0, to follow a normal prior distribution with mean 0 and diffuse variance v_1. The prior distributions of the two regression coefficient vectors, $\boldsymbol{\beta}'_\lambda = (\beta_{\lambda,1}, \ldots, \beta_{\lambda,p})$ and $\boldsymbol{\beta}'_\mu = (\beta_{\mu,1}, \ldots, \beta_{\mu,p})$, regulate the variable selection procedure within EMVS. We set

$$\pi(\boldsymbol{\beta}_\lambda \mid \boldsymbol{\gamma}_\lambda, v_0, v_1) = N_p(\mathbf{0}, \boldsymbol{D}_{\boldsymbol{\gamma}_\lambda}),$$

where $\mathbf{0}$ is a p-dimensional vector of zeros, and v_0 and v_1 are pre-set variances of exclusion and inclusion, respectively (Ročková and George, 2014; George and McCulloch, 1993). Setting v_0 small drives unassociated covariate regression coefficients to zero and v_1 large allows associated covariate regression coefficients to be freely estimated. $\boldsymbol{D}_{\boldsymbol{\gamma}\lambda}$ is a $p \times p$ diagonal matrix with each D_{rr} term equal to $(1 - \gamma_{\lambda,r})v_0 + \gamma_{\lambda,r}v_1$. The prior for $\boldsymbol{\beta}_\mu$ is defined similarly. The $2p$-dimensional inclusion parameter vector, $\boldsymbol{\gamma}' = (\boldsymbol{\gamma}'_\lambda, \boldsymbol{\gamma}'_\mu)$, where $\boldsymbol{\gamma}'_\lambda = (\gamma_{\lambda,1}, \ldots, \gamma_{\lambda,p})$, $\boldsymbol{\gamma}'_\mu = (\gamma_{\mu,1}, \ldots, \gamma_{\mu,p})$, and $\boldsymbol{\gamma} \in \{0, 1\}$, is treated as missing and follows the iid Bernoulli distribution,

$$\pi(\boldsymbol{\gamma} \mid \theta) = \theta^{\sum_{r=1}^{p}(\gamma_{\lambda,r} + \gamma_{\mu,r})}(1-\theta)^{2p - \sum_{r=1}^{p}(\gamma_{\lambda,r} + \gamma_{\mu,r})},$$

where $\gamma_{\lambda,r}$ (or $\gamma_{\mu,r}$) = 1 indicates the inclusion of covariate $x_{\lambda,r}$ (or $x_{\mu,r}$) in the model. We set the prior distribution of the sparsity parameter $\theta \in [0, 1]$ to a weakly informative, conjugate $beta(a, b)$, with $a = b = 2$, to remove any boundary issues during estimation, as identified in Koslovsky et al. (2016). Note that we parameterize θ as an overall sparsity parameter for both $\gamma_{\lambda,r}$ and $\gamma_{\mu,r}$. Here, we assume that the covariates' inclusion is exchangeable, which places no restrictions on the complexity for the two transition rates, λ^{ij} and μ^{ij}. Alternative prior specifications that can accommodate structural information regarding the covariates are available (Ročková and George, 2014).

To execute our method, we iteratively determine the conditional expectation of the log posterior distribution, termed the Q-function, with respect to the conditional distribution of the missing $\boldsymbol{\gamma} | \boldsymbol{\beta}^{(k)}, \lambda_0^{(k)}, \mu_0^{(k)}, \theta^{(k)}, \mathbf{y}$ (E-step), and then maximize with respect to the parameters, $\boldsymbol{\Phi}' = (\boldsymbol{\beta}, \lambda_0, \mu_0, \theta)$, (M-step) until convergence, where $\boldsymbol{\beta}' = (\boldsymbol{\beta}'_\lambda, \boldsymbol{\beta}'_\mu)$.

2.2 E-Step

The Q-function, for iteration $k + 1$, is defined as

$$Q[\mathbf{\Phi} \mid \mathbf{\Phi}^{(k)}] = E_{\gamma \mid \cdot}\left[\log\left(\pi(\mathbf{\Phi}, \boldsymbol{\gamma} \mid \boldsymbol{y}) \mid \mathbf{\Phi}^{(k)}, \boldsymbol{y}\right)\right] = \sum_{\gamma} \log\left(\pi(\mathbf{\Phi}, \boldsymbol{\gamma} \mid \boldsymbol{y})\right) \times \pi(\boldsymbol{\gamma} \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}),$$

