Received: 10 August 2023 | Revised: 17 June 2024 | Accepted: 28 June 2024

DOI: 10.1002/sim.10171

RESEARCH ARTICLE

Statistics in Medicine WILEY

A latent variable approach to jointly modeling longitudinal and cumulative event data using a weighted two-stage method

Madeline R. Abbott[1] | **Inbal Nahum-Shani**[2] | **Cho Y. Lam**[3] | **Lindsey N. Potter**[3] | **David W. Wetter**[3] | **Walter H. Dempsey**[1]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA

[2]Institute for Social Research, University of Michigan, Ann Arbor, Michigan, USA

[3]Department of Population Health Sciences and Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah, USA

Correspondence
Madeline R. Abbott, Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA.
Email: mrabbott@umich.edu

Funding information
National Institute on Drug Abuse, Grant/Award Numbers: F31DA057048, P50DA054039, R01DA039901; National Center on Minority Health and Health Disparities, Grant/Award Number: R01MD010362; National Cancer Institute, Grant/Award Numbers: P30CA042014, U01CA229437; Huntsman Cancer Foundation

Ecological momentary assessment (EMA), a data collection method commonly employed in mHealth studies, allows for repeated real-time sampling of individuals' psychological, behavioral, and contextual states. Due to the frequent measurements, data collected using EMA are useful for understanding both the temporal dynamics in individuals' states and how these states relate to adverse health events. Motivated by data from a smoking cessation study, we propose a joint model for analyzing longitudinal EMA data to determine whether certain latent psychological states are associated with repeated cigarette use. Our method consists of a longitudinal submodel—a dynamic factor model—that models changes in the time-varying latent states and a cumulative risk submodel—a Poisson regression model—that connects the latent states with the total number of events. In the motivating data, both the predictors—the underlying psychological states—and the event outcome—the number of cigarettes smoked—are partially unobservable; we account for this incomplete information in our proposed model and estimation method. We take a two-stage approach to estimation that leverages existing software and uses importance sampling-based weights to reduce potential bias. We demonstrate that these weights are effective at reducing bias in the cumulative risk submodel parameters via simulation. We apply our method to a subset of data from a smoking cessation study to assess the association between psychological state and cigarette smoking. The analysis shows that above-average intensities of negative mood are associated with increased cigarette use.

KEYWORDS
dynamic factor model, ecological momentary assessment, joint model, smoking cessation, two-stage estimation

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2024 The Author(s). *Statistics in Medicine* published by John Wiley & Sons Ltd.

1 | INTRODUCTION

Widespread use of mobile health (mHealth) technology, ranging from smartphones to wearable devices, has increased the availability of intensive longitudinal data (ILD). Ecological momentary assessment (EMA), which is a data collection method often used in mHealth studies, consists of repeatedly sampling individuals' current states and contexts. This method allows for the collection of data in real time and in individuals' natural environments. When outcomes of interest are measured multiple times per day, researchers can record rich data that capture temporal variations in individuals' current states and contexts. These EMA data can be particularly useful in helping researchers understand what factors (eg, psychological states or environmental stimuli) represent risk for adverse health events (eg, smoking, sedentarism, unhealthy eating). We consider the setting in which researchers are interested in examining the association between multiple dynamic latent factors—which are measured through a larger number of observed longitudinal outcomes—and the risk of repeated adverse health events, where these events are indirectly observed as the total number of events across some interval of time.

In many research areas, longitudinal outcomes are measured with error and not necessarily at the same time as the adverse health events. Thus, simply including the longitudinal outcomes as time-varying predictors in a time-to-event model, such as a Cox proportional hazards model, is not appropriate.[1] A more suitable standard alternative is to jointly fit a longitudinal and time-to-event model. However, existing joint longitudinal-survival models are limited in their ability to flexibly model multiple longitudinal outcomes in a computationally efficient manner. This computational challenge is of particular concern in the setting of ILD when many longitudinal outcomes are measured simultaneously.

Additionally, rather than recording the timing of each individual adverse health event, the mHealth study motivating this work was designed such that we only observe cumulative numbers of events over windows of time, further complicating our setting. Here, we take an important first step to address these challenges—stemming from both the partial observation of the event outcome and the computational complexity due to ILD—by developing a practically useful approach for modeling the association between multiple longitudinal factors and the risk of repeated adverse health events. Our approach aims to take advantage of existing software while at the same time balance flexibility, computational complexity, and scientific utility.

1.1 | Related work

Much of the existing work within the field of joint longitudinal-survival models has focused on combining univariate longitudinal processes with a single time-to-event outcome. Modeling multivariate longitudinal outcomes can rapidly increase the computational cost of joint estimation. Aiming to address the computational challenges of joint modeling with multivariate longitudinal data, Rustand et al[2] propose an approximate Bayesian method based on the integrated nested Laplace approximation. Adaptation of this technique to fit our specific model is less than straightforward, particularly due to our use of a multivariate continuous-time stochastic process. An alternative sampling-based approach to modeling time-varying covariates and survival outcomes was proposed by Rathbun et al.[3] This work was motivated by a desire to improve computational efficiency and avoid specification of a model for the longitudinal process; however, it requires many simplifying assumptions—such as lack of measurement error—that do not generally hold with use of EMAs in practice.

An alternative strategy to joint estimation of a longitudinal-survival model for multiple longitudinal outcomes, a two-stage approach is commonly used. Two stage estimation, however, has the well-documented risk of introducing bias into estimates of model coefficients.[1,4,5] Incorporating weights has been shown to help reduce this bias. For example, Mauff et al[5] recently proposed a two-stage method that leverages weights based on importance sampling within a Bayesian framework.

To lower the high computational cost associated with directly modeling the observed multivariate longitudinal outcomes, we reduce complexity by modeling a smaller number of latent longitudinal factors via a dynamic factor model. Factor models have long been used in the fields of psychology and behavioral science to represent highly correlated outcomes as simpler underlying states, where these states can be interpreted as affect, motivation, vulnerability, and so forth.[6-8] Numerous variations of latent variable models have been proposed in the statistical literature, many of which involve modeling the evolution of the latent factors over time.[9-18] Extending this dimension-reduction strategy to the joint model setting, McCurdy et al,[19] Liu et al,[20] and Larsen[21] have proposed combining factor models and hazard

models to jointly model longitudinal and time-to-event data. Liu et al[20] modeled the evolution of the latent factors using mixed models while McCurdy et al[19] and Larsen[21] did not model correlation over time. Muthén and Muthén[22] and Asparouhov and Muthén[23] have implemented methods to fit models like that proposed by Larsen[21] in their software Mplus but suggest an approach to estimation that relies on numerical integration. We avoid the computational complexity of numerical integration by using a two-stage approach. Furthermore, our use of a lower-dimensional stochastic process to model the evolution of the latent factors falls beyond the scope of the longitudinal models considered in these prior works.