(4)

where $\pi(\mathbf{\Phi}, \boldsymbol{\gamma}|\boldsymbol{y})$ is the complete posterior distribution and $E_{\boldsymbol{\gamma}|\mathbf{\Phi}^{(k)},\boldsymbol{y}} = E_{\boldsymbol{\gamma}|\boldsymbol{\beta}^{(k)},\theta^{(k)}}$, which we denote as $E_{\boldsymbol{\gamma}|\cdot}$. Here, $\pi(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(k)}, \theta^{(k)})$ is the posterior probability distribution for inclusion, which is equivalent to the complete posterior divided by the observed posterior. Explicitly, Eq. 4 is defined as

$$\begin{aligned} Q[\mathbf{\Phi} \mid \mathbf{\Phi}^{(k)}] = C + \sum_{i=1}^{m}\sum_{j=2}^{n_i}\Big[\log P_{y_i(t_{i,j-1}),\, y_i(t_{i,j})}(\delta_{ij} \mid \boldsymbol{x}_{i,j-1})\Big] - \frac{1}{2v_1}(\lambda_0^2 + \mu_0^2) \\ + \sum_{r=1}^{p}\Bigg[-\frac{1}{2}\Bigg\{\beta_{\lambda,r}^2 E_{\boldsymbol{\gamma}\mid\cdot}\left[\frac{1}{v_0(1-\gamma_{\lambda,r}) + v_1\gamma_{\lambda,r}}\right] + \beta_{\mu,r}^2 E_{\boldsymbol{\gamma}\mid\cdot}\left[\frac{1}{v_0(1-\gamma_{\mu,r}) + v_1\gamma_{\mu,r}}\right]\Bigg\} \\ + E_{\boldsymbol{\gamma}\mid\cdot}[\gamma_{\lambda,r} + \gamma_{\mu,r}]\log\left(\frac{\theta}{1-\theta}\right)\Bigg] + (a-1)\log\theta + (b+2p-1)\log(1-\theta) \end{aligned}$$

where C is a constant term.

For the E-step, we evaluate the conditional expectations within the Q-function at the current iteration, k. The conditional expectation of the inclusion parameter, $E_{\boldsymbol{\gamma}|\cdot}[\gamma_{\lambda,r}]$, is defined as

$$\begin{aligned} E_{\boldsymbol{\gamma}\mid\cdot}[\gamma_{\lambda,r}] &= P(\gamma_{\lambda,r} = 1 \mid \boldsymbol{\beta}^{(k)}, \theta^{(k)}) \\ &= \frac{\pi(\beta_{\lambda,r}^{(k)} \mid \gamma_{\lambda,r} = 1) P(\gamma_{\lambda,r} = 1 \mid \theta^{(k)})}{\pi(\beta_{\lambda,r}^{(k)} \mid \gamma_{\lambda,r} = 0) P(\gamma_{\lambda,r} = 0 \mid \theta^{(k)}) + \pi(\beta_{\lambda,r}^{(k)} \mid \gamma_{\lambda,r} = 1) P(\gamma_{\lambda,r} = 1 \mid \theta^{(k)})} = p_{\lambda,r}^{*} \end{aligned}$$

where $P(\gamma_{\lambda,r} = 1|\theta^{(k)}) = \theta^{(k)}$. The other conditional expectation is the average of the precisions, $1/v_0$ and $1/v_1$, weighted by the expected probability of inclusion, $p_{\lambda,r}^{*}$,

$$E_{\boldsymbol{\gamma}\mid\cdot}\left[\frac{1}{v_0(1-\gamma_{\lambda_r}) + v_1\gamma_{\lambda_r}}\right] = (1 - p_{\lambda_r}^{*})\frac{1}{v_0} + p_{\lambda_r}^{*}\frac{1}{v_1}.$$

The conditional expectations of $E_{\boldsymbol{\gamma}|\cdot}[\gamma_{\mu,r}]$ and $E_{\boldsymbol{\gamma}\mid\cdot}\left[\frac{1}{v_0(1-\gamma_{\mu_r}) + v_1\gamma_{\mu_r}}\right]$ are defined similarly.

2.3 M-Step

When applying the EM algorithm, the maximization of the Q-function is often complicated when closed-form solutions do not exist. The expectation conditional-maximization algorithm (ECM) replaces the traditional M-step with multiple conditional maximization steps (CM-steps), conditioned on the subset of parameters being estimated (Meng and Rubin, 1993). This common, alternative approach simplifies and stabilizes maximization, because the Q-function is maximized over a lower dimension of parameters (Meng and Rubin, 1993). Even after conditioning each maximization step, closed-form solutions often are still unobtainable. Thus, researchers rely on iterative procedures, including the Newton-Raphson algorithm. Such is the case for our method. We define the CM-steps as follows:

CM-step 1: Obtain $\lambda_0^{(k+1)}$ and $\boldsymbol{\beta}_\lambda^{(k+1)}$ by maximizing Eq. 4, conditioned on $\mu_0^{(k)}, \boldsymbol{\beta}_\mu^{(k)}$, and $\theta^{(k)}$ using one step of the Newton-Raphson algorithm.

CM-step 2: Obtain $\mu_0^{(k+1)}$ and $\boldsymbol{\beta}_\mu^{(k+1)}$ by maximizing Eq. 4, conditioned on $\lambda_0^{(k+1)}, \boldsymbol{\beta}_\lambda^{(k+1)}$, and $\theta^{(k)}$ using one step of the Newton-Raphson algorithm.

CM-step 3: Update the estimate of θ with the closed-form solution,

$$\theta^{(k+1)} = \frac{\sum_{r=1}^{p} (p^*_{\lambda,r} + p^*_{\mu,r}) + a - 1}{a + b + 2p - 2}.$$

The ECM algorithm stops when the absolute value of the difference between the log-likelihood distribution evaluated at the current and next step of the algorithm falls below a set threshold (Wu, 1983). See Web Appendix A for details regarding the convergence stopping rule. Once the algorithm has converged, the final estimates, $\hat{\boldsymbol{\Phi}}$, maximize Eq. 4. Inclusion is determined if $E_{\boldsymbol{\gamma}|\hat{\boldsymbol{\Phi}}}[\gamma_r] \geq 0.5$ (Ročková and George, 2014). In practice, the performance of the EM algorithm is sensitive to initialization, and convergence is not guaranteed at the global mode. Thus, we use a deterministic annealing variant to reduce the algorithm's dependence on initialization, similar to (Ročková and George, 2014; Koslovsky et al., 2016). See Web Appendix B for details of the deterministic annealing variant's formulation.

2.4 Variance Estimation

To estimate variances of the unknown parameters, $\boldsymbol{\Phi}$, in our proposed method, we use Louis's method (Louis, 1982), which relies on the missing information principle (Orchard et al., 1972),

$$\text{Observed Information} = \text{Complete Information} - \text{Missing Information}.$$

Louis's method formulates the observed information matrix in terms of the second derivative of the Q-function and the variance of the first derivative of the posterior distribution with respect to the missing information, $\boldsymbol{\gamma}$. Following the Bayesian central limit theorem, the posterior distribution of the unknown parameters can be estimated assuming a normal

distribution with mean equal to the posterior mode and variance equal to the inverse observed information matrix (Carlin and Louis, 2008). The estimated variance of the unknown parameters is expressed as

$$\widehat{\text{Var}}(\widehat{\boldsymbol{\Phi}}) = \frac{1}{I_{obs}(\widehat{\boldsymbol{\Phi}})} = \left[-\frac{\partial^2 Q\left[\boldsymbol{\Phi}\middle|\widehat{\boldsymbol{\Phi}}\right]}{\partial \boldsymbol{\Phi}\, \partial \boldsymbol{\Phi}'} - \text{var}\left\{ \frac{\partial \log \pi(\boldsymbol{\Phi}, \boldsymbol{\gamma} \mid \boldsymbol{y})}{\partial \boldsymbol{\Phi}} \middle| \widehat{\boldsymbol{\Phi}}, \boldsymbol{y} \right\} \right]^{-1}. \quad (5)$$

Details of this derivation are found in Web Appendix C. To avoid any boundary issues when calculating 95% credible intervals for $\theta \in [0, 1]$, we apply a logit transformation to the posterior mode and assume it follows a normal distribution with mean $logit(\hat{\theta})$ and variance $\widehat{\text{Var}}(\hat{\theta})/(\hat{\theta}(1-\hat{\theta}))^2$.

3. Simulation Study

To evaluate the performance of our method, we apply it to multiple simulated data sets in a variety of research scenarios. Details of the data generation, evaluation methods, and results of the simulation study can be found in Web Appendices D – F, respectively. Briefly, we examined the performance of our method in various scenarios, with different sample sizes ($m = 100$ and $m = 150$), numbers of equally (randomly) spaced assessment times ($n_i = 30$ or $n_i = 70$), and exchangeable correlation structures between covariates ($\rho = 0$ and $\rho = 0.75$). For the special case of equally spaced assessment times, we compared our model to EMVS and the LASSO for logistic regression models (Koslovsky et al., 2016; Tibshirani, 1996).