Importantly, our proposed approach differs from the standard joint model setting due to the partially unobservable nature of our event outcome. Rather than observing the time of each event (or censoring), we observe total numbers of events over defined windows of time. As a result, our joint model consists of a longitudinal submodel and an event submodel suitable for this count data—specifically, a Poisson regression model.

1.2 | Main contributions and outline

Overall, there exists a need for further development of statistical methods well-suited for ILD that can connect multiple longitudinal outcomes with the risk of adverse events in a flexible and interpretable manner. Rather than attempting to jointly estimate the longitudinal and risk submodels in this work, we take a two-stage estimation approach, using weights to alleviate potential bias that can result from non-joint estimation. Our main contribution is the development of a scientifically useful approach to modeling multiple highly variable longitudinal outcomes and their associations with recurrent events; we do so by combining a dynamic factor model and a Poisson regression model for cumulative risk via an approximate method that leverages existing software in a practically useful way.

The remainder of this paper is organized as follows: in Section 2, we describe the motivating data collected as part of a smoking cessation study that uses EMAs; in Section 3, we outline our proposed method; in Section 4, we demonstrate its performance via simulation; in Section 5, we use our method to analyze a subset of data recorded in the smoking cessation study; and in Section 6, we discuss our results and provide concluding remarks.

2 | MOTIVATING DATA

Data motivating this work come from a longitudinal study examining the influence of contextual, biobehavioral/psychosocial, and demographic factors; social history; and acute momentary precipitants on smoking cessation among 302 African Americans attempting to quit (Break Free II; R01MD010362). Participants answered EMAs on a pre-programmed smartphone from 4 days prior to their quit date through 10 days post-quit. Smartphones were programmed to deliver up to four random EMAs per day. Participants also answered event-contingent EMAs based on whether wearable chest and wrist sensors detected possible smoking or stress.[24,25] Each EMA assessed smoking behaviors, mood, and other cognitive, interpersonal, and contextual factors. All participants received nicotine patch therapy, self-help materials, and brief quitting advice[26] and were compensated for their time. Detailed information about study procedures have been published in Potter et al.[27] Our work uses data that were available at the time of drafting this manuscript from 218 participants and focuses on their psychological state (measured via emotion items) and smoking (measured via the number of cigarettes recently smoked) self-reported using EMAs over the 10 day post-quit period. The total number of random EMAs to which individuals responded over the post-quit period varied from 2 to 47 (average = 17). Due to the non-random delivery of event-contingent EMAs, we only use event outcome (smoking) data from these EMAs and do not use information about the longitudinal process (emotions). Data collected from three individuals are illustrated in Figure 1.

An important characteristic of these data is the way that smoking information is collected. Each EMA—whether random or event-contingent—prompts the individual to respond to a series of questions that capture their cigarette use in the period since the previous EMA. We use the data to partition the days into non-overlapping time intervals. In each interval, we either observe the count outcome (taking a 0 or positive value) or we do not observe the count outcome and mark it as unobserved. The unobserved windows include sleep times when the individual is assumed not "at-risk." The partitions during which counts are observed are called "event intervals." A general illustration of this data structure is provided in the Supporting Information (see Web Figure 1).