We evaluated the performance of our method based on the average false positive (FP) and false negative (FN) rates with FPR = FP/(FP + TN) and FNR= FN/(FN+TP), where TP and TN are true positives and true negatives, respectively. Additionally, we assessed the bias (average of the posterior modes minus true values), the Monte Carlo error of the posterior modes (MCE), the square root of the average of the posterior variances estimated with Louis's method (SE), the coverage probability (CP) of the 95% equal-tail credible intervals, and the average mean squared error of the steady-state probability of transition from a non-smoking to smoking state (MSE).

We found the performance of our method with both randomly and equally spaced assessment times improved with larger sample sizes, larger number of assessments observed, and lower correlation structures. With equally spaced assessment times, our method outperformed or showed relatively equivalent performance to the LASSO for FPR and FNR in every setting and comparable performance to EMVS for logistic regression models. Additionally in all scenarios, our method correctly included associated covariates and correctly excluded unassociated covariates in about 99% of the simulations on average. Also, CPs fell around 92% on average. As the sample size and number of assessments increased, we observed the MCE approach the SE. The MSE for the steady-state probability of transition from a non-smoking to a smoking state was around 0.04 for all scenarios. Overall, our method demonstrated encouraging performance across the simulation scenarios,

justifying its use for identifying risk factors associated with transitions between smoking states in our application data.

4. Application

Our method was developed to analyze intensive longitudinal data collected with EMA from the PREVAIL study (Kendzor et al., 2015), which demonstrated the effectiveness of a contingency management (CM) treatment to promote smoking cessation. At the beginning of the study, 146 of 222 screened individuals met the eligibility requirements and were randomized into treatment groups. One group received usual smoking cessation care from a Dallas based, safety-net hospital ($n = 71$), and the other received usual care as well as the contingency management ($n = 75$), which offered small financial incentives to encourage abstinence. A week before the scheduled quit date, baseline measures were taken and individuals were taught how to complete assessments on a study provided smart phone. Each individual logged his/her smoking behaviors on the smart phone over a 2-week period (1 week prior and 1 week after the scheduled smoking quit date). Individuals were prompted with 4 random assessments per day, which collected information regarding their urge to smoke, affect, social environment, abstinence self-efficacy, cigarette availability, and location. Since assessment times were randomly prompted by the smart phone, they were considered non-informative to the non-smoking/smoking process (Gruger et al., 1991).

As mentioned, our main objective in this analysis is to identify risk factors associated with transitioning between smoking and non-smoking states after the scheduled quit date in this cohort. An individual's smoking status was deemed to be in a smoking state if they reported smoking since their previous assessment. Potential risk factors consist of both baseline measurements and EMA items (Table 1). Related analyses have summarized positive and negative affect items by taking the average of each set of items (Businelle et al., 2014). Here, we are interested in selecting individual components of positive affect (e.g., happy, calm) and negative affect (e.g., irritable, frustrated/angry, sad, worried, miserable) measures. Four individuals were dropped from the analysis because responses to the set of items in Table 1 were missing. A total of 3091 assessments collected after quit date (on average 41 per individual ranging from 3 to 51) were analyzed.

Before performing variable selection, we assessed the feasibility of a MSM for the application data. Since assessments were collected frequently, we determined the MSM's overall fit by plotting the observed and estimated prevalences, obtained by fitting a full model, of each state over time (Titman and Sharples, 2010). We tested the assumption of time homogeneity for the Markov process by comparing the full model to an alternative model with piecewise constant transition intensities using a likelihood ratio test. We tested the Markov assumption by comparing the full model to an alternative model that included the state occupied two assessments prior as a covariate using a likelihood ratio test. Throughout the observation window, the observed and estimated smoking and non-smoking prevalences fell around 20% and 80%, respectively (Web Appendix Figure 1). Additionally, we failed to reject the null hypotheses that the baseline hazards were constant and the Markov assumption was upheld at the 0.05 α-level with p-values of 0.17 and 0.42, respectively. To perform variable selection on the data, we initialized and parameterized our

model similar to the methods found in Web Appendix E. The variance of exclusion and inclusion were set to $v_0 = 0.0006$ and $v_1 = 0.5$, respectively. Continuous covariates were standardized to mean 0 and variance 1 before selection. Covariates were included in the model if the conditional expectation of their respective inclusion indicator was greater than or equal to 0.50. In this analysis, all included covariates had an $E_{\boldsymbol{\gamma}|\hat{\boldsymbol{\Phi}}}[\gamma_r] = 1$, and all excluded covariates had an $E_{\boldsymbol{\gamma}|\hat{\boldsymbol{\Phi}}}[\gamma_r] < 0.06$.