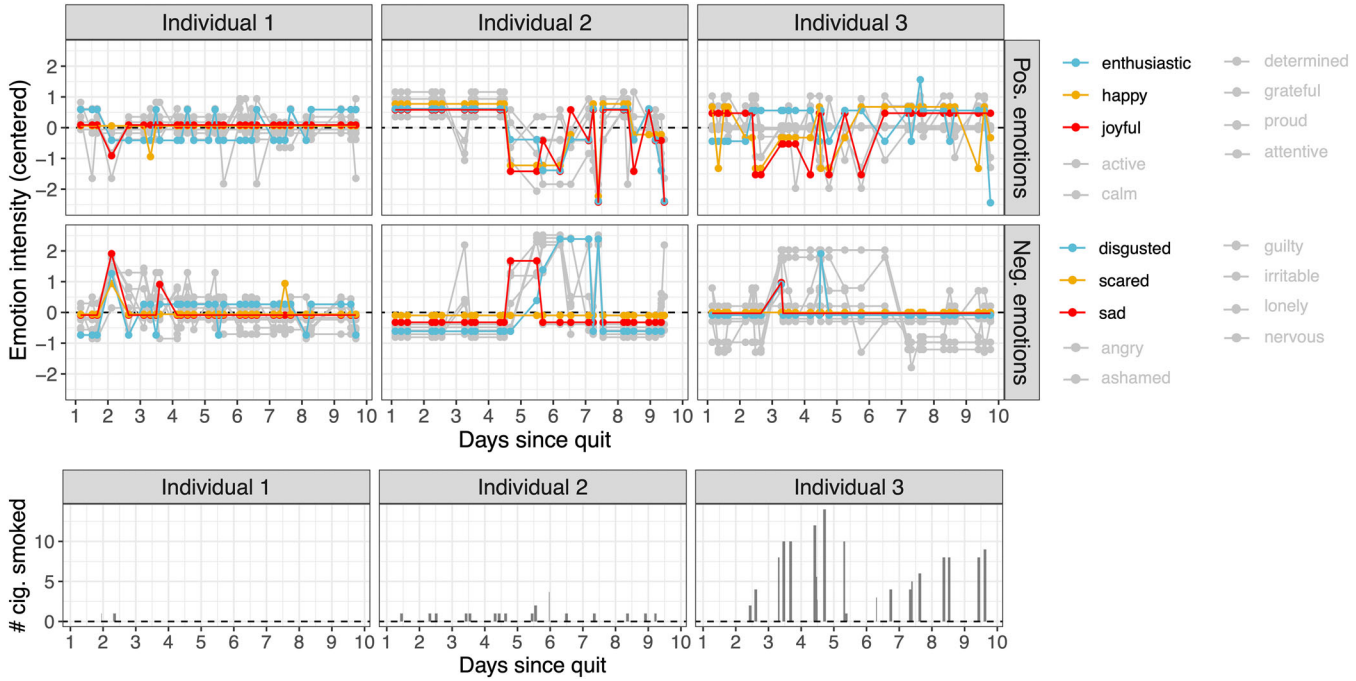


FIGURE 1 Plot of self-reported emotions from the random ecological momentary assessments (EMAs) and self-reported cigarette use from all EMAs for three individuals in the smoking cessation study. A subset of the 18 positive and negative emotions are highlighted for illustrative purposes only. In the plot summarizing reported cigarette use, the width of each bar corresponds to the interval over which an individual reported smoking at least one cigarette and the height of the bar corresponds to the total number of cigarettes reported smoked. Note that any smoking reported at a rate greater than 8 cigarettes per hour has been truncated at 8 cigarettes per hour. This figure is similar to Figure 1 in Abbott et al[28] because both this method and the method in Abbott et al are motivated by the same dataset; however, the goals of these two papers are distinct.

3 | METHODS

We propose a cumulative risk model for the summarized version of the event outcome (eg, cigarette counts), which we model as a function of the partially latent predictors (eg, psychological state). Although we could treat each individual EMA item (eg, emotion) as a measurement of its own latent factor, we instead take a dimension-reduction approach to modeling these observed longitudinal outcomes (eg, we assume that the emotions are measurements of psychological states capturing positive and negative affect). Advantages of this dimension-reduction approach are twofold: (i) it improves interpretability of the cumulative risk model while minimizing information loss due to the high correlation between many observed longitudinal measurements and (ii) it increases computational efficiency by using a low-dimensional latent process to summarize the many measured longitudinal outcomes, rather than directly modeling them via a higher-dimensional process.

Overall, our proposed approach consists of two parts: (i) a longitudinal submodel that summarizes the observed longitudinal outcome (eg, emotions) as lower dimensional dynamic latent factors (eg, psychological states) and (ii) a cumulative risk submodel that links time-varying latent factors with risk of recurrent events (eg, cigarette use).

In the description of these submodels, we use the following notation: Suppose that we have N independent individuals, indexed by $i = 1, \dots, N$. For each individual, a set of K longitudinal outcomes, $X_i(t_{ij}) = \left(X_{i1}(t_{ij}), \dots, X_{ijK}(t_{ij})\right)^\top$, is collected at each measurement occasion t_{ij}, where $j = 1, \dots, n_i$. This stacked vector of longitudinal outcomes is denoted as $X_i = \left(X_i^\top(t_{i1}), \dots, X_i^\top(t_{ij}), \dots, X_i^\top(t_{in_i})\right)^\top$. At each measurement occasion, individuals also report event intensity data, denoted by $Y_i(t_{ij'})$ for individual i at time j'. Note that we do not require that measurement occasions are evenly spaced or that the set measurement occasions for the longitudinal outcomes X and event intensity outcomes Y be the same within the same individual. Our notation currently assumes that no longitudinal outcomes are missing within a given measurement occasion, but our model could easily be adapted to allow for such a scenario.

3.1 | Longitudinal submodel

To summarize the observed emotions as a smaller number of latent psychological states, we take the approach described in Abbott et al.[28] We first use a dynamic factor model to summarize the K observed EMA items (emotions) as a low-dimensional multivariate latent process (representing psychological states), given by

$$X_i(t_{ij}) = \Lambda \eta_i(t_{ij}) + u_i + \epsilon_i(t_{ij})$$

where Λ is the $K \times p$ time-invariant loadings matrix; $\eta_i(t_{ij})$ is the value of the p-dimensional latent process at time t_{ij}, where $p \ll K$; $u_i \sim N(0, \Sigma_u)$ is a random intercept with diagonal covariance matrix Σ_u; and $\epsilon_i(t_{ij}) \sim N(0, \Sigma_\epsilon)$ accounts for error in the measurement of the longitudinal outcome, with covariance matrix Σ_ϵ also assumed to be diagonal. We make the simplifying assumption that each observed longitudinal outcome loads onto only a single latent factor; this means that each row of Λ only contains one non-zero element (and we furthermore assume that the locations of the non-zero elements are known). We make this assumption to clarify the interpretation of the model, but this assumption could be relaxed. Domain knowledge (eg, behavioral science theories) can be incorporated into the structure of this submodel by using it to specify the location of the non-zero elements within the loadings matrix. We include the outcome-specific random intercepts to account for differences in underlying levels of reported responses across individuals. By using a diagonal Σ_u, we assume that these levels are independent across outcomes and instead we capture the correlation between the observed responses using a multivariate stochastic process for the latent factors.

The temporal evolution of the p-dimensional latent process is then modeled with a continuous-time multivariate Ornstein-Uhlenbeck (OU) stochastic process, which can be thought of as a continuous-time analog of a multivariate autoregressive process. Use of this process allows us to capture correlation within and between multiple latent factors over time. The OU process implies that, for individual i, the marginal distribution of the latent factors across all times $t_{i1}, \ldots, t_{in_i}$ is

$$\eta_i \sim N(0, \Psi_i)$$

where $\eta_i = \left(\eta_i^\top(t_{i1}), \ldots, \eta_i^\top(t_{in_i})\right)^\top$ and Ψ_i is a $(pn_i) \times (pn_i)$ covariance matrix parameterized by two $p \times p$ matrices of parameters, θ and σ. The exact form of the covariance matrix is given in the Supporting Information (Section B). To make this dynamic factor model identifiable, we model the OU process on the correlation scale; for additional details, see Abbott et al.[28]

If, for example, $p = 2$, then the OU process parameters are

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{bmatrix} \text{ and } \sigma = \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix}$$

We assume that the off-diagonal elements of σ are 0, which effectively forces all correlation between the latent factors to be captured via θ; diagonal elements of σ are assumed to be positive. The diagonal elements of θ must also be positive but the off-diagonals can take positive or negative values; θ is also constrained to have eigenvalues with positive real parts to ensure that the latent process is mean-reverting.[29] Lastly, we assume that this process has a marginal mean of 0 and that it is stationary.

Together, these two submodels imply that the marginal distribution of the observed longitudinal outcome X_i for individual i is

$$X_i \sim N\left(0, \Sigma_i^*\right) \text{ where } \Sigma_i^* = \left(I_{n_i} \otimes \Lambda\right)\Psi_i\left(I_{n_i} \otimes \Lambda\right)^\top + \left(J_{n_i} \otimes \Sigma_u\right) + \left(I_{n_i} \otimes \Sigma_\epsilon\right)$$

where Ψ_i is the marginal correlation matrix of the OU process parameterized by θ and σ, I_{n_i} is an identity matrix of dimension n_i, J_{n_i} is a $n_i \times n_i$ square matrix of ones, and $\otimes$ denotes the Kronecker product.

This approach to jointly modeling the observed longitudinal outcomes as dynamic latent factors and modeling the evolution of the latent factors as a multivariate OU stochastic process has the advantage of being both a flexible model and returning unbiased estimates of the parameters that characterize the latent process.

3.2 | Cumulative risk submodel

The coarseness of the adverse health event measurements makes it challenging to estimate the association between latent longitudinal factors and the instantaneous risk of an event. Instead of modeling instantaneous risk, we propose a cumulative risk model for the association between the average value of the latent factors and total number of events across a small window of time, or event interval. Specifically, we propose using a Poisson regression model to connect the latent factors with risk of events.

Let $Y_i(t_{ij})$ denote the cumulative number of events recorded between times $t_{i,j-1}$ and t_{ij}. To model the intensity of events over the interval from $t_{i,j-1}$ to t_{ij} as a function of the latent process, we use

$$Y_i(t_{ij}) \sim \text{Poi}\big((t_{ij} - t_{i,j-1})\lambda_i(t_{ij})\big) \text{ where } \log\big(\lambda_i(t_{ij})\big) = \beta_0 + \beta_\eta^\top \left[\frac{1}{(t_{ij} - t_{i,j-1})}\int_{t_{i,j-1}}^{t_{ij}} \eta_i(s)ds\right]$$

Following the bivariate latent process given as an example in Section 3.1, $\eta_i(s)$ is a length-2 vector that contains the values of $\eta_{i1}(s)$ and $\eta_{i2}(s)$ for individual i at time s over the interval from $t_{i,j-1}$ to t_{ij}. We later denote the average value of the latent process over this interval as $\overline{\eta}_i(t_{ij})$. This allows us to re-write our model as

$$Y_i(t_{ij}) \sim \text{Poi}\big((t_{ij} - t_{i,j-1})\lambda_i(t_{ij})\big) \text{ where } \log\big(\lambda_i(t_{ij})\big) = \beta_0 + \beta_\eta^\top \overline{\eta}_i(t_{ij})$$

Because we assume a stationary OU process for the latent factors, we can derive the limiting distribution of this integral, which is also Gaussian. Details on this limiting distribution are given in the Supporting Information (Section C). The association between the cumulative event outcome and $\overline{\eta}_i(t_{ij})$ is captured through coefficient vector β_η.

3.3 | Estimation

In our two-stage approach to estimation, we first fit the longitudinal submodel using an iterative block coordinate descent algorithm. This block-wise approach was developed in Abbott et al[28] with the goal of improving the computational efficiency of estimation by leveraging the availability of analytic gradients for a subset of parameters. Using results from the first stage of estimation, we then fit a weighted version of the cumulative risk model via R's svyglm function in the survey package[30] where the weights account for uncertainty in the average values of the latent trajectories. This approach to the second stage of estimation, which is based on a Monte Carlo Expectation Maximization (MCEM) algorithm with importance sampling-based weights, involves iterating between updating the cumulative risk model coefficients and then updating the weights.

We briefly describe the motivation for using these weights: in an ideal setting, we would like to sample values of the latent process from the conditional distribution given the event outcome, $\eta_i \mid Y_i$. However, sampling from this distribution is difficult. We use weights that are proportional to this conditional distribution to adjust for the fact that, in practice, the values of the latent process are sampled from the marginal distribution, rather than the conditional distribution. By applying these estimated weights to up-weight or down-weight certain values of the latent process sampled in stage 1, we avoid the need to re-sample these values in stage 2. Additional discussion of this approximate approach is provided in the Supporting Information (Section D).

The two stages that make up the full estimation approach are outlined below, along with a definition of the weights.

Stage 1:

1. Let $\Theta_L = (\Lambda, \Sigma_u, \Sigma_\epsilon, \theta, \sigma)$ be a vector containing all parameter matrices for the longitudinal submodel. Then, initialize parameters in the longitudinal submodel: consider both empirical initial values (based on fitting simpler models) and default initial values and select the set of initial values that correspond to the higher log-likelihood for the longitudinal submodel.
2. Estimate the longitudinal submodel via an iterative block coordinate descent algorithm.
3. Predict the factor scores $\widehat{\eta}_i$ using the parameter estimates, denoted $\widehat{\Theta}_L$, from the fitted longitudinal submodel where $\widehat{\eta}_i = E\left(\eta_i | X_i; \widehat{\Theta}_L\right) = \widehat{\Psi}_i\left(I_i \otimes \widehat{\Lambda}^\top\right)\widehat{\Sigma}_i^{*^{-1}} X_i$. These factor scores are observed at longitudinal measurement occasions (eg, times of the random EMAs).

4. Augment observed data by drawing a value for the event interval-specific average of the latent factors given known values of the factors scores, $\widehat{\eta}_i$, at each endpoint of the interval; that is, sample $\overline{\eta}_i(t_{ij}) \mid \widehat{\eta}_i(t_{i,j-1}), \widehat{\eta}_i(t_{ij}) \sim N(\overline{\mu}, \overline{\Sigma})$. For the form of $\overline{\mu}$ and $\overline{\Sigma}$, see the Supporting Information (Section C). We refer to these generated average values that augment the observed data as "synthetic values." For each of the N individuals, we consider $r = 1, \ldots, R$ possible synthetic values per event interval.

At this point, our data consists of two parts: $Y_i(t_{ij})$, the cumulative number of cigarettes smoked in the interval $(t_{i,j-1}, t_{ij}]$; and $\overline{\eta}_i^{(r)}(t_{ij})$, $r = 1, \ldots, R$, the set of R possible average values of the latent process across each interval.