We present results based on each risk factor's inclusion or exclusion via EMVS, however not all risk factor's remained influential (95% CI for hazard ratio contains 1) after accounting for estimation and selection uncertainty (Table 2). We found results that were consistent with previous research analyzing the relation between risk factors and smoking behaviors after a quit attempt. CM was previously shown to be an effective means of increasing smoking abstinence after a quit attempt in this cohort (Kendzor et al., 2015). In this analysis, we found that CM was associated with a decrease in the transition rate from N → S after the quit date. Addiction level has previously been associated with relapse (Zhou et al., 2009). In this analysis, the Heaviness of Smoking Index (HSI) served as a proxy for addiction level and was found to be associated with a decrease in transition rates from N → S and S → N. Consistent with Zhou et al. (2009), we found that baseline education level was not associated with relapse. This analysis also did not find any association with transition rates. Age has been shown to be associated with a decrease in the odds of relapse (Zhou et al., 2009). We found that age reduced the transition rate from N → S and S → N. Also, environmental factors, such as having cigarettes available and being around someone who is smoking, have been associated with smoking behaviors (Zhou et al., 2009). Here, having cigarettes available increased the transition rate from N → S and decreased the transition from S → N. Negative and positive affect as well as urge to smoke are commonly identified as risk factors associated with smoking lapse and relapse after a quit attempt (Piasecki, 2006; Vasilenko et al., 2014; Shiffman et al., 2002; Zhou et al., 2009). In our analysis, we found that the being calm (a positive affect item) and worried (a negative affect item) were associated with a reduction in both transitions after the quit attempt. While urge to smoke is considered a defining characteristic of addiction (Kassel and Shiffman, 1992; Shiffman et al., 1997), its association with smoking behaviors is often inconsistent (Wray et al., 2013). Here, we found that urge was associated with an increase in transition between N → S and S → N after the quit date. Self-efficacy to abstain is commonly shown to be associated with smoking behaviors around a quit attempt (Smit et al., 2014; Shiffman et al., 2000). This analysis identified self-efficacy as being associated with a decrease in transition rate from N → S. Additionally, being in a car has been associated with a reduction in the odds of smoking (Shiffman et al., 2002). We found it to be associated with a decrease in both transitions rates. During ad-lib smoking, being at work and being outside have been associated with a decrease and increase in smoking, respectively. However in this study, being at work was not found to be associated with any transition, but being outside was associated with an decrease in transition from N → S and S → N. For two of the risk factors (self-efficacy to abstain and having cigarettes available), our method was able to differentiate between a risk factor's relation with transitioning from S → N and N → S. These results demonstrate how EMVS for a MSM can reveal intricacies in complex behavioral processes that may elude other methods.

5. Discussion

To our knowledge, we developed the first variable selection method for MSMs with interval-censored data. Using EMA data, we demonstrated the usefulness of our method in practice by identifying potential risk factors associated with transitions between discrete smoking states in a cohort of socioeconomically disadvantaged individuals. In future studies, this method could be used to identify multiple predictor variables for lapse in real-time that could trigger the delivery of tailored interventions at the critical time after a quit attempt.

In this work, we show the usefulness of a variable selection method on a two-state Markov model, but the method is generalizable to other state spaces. However, a major challenge of modeling MSMs is model estimation when the number of potential transitions in the state space and covariates increases (Saint-Pierre et al., 2003). Extending our method to MSMs with larger state spaces would require adjusting the likelihood function component in the Q-function. For three- and four-state models, we conjecture that estimation times would not increase significantly, since closed-formed solutions exist for the transition probabilities (Li and Chan, 2006; Chan, 2017). Thereafter, we expect computational cost to depend more on the optimization routine employed. In practice, researchers often ignore interval-censoring and assume that exact transition times are known, which may bias parameter estimates (Sutradhar et al., 2010). Therefore, variable selection methods designed for datasets in which the exact transition time are known are not appropriate for analyzing data structures found in this study. However, future work could incorporate the attractive features of these methods, including selection of non-linear covariate effects into the EMVS framework (Reulen and Kneib, 2015, 2016).