Stage 2:

1. Fit the cumulative risk model.
   a. Initialize coefficients β in the cumulative risk model by fitting a model with uniform weights. That is, fit

$$Y_i^{(r)}(t_{ij}) \sim Poi\Big((t_{ij} - t_{i,j-1})\lambda_i^{(r)}(t_{ij})\Big) \text{ where } \log\Big(\lambda_i^{(r)}(t_{ij})\Big) = \beta_0 + \beta_\eta^\top \overline{\eta}_i^{(r)}(t_{ij})$$

   and $Y_i^{(r)}(t_{ij}) = Y_i(t_{ij}) \forall r = 1, \ldots, R$.

   b. Given estimates $\widehat{\beta} = \left(\widehat{\beta}_0, \widehat{\beta}_\eta^\top\right)$, update the weights, $w_{ij}^{(r)}$. Calculate weights for each set of synthetic latent factor averages (r) where the weight for each event interval j is calculated as

$$\widetilde{w}_{ij}^{(r)} = \Pr\left(Y_i(t_{ij}) = y \,\middle|\, \overline{\eta}_i^{(r)}(t_{ij}), \widehat{\beta}\right) \propto e^{-(t_{ij}-t_{i,j-1})\lambda_i^{(r)}(t_{ij})}(t_{ij} - t_{i,j-1})\left(\lambda_i^{(r)}(t_{ij})\right)^y$$

   Weights are normalized such that $w_{ij}^{(r)} = \widetilde{w}_{ij}^{(r)} / \sum_r \widetilde{w}_{ij}^{(r)}$ and then again such that $\sum_i \sum_j \sum_r w_{ij}^{(r)} = N$.

   c. Update the coefficients in the cumulative risk model by refitting

$$Y_i(t_{ij}) \sim \text{Poi}\Big((t_{ij} - t_{i,j-1})\lambda_i^{(r)}(t_{ij})\Big) \text{where} \log\Big(\lambda_i^{(r)}(t_{ij})\Big) = \beta_0 + \beta_\eta^\top \overline{\eta}_i^{(r)}(t_{ij})$$

   using weights, $w_{ij}^{(r)}$, which upweight trajectories (indexed by r) that are more likely given the observed event counts and the current cumulative risk model parameter estimates $\widehat{\beta}$.

   d. Let k index iterations. Then iterate between steps b. and c. until convergence, where convergence is defined as a small change in the parameter estimates,

$$|\widehat{\beta}^{(k)} - \widehat{\beta}^{(k-1)}| < 1 \times 10^{-6}$$

2. Estimate standard errors (SEs) using the final parameter estimates $\widehat{\beta}$ and weights via the approach described in Section 3.4.

When implementing stage 2 in practice, we can assume a quasi-Poisson distribution to account for potential overdispersion in the data.

3.4 | Inference

Estimation of SEs for the longitudinal submodel parameters in stage 1 is straightforward using a Fisher information-based approach.[28] Our approach to stage 2 estimation, however, presents a challenging setting for inference due to the need to accurately capture additional uncertainty from the synthetic average values of the latent factors and the estimated weights in the cumulative risk model. Here, we focus on inference for stage 2 and present two approaches to SE estimation often

applied in multiple imputation settings that account for uncertainty in parameter estimates in different manners: standard Rubin's rule for combining variance estimates[31] and a bootstrap-based approach for estimating variance in multiple imputation settings with uncongeniality or model misspecification.[32]

When applying Rubin's rule, we generate M imputed datasets, each of which consists of R sets of synthetic average values for the latent factors. We then apply Rubin's rule in a standard manner to combine point estimates and variances across the M imputed datasets. Variance estimates from each imputed dataset are calculated using robust sandwich estimators to account for uncertainty due to weighting of the R synthetic values. Because we require generating multiple average values per event interval to estimate the weights that are used to help alleviate potential bias in our two-stage method, our data do not have the standard structure of multiply imputed datasets. We found that this implementation of Rubin's rule resulted in anticonservative estimates of SEs and so we attempted to improve our estimates of uncertainty through use of the bootstrap-based approach.

Bartlett and Hughes[33] summarize various approaches for combining bootstrapping and multiple imputation that aim to improve coverage of confidence intervals in settings of uncongeniality or model misspecification. Among these approaches is a computationally efficient adaptation of bootstrapping followed by imputation that was originally proposed in von Hippel and Bartlett.[32] In this approach, called the von Hippel approach, data are bootstrapped, multiple imputation is applied to each bootstrapped dataset, and then the variance is estimated as a weighted sum of the mean sum of squares within and between bootstraps. In this adaptation, the authors suggest the multiple for imputation can be low ($M = 2$) for a reasonable number of bootstrapped samples ($B = 200$) and nominal coverage can still be achieved. We apply this method in our setting using the suggested number of replicates for imputation and bootstrapping. As in our implementation of Rubin's rule, we consider a single imputed dataset to consist of R synthetic average values per event interval. To pool point estimates, we simply average across results from all $M \times B$ imputed bootstrapped datasets.

4 | SIMULATION STUDY

We conduct a simulation study to assess the performance of our two-stage method. Motivated by our mHealth data, here we assume $K = 18$ observed longitudinal outcomes (emotions) are measurements of $p = 2$ underlying latent factors. These longitudinal outcomes are observed for $N = 200$ individuals at n_i measurement occasions, where $n_i \sim Uniform(10, 20)$. The gaps between consecutive measurements range from 1 to 2 units of time and the timing of the measurement occasions define the endpoints of the event intervals. To generate *complete data* for each individual, we take the following approach:

1. Set true values of θ^* and σ^*, which characterize the underlying bivariate OU stochastic process, and generate values of the latent factors at each measurement occasion.
    a. We consider two sets of true values for θ^* and σ^*; one set corresponds to an OU process with higher autocorrelation and the other set corresponds to an OU process with lower autocorrelation.
2. Using the measurements of the true latent factors, generate true values for the observed longitudinal outcome, X_i, also at the measurement occasions.
    a. In this simulation study, we consider only a single set of true values for the parameters $\Lambda^*, \Sigma_u^*, \Sigma_\epsilon^*$ in the longitudinal submodel. We assume that each row of Λ^* only contains one non-zero element and that Σ_u^* and Σ_ϵ^* are diagonal.
3. Using true values of θ^* and σ^*, generate a single set of true average values of the latent process ($\overline{\eta}_i^*$), given the values of the factors at the measurement occasions, across each event interval.
4. Then, generate cumulative numbers of events that occurred across each event interval using the average value of each trajectory within the bivariate latent process by drawing from a Poisson distribution with true mean,

$$\lambda_i^*(t_{ij}) = (t_{ij} - t_{i,j-1}) \exp\left\{\beta_0^* + \beta_1^* \overline{\eta}_{i1}^*(t_{ij}) + \beta_2^* \overline{\eta}_{i2}^*(t_{ij})\right\}$$

where $\beta_0^*, \beta_1^*, \beta_2^*$ are chosen such that, given $\overline{\eta}_1^*$ and $\overline{\eta}_2^*$, the average event rate is close to the observed smoking rate in the mHealth data described in Section 2.

From these *complete data*, we then assume the *observed data* consist of: (a) the timing of the measurement occasions that define the event intervals, (b) the longitudinal outcomes measured at those occasions, and (c) the cumulative number of events that occurred within each event interval. For the true values of the parameters used to generate these data, see the

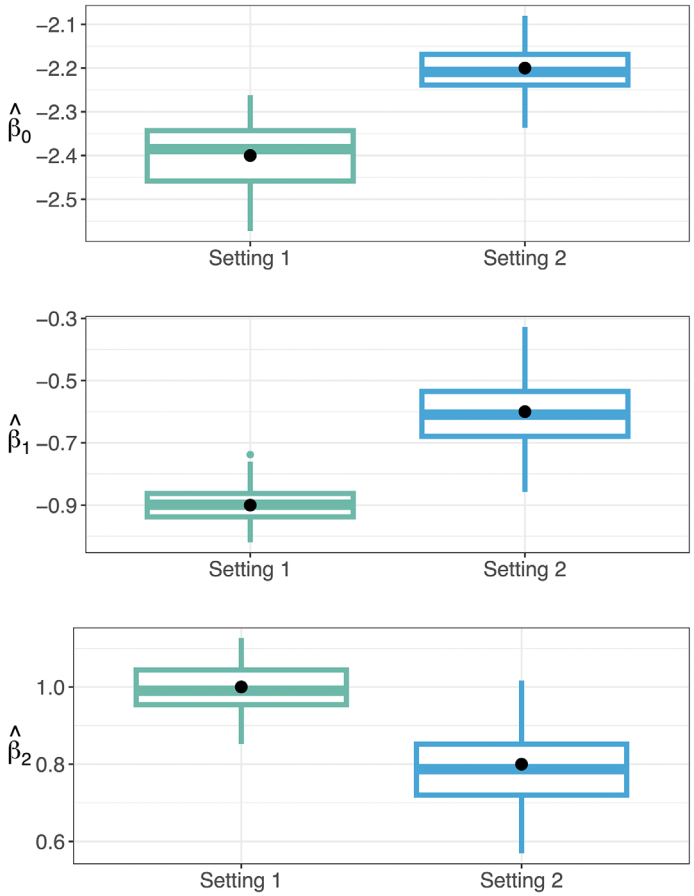


FIGURE 2 Box plots of point estimates of cumulative risk model parameters from simulation study. True values are indicated with black dots. Setting 1 corresponds to data generated from a true stochastic process with higher correlation and setting 2 corresponds to data generated with lower correlation.

Supporting Information (Section E). We then carry out the estimation and inference approaches described in Sections 3.3 and 3.4.

In our simulations, we generate $R = 50$ possible values for the average of the latent factors over each event interval (ie, $\overline{\eta}_i^{(r)}(t_{ij}), r = 1, \ldots, R = 50$). When calculating SEs using the scaled von Hippel approach, we assume $M = 2$ imputed datasets (each of which consist of $R = 50$ synthetic average values per event interval) and $B = 200$ bootstrapped samples, as recommended in Bartlett and Hughes.[33]

We simulate 100 datasets and summarize results across these replicates. We only consider 100 replicates because bootstrapping is computationally expensive. Point estimates from stage 1 are summarized in Web Figure 3. For stage 2, we report point estimates in Figure 2 and summarize coverage rates for confidence intervals calculated using SE estimates from Rubin's rule and the von Hippel approach in Table 1. We find that we consistently recover unbiased estimates of the coefficients in all submodels. For the cumulative risk submodel, coverage rates of the confidence interval show that, in our setting, Rubin's rule returns confidence intervals that are occasionally anticonservative while the von Hippel approach results in confidence intervals with closer-to-nominal coverage. Based on additional empirical results, we find that a large value of R (eg, $R = 50$) helps reduce bias in our point estimates. Additional discussion of the choice of R is provided in the Supporting Information (Section E.2).

5 | APPLICATION TO SMOKING CESSATION DATA

To illustrate this method, we apply it to a subset of data collected from the Break Free II study described earlier (Section 2). Our event outcome for the cumulative risk model is the self-reported total number of cigarettes smoked over repeated

TABLE 1 Coverage rates (%) for 95% confidence intervals for the cumulative risk submodel parameters in the simulation study.

		Setting 1: high correlation		Setting 2: low correlation	
CR	**Coef.**	**RR**	**VH**	**RR**	**VH**
80%	β_0	**69**	**71**	80	82
	β_1	82	85	73	81
	β_2	80	82	**67**	73
95%	β_0	**89**	**89**	94	94
	β_1	96	96	91	94
	β_2	95	99	**85**	92

Note: Coverage rates (CRs) are averaged across **100 datasets**. CRs that fall outside of the range of values expected based on a 95% binomial proportion confidence interval are in bold font. Given that we have calculated the CR for 12 coefficients in each setting (using 80% and 95% confidence intervals), we would expect 0-1 of the CRs to fall outside of the expected range of values from the 95% binomial proportion confidence interval.
Abbreviations: RR, Rubin's rule; VH, von Hippel.

event intervals, where these intervals of time are defined by individuals' responses to the smoking-related EMA questions. Due to uncertainty surrounding the exact time at which each individual attempts to quit smoking, we restrict the cumulative risk model to use only data that were collected 24 hours after the designated quit date. This approach is intended to minimize the accidental inclusion of cigarette use that preceded quit. Furthermore, if a participant failed to respond to any EMA for a long period of time (ie, more than 24 hours), we designated them only "at-risk" within the 24 hours prior to the most recent EMA. As a result, we assumed that data were missing-at-random when fitting the cumulative risk model. We also truncated self-reported cigarette use at a maximum rate of 8 cigarettes per hour; this value of 8 was selected based on domain knowledge of how many cigarettes a heavy smoker could reasonably smoke in an hour.