One of the main objectives of intensive longitudinal data analysis is to identify or re-affirm complex relations between potential risk factors and behavioral outcomes over time (Walls and Schafer, 2005). While EMVS for MSMs with interval-censored data shows promise for identifying these relations, no variable selection method is a panacea. In practice, we suggest using our method coupled with other intensive longitudinal data analyses approaches (Walls and Schafer, 2005; Tan et al., 2012) to provide a deeper perspective on the intricacies of the behavioral process.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research is supported by the University of Texas School Health Science at Houston Center School of Public Health, Cancer Education and Career Development Program National Cancer Institute/NIH Grant R25 CA57712 predoctoral fellowship to Matthew D. Koslovsky; the University of Texas Health Science Center at Houston School of Public Health, Training Program in Biostatistics National Institute of General Medical Sciences/NIH Grant T32GM074902 predoctoral traineeship to Matthew D. Koslovsky; and the Michael & Susan Dell Foundation, Michael & Susan Dell Center for Healthy Living, The University of Texas School of Public Health, Austin Regional Campus. The parent study was primarily supported by the University of Texas Health Science Center, School of Public Health with additional support from American Cancer Society Grants MRSGT-10-104-01-CPHPS to Darla E. Kendzor and MRSGT-12-114-01-CPPB to Michael S. Businelle. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

References

Aguirre-Hernández R, Farewell V. A Pearson-type goodness-of-fit test for stationary and time-continuous Markov regression models. Statistics in Medicine. 2002; 21:1899–1911. [PubMed: 12111896]

Businelle MS, Ma P, Kendzor DE, Reitzel LR, Chen M, Lam CY, Bernstein I, Wetter DW. Predicting quit attempts among homeless smokers seeking cessation treatment: an ecological momentary assessment study. Nicotine & Tobacco Research. 2014; 16:1371–1378. [PubMed: 24893602]

Carlin, BP., Louis, TA. Bayesian Methods for Data Analysis. CRC Press; Boca Raton, FL: 2008.

Chan, W. Appendix: Derivations of transition probabilities for a four-state continuous time Markov chain. 2017. https://drive.google.com/open?id=0BAhiwb6HTQzTXBuWmo3dldKRE0

Chatfield, C. Journal of the Royal Statistics Society A158. 1995. Model uncertainty, data mining and statistical inference; p. 419-466.

Cox DR. Regression models and life-tables. Journal of the Royal Statistics Society. 1972; B34:187–220.

Cox, DR., Miller, HD. The Theory of Stochastic Processes. Vol. 134. CRC Press; Boca Raton, FL: 1977.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological). 1977; 39:1–38.

Farewell VT, Tom BD. The versatility of multi-state models for the analysis of longitudinal data with unobservable features. Lifetime Data Analysis. 2014; 20:51–75. [PubMed: 23225140]

Gelman, A., Carlin, JB., Stern, HS., Rubin, DB. Bayesian Data Analysis. 3. Taylor & Francis; Boca Raton, FL: 2014.

George EI, McCulloch RE. Variable selection via Gibbs sampling. Journal of the American Statistical Association. 1993; 88:881–889.

Gruger J, Kay R, Schumacher M. The validity of inferences based on incomplete observations in disease state models. Biometrics. 1991; 47:595–605. [PubMed: 1912263]

Jones RH, Xu S, Grunwald GK. Continuous time Markov models for binary longitudinal data. Biometrical Journal. 2006; 48:411–419. [PubMed: 16845905]

Kalbfleisch J, Lawless JF. The analysis of panel data under a Markov assumption. Journal of the American Statistical Association. 1985; 80:863–871.

Kassel JD, Shiffman S. What can hunger teach us about drug craving? a comparative analysis of the two constructs. Advances in Behaviour Research and Therapy. 1992; 14:141–167.

Kay R. A Markov model for analysing cancer markers and disease states in survival studies. Biometrics. 1986; 14:855–865.

Kendzor DE, Businelle MS, Poonawalla IB, Cuate EL, Kesh A, Rios DM, Ma P, Balis DS. Financial incentives for abstinence among socioeconomically disadvantaged individuals in smoking cessation treatment. American Journal of Public Health. 2015; 105:1198–1205. [PubMed: 25393172]

Koslovsky MD, Swartz MD, Leon-Novelo L, Chan W, Wilkinson AV. Using the EM algorithm for Bayesian variable selection in logistic regression models with related covariates. 2016 In Revisions.

Li YP, Chan W. Analysis of longitudinal multinomial outcome data. Biometrical Journal. 2006; 48:319–326. [PubMed: 16708781]

Louis TA. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological). 1982; 44:226–233.