The observed outcomes used for the longitudinal submodel included the intensity of 18 emotions (listed in Figure 1). Although each random EMA assessed the intensity of a total of 23 emotions, our illustrative analysis included only 18 emotions due to computational constraints. These 18 emotions were selected by fitting the longitudinal submodel to the 6 emotions with the highest loadings based on a cross-sectional factor model and then gradually adding emotions from the remaining 17 until computational cost became prohibitive. We assume that the 18 emotions are observed measurements of two latent factors that represent the psychological constructs of positive affect and negative affect.[34] We also assume that structural zeros within the loadings matrix of the longitudinal submodel are known, meaning that positive emotions are measurements of only positive affect and negative emotions are measurements of only negative affect. Thus, correlation between positive and negative affect is captured entirely through the multivariate latent process representing these unobservable affective states. The cumulative risk model is,

$$Y_i(t_{ij}) \sim Poi\big((t_{ij} - t_{i,j-1})\lambda_i(t_{ij})\big)$$

$$\log\big(\lambda_i(t_{ij})\big) = \beta_0 + \beta_1 PA_i(t_{ij}) + \beta_2 NA_i(t_{ij}) + \beta_3 q_{ij} + \beta_4 PA_i(t_{ij}) \times q_{ij} + \beta_5 NA_i(t_{ij}) \times q_{ij}$$

where PA indicates average positive affect across event interval t_{ij}, NA represents average negative affect, time q_{ij} corresponds to the midpoint time of event interval t_{ij} and is in units of weeks since quit, and the outcome $Y_i(t_{ij})$ has units of cigarettes smoked per hour. As a result of the identifiability constraint on the longitudinal submodel, a one-unit change in PA or NA corresponds to one standard deviation.

We previously applied the stage 1 longitudinal submodel to all available post-quit data in Abbott et al[28] and use these existing results here. Because we previously considered all post-quit emotion data when fitting the longitudinal submodel, our analyses included $N = 218$ individuals. However, since in the current illustrative analysis we have restricted the data to 24 hours post-quit, our sample size reduces to the subset of $N = 214$ individuals because responses for four individuals were only available in the first 24 hours post-quit. As these individuals contributed such limited information, we do not believe that the impact of including/excluding their emotion data when estimating the population-level parameters of the longitudinal submodel would substantially impact the results. Point estimates and confidence intervals for the longitudinal submodel are given in the Supporting Information (Web Figure 6). Thus, we use results from previous application of the longitudinal submodel as our stage 1 results here.

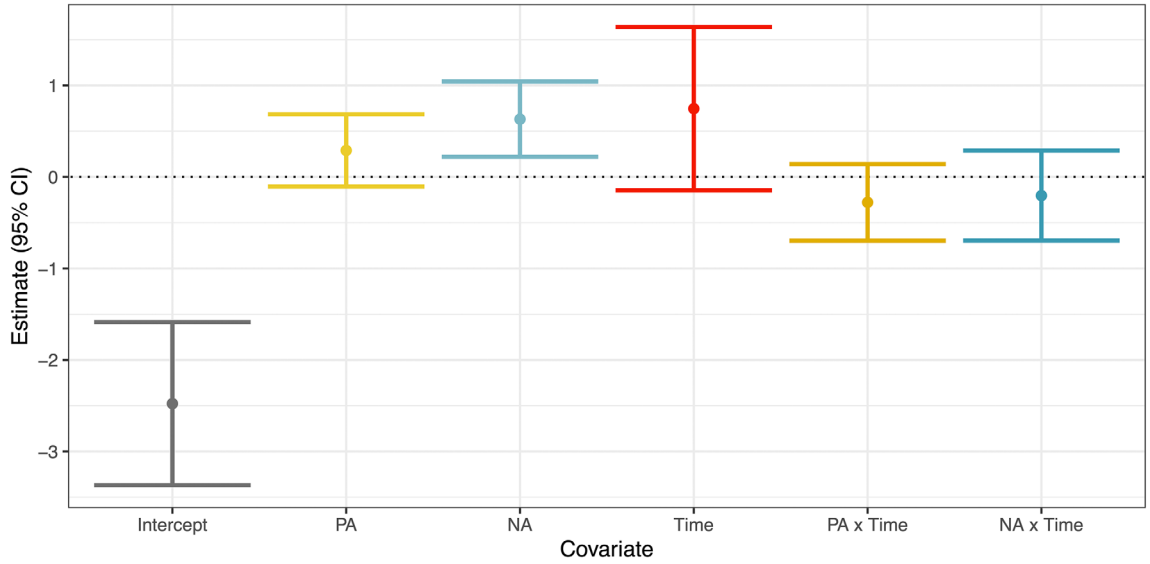


FIGURE 3 Point estimates and 95% confidence intervals (CIs) for cumulative risk model parameters applied to mHealth smoking cessation study data. Error bars correspond to 95% confidence intervals calculated from von Hippel standard error estimates using $R = 50, M = 2$, and $B = 200$. Time is in units of weeks since quit. Abbreviations: NA, negative affect; PA, positive affect.

In our estimation algorithm, we generate $R = 50$ synthetic average values of the latent factors for each event interval. For inference, we assume $M = 2$ and $B = 200$ as in the simulation study.

In Figure 3, we report point estimates and confidence intervals (on the log scale) from the fitted cumulative risk model. These confidence intervals are based on SEs from the von Hippel approach. The point estimates indicate that above average levels of negative affect (where a value of 0 corresponds to the average) are associated with increased smoking but the association between negative affect and smoking decays over time. The positive association between higher negative affect and increased smoking could be explained by smoking being used as a tool for coping with high stress or negative feelings. But given the observational nature of these data, and since the relative ordering of affective state and cigarettes smoked is unknown, we can only draw conclusions about associations and not causal effects. As such, other explanations may be plausible as this application does not distinguish between affect as a driver of cigarette use vs changes in affect resulting from cigarette use.