Ma J, Chan W, Tsai CL, Xiong M, Tilley BC. Analysis of transtheoretical model of health behavioral changes in a nutrition intervention study-a continuous time Markov chain model with Bayesian approach. Statistics in Medicine. 2015; 34:3577–3589. [PubMed: 26123093]

Marshall G, Jones RH. Multi-state models and diabetic retinopathy. Statistics in Medicine. 1995; 14:1975–1983. [PubMed: 8677398]

McDermott P, Snyder J, Willison R. Methods for Bayesian variable selection with binary response data using the EM algorithm. 2016 arXiv preprint arXiv:1605.05429.

McLachlan, G., Krishnan, T. The EM Algorithm and Extensions. 2. John Wiley & Sons; Hoboken, NJ: 2007.

Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika. 1993; 80:267–278.

Orchard, T., Woodbury, MA. A missing information principle: theory and applications. Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability; Berkeley, CA: University of California Press; 1972. p. 697-715.

Pan SL, Wu HM, Yen AMF, Chen THH. A Markov regression random-effects model for remission of functional disability in patients following a first stroke: A Bayesian approach. Statistics in Medicine. 2007; 26:5335–5353. [PubMed: 17676712]

Piasecki TM. Relapse to smoking. Clinical Psychology Review. 2006; 26:196–215. [PubMed: 16352382]

Pinsky, M., Karlin, S. An Introduction to Stochastic Modeling. Academic Press; Burlington, MA: 2010.

Reulen, H., Kneib, T. Technical report. University of Goettingen; 2015. Structured fusion lasso penalised multi-state models.

Reulen H, Kneib T. Boosting multi-state models. Lifetime Data Analysis. 2016; 22:241–262. [PubMed: 25990764]

Ročková V, George EI. EMVS: The EM approach to Bayesian variable selection. Journal of the American Statistical Association. 2014; 109:828–846.

Saint-Pierre P, Combescure C, Daures J, Godard P. The analysis of asthma control under a Markov assumption with use of covariates. Statistics in Medicine. 2003; 22:3755–3770. [PubMed: 14673936]

Shiffman S, Balabanis MH, Paty JA, Engberg J, Gwaltney CJ, Liu KS, Gnys M, Hickcox M, Paton SM. Dynamic effects of self-efficacy on smoking lapse and relapse. Health Psychology. 2000; 19:315. [PubMed: 10907649]

Shiffman S, Engberg JB, Paty JA, Perz WG, Gnys M, Kassel JD, Hickcox M. A day at a time: predicting smoking lapse from daily urge. Journal of Abnormal Psychology. 1997; 106:104. [PubMed: 9103722]

Shiffman S, Gwaltney CJ, Balabanis MH, Liu KS, Paty JA, Kassel JD, Hickcox M, Gnys M. Immediate antecedents of cigarette smoking: an analysis from ecological momentary assessment. Journal of Abnormal Psychology. 2002; 111:531. [PubMed: 12428767]

Shiffman S, Hufford M, Hickcox M, Paty JA, Gnys M, Kassel JD. Remember that? A comparison of real-time versus retrospective recall of smoking lapses. Journal of Consulting and Clinical Psychology. 1997; 65:292. [PubMed: 9086693]

Smit ES, Hoving C, Schelleman-Offermans K, West R, de Vries H. Predictors of successful and unsuccessful quit attempts among smokers motivated to quit. Addictive Behaviors. 2014; 39:1318–1324. [PubMed: 24837754]

Sutradhar R, Barbera L, Seow H, Howell D, Husain A, Dudgeon D. Multistate analysis of interval-censored longitudinal data: Application to a cohort study on performance status among patients diagnosed with cancer. American Journal of Epidemiology. 2010; 173:384.

Tan X, Shiyko MP, Li R, Li Y, Dierker L. A time-varying effect model for intensive longitudinal data. Psychological Methods. 2012; 17:61. [PubMed: 22103434]

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological). 1996; 58:267–288.

Titman AC, Sharples LD. Model diagnostics for multi-state models. Statistical Methods in Medical Research. 2010; 19:621–651. [PubMed: 19654169]

Ueda N, Nakano R. Deterministic annealing EM algorithm. Neural Networks. 1998; 11:271–282. [PubMed: 12662837]

Vasilenko SA, Piper ME, Lanza ST, Liu X, Yang J, Li R. Time-varying processes involved in smoking lapse in a randomized trial of smoking cessation therapies. Nicotine & Tobacco Research. 2014; 16:S135–S143. [PubMed: 24711627]

Walls, TA., Schafer, JL. Models for intensive longitudinal data. Oxford University Press; New York, NY: 2005.

Wray JM, Gass JC, Tiffany ST. A systematic review of the relationships between craving and smoking cessation. Nicotine & Tobacco Research. 2013; 15:1167–1182. [PubMed: 23291636]

Wu CJ. On the convergence properties of the EM algorithm. The Annals of Statistics. 1983; 11:95–103.