Figure 4 illustrates trends in the association between smoking and positive and negative affect over time. This figure allows us to evaluate changes over time in expected smoking rates across hypothetical individuals with (a) average positive affect and average negative affect (PA = 0, NA = 0), (b) high positive affect and low negative affect (PA = 1, NA = −1), or (c) low positive and high negative affect (PA = −1, NA = 1). Based on these results, the difference in expected smoking rates by affective state increases over time; however, all 95% confidence intervals overlap. Note that confidence intervals presented in this figure are point-wise confidence intervals calculated using SEs from the von Hippel approach.

6 | DISCUSSION

In this paper, we proposed a two-stage approach for analyzing longitudinal EMA data to determine how longitudinal outcomes (in our case, emotions) are associated with increased risk of repeated events (here, intensity of cigarette use). We present a two-stage approach to fitting a longitudinal submodel—a dynamic factor model—that summarizes the time-varying dynamics of multiple measured longitudinal outcomes as a lower dimensional stochastic process and a cumulative risk model—a Poisson regression model—that captures the association between latent factors represented by the stochastic process and a count outcome. We use weights to address the partially unobservable nature of both the outcome and predictors and to reduce potential bias in our two-stage approach to estimation.

Readers may be concerned that we take a two-stage approach to fitting our model, which is commonly known to result in biased estimates of coefficients. We argue, however, that the focus of this paper is on the development of a

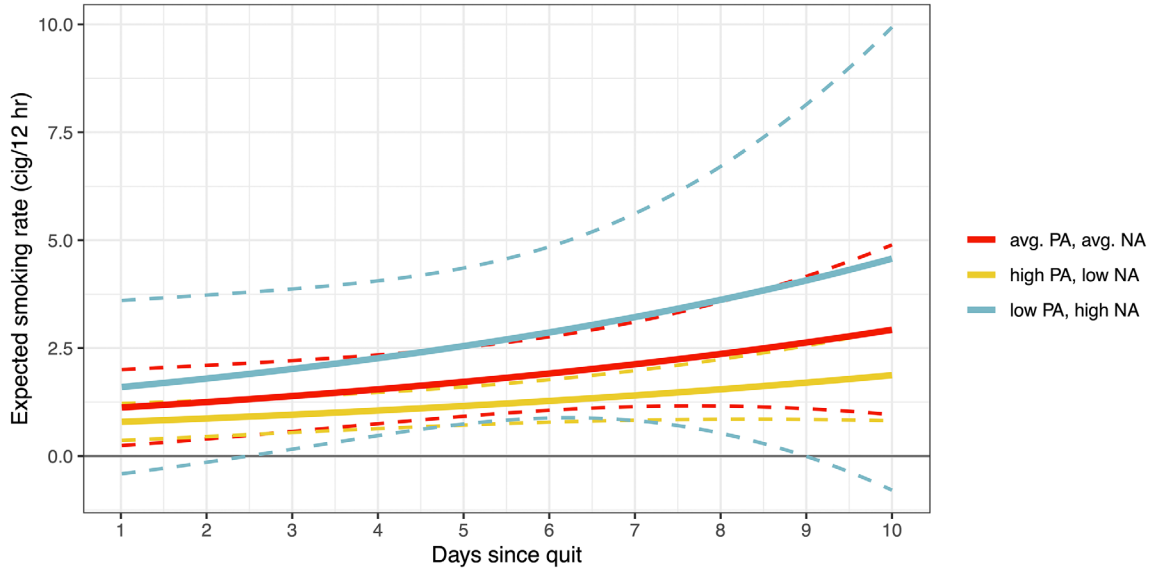


FIGURE 4 Expected smoking rate (in units of cigarettes per 12 hour interval) using estimated coefficients and assuming constant values of either (a) average positive affect (PA = 0) and average negative affect (NA = 0), (b) above-average positive affect (PA = 1) and below-average negative affect (NA = −1), or (c) below-average positive affect (PA = −1) and above-average negative affect (NA = 1). This figure illustrates the difference in expected smoking rates at a fixed time point under different hypothetical levels of positive and negative affect. Dotted lines indicate 95% point-wise confidence intervals estimated using von Hippel SEs.

practical method and this two-stage approach allows us to leverage existing methods and software. We include weights in our cumulative risk submodel, which seem to help alleviate bias. Readers might also express concern about our use of MCEM, which relies on approximations rather than direct maximization of the likelihood from our model. Because we assume that our count outcome follows a Poisson distribution, the complexity of our likelihood increases. We take advantage of the approximations with MCEM to greatly simplify the computations involved in the second stage of estimation. Our empirical investigations indicate that these approximations are close enough to yield unbiased estimates in our setting; however, future work could include the development of a computationally efficient approach for directly maximizing the likelihood in the two-stage approach or jointly estimating the longitudinal and cumulative risk submodels.

One might argue that if computational cost is a concern that inhibits use of joint models in practice, then why not fit a three-part model consisting of a cross-sectional factor model, a stochastic process fit to factor scores, and then the risk model? Through simulation, we found that treating the predicted factor scores as observed values of the latent process and then trying to estimate the OU process using those predicted factor scores resulted in poor estimates of the stochastic process parameters and thus biased coefficient estimates in our risk model. Stage 1 estimation does contribute a large portion of our method's computation time (see Web Figure 5 for computation by stage). However, the method that we use for the first stage of estimation was shown in Abbott et al[28] to require substantially less computation time than an alternative approach to fitting a similar model proposed in Tran et al.[29] Although both the first stage of estimation and the bootstrapping within the second stage of estimation requires a significant amount of time, stage 2 only takes approximately 6 minutes to return point estimates in the simulation study. Our ability to use off-the-shelf software in stage 2 means that the computational cost of this stage is low enough to easily allow for estimation of bootstrap-based SEs. That is, while the bootstrap-based method for SE estimation is computationally demanding, accurate estimation of SEs is only possible because each subroutine (ie, fitting each weighted Poisson model) is fast enough and easy to implement. Thus, exploration of various combinations, interactions, and parameterizations of the latent states in the cumulative risk model is feasible from a practical point of view.

Through simulation, we found that taking a standard likelihood-based approach to inference of the cumulative risk submodel parameters via Fisher information resulted in underestimates of SEs and thus confidence intervals with low coverage. To improve our ability to fully capture the uncertainty of these point estimates, we