Zhao K, Lian H. The Expectation–Maximization approach for Bayesian quantile regression. Computational Statistics & Data Analysis. 2016; 96:1–11.

Zhou X, Nonnemaker J, Sherrill B, Gilsenan AW, Coste F, West R. Attempts to quit smoking and relapse: factors associated with success or failure from the ATTEMPT cohort study. Addictive Behaviors. 2009; 34:365–373. [PubMed: 19097706]

Biometrics. Author manuscript; available in PMC 2018 June 27.

Table 1

Description of potential risk factors for transitioning between smoking states.

Baseline Measures			
Measure		**Scale**	**Coded As**
Heaviness of Smoking Index (HSI)		0–6	Continuous
Education level completed		Years	Continuous
Age		Years	Continuous
Race/Ethnicity		Non-Hispanic White or Black/Other	Binary indicator with “Non-Hispanic White” as reference.
Contingency Management (CM)		Yes or No	Binary indicator with “No” as reference
EMA Items			
Item Type	Item	Assessment Scale	Coded As
Urge to smoke	“I have an urge to smoke”	1–5 Likert	Continuous
Affect	“I feel happy”	1–5 Likert	Continuous
	“I feel calm”	1–5 Likert	Continuous
	“I feel irritable”	1–5 Likert	Continuous
	“I feel frustrated/angry”	1–5 Likert	Continuous
	“I feel sad”	1–5 Likert	Continuous
	“I feel worried”	1–5 Likert	Continuous
	“I feel miserable”	1–5 Likert	Continuous
Social Setting	“Is anyone you are, interacting with smoking?” **	Yes or No	Binary indicator with “No” as reference
Abstinence Self-efficacy	“I am confident in my ability to AVOID smoking”	1–5 Likert	Continuous
Cigarette Availability	“Cigarettes are available to me”	1–5 Likert	Continuous
Location	Being outside	Yes or No	Binary indicator with “Not outside” as reference
	In a car/truck	Yes or No	Binary indicator with “Not in a car or truck” as reference
	At work	Yes or No	Binary indicator with “Not at work” as reference

** Prompted if individual answered “Yes” to “Are you interacting with people?”

Table 2

Application Results

Hazard rates and 95% equal-tail credible intervals (CI) of potential risk factors for transitioning between N → S and S → N states after the scheduled quit attempt.

Risk Factor	After Quit Attempt	
	N → S (95% CI)	S → N (95% CI)
Baseline hazard	0.081 (0.062, 0.107)	0.321 (0.248, 0.416)
HSI	0.727 (0.587, 0.901) **	0.796 (0.623, 1.016) *
Education level	0.976 (0.930, 1.023)	1.010 (0.965, 1.057)
Age	0.673 (0.535, 0.845) **	0.568 (0.447, 0.723) **
Race/Ethnicity	1.002 (0.954, 1.052)	0.995 (0.948, 1.045)
CM	0.569 (0.452, 0.717) **	1.002 (0.954, 1.053)
Urge	1.565 (1.231, 1.988) **	1.264 (0.997, 1.602) *
Happy	1.000 (0.954, 1.048)	0.998 (0.952, 1.046)
Calm	0.615 (0.477, 0.791) **	0.653 (0.517, 0.824) **
Irritable	0.995 (0.949, 1.044)	1.002 (0.955, 1.050)
Frustrated	0.999 (0.952, 1.047)	0.993 (0.947, 1.041)
Sad	1.002 (0.956, 1.051)	0.995 (0.950, 1.043)
Worried	0.503 (0.381, 0.665) **	0.532 (0.409, 0.692) **
Miserable	1.007 (0.960, 1.056)	0.987 (0.942, 1.035)
Interacting w/ smoker	1.001 (0.954, 1.052)	0.998 (0.950, 1.048)
Self-efficacy	0.711 (0.631, 0.802) **	0.998 (0.950, 1.047)
Cigarettes available	1.446 (1.151, 1.819) **	0.772 (0.613, 0.972) **
Being outside	0.572 (0.357, 0.917) **	0.293 (0.180, 0.476) **
In a car/truck	0.780 (0.455, 1.338) *	0.527 (0.296, 0.938) **
At work	0.999 (0.951, 1.049)	1.001 (0.953, 1.051)

** Risk factor selected by EMVS and CI does not contain hazard ratio equal to 1

* Risk factor selected by EMVS and CI does contain hazard ratio equal to 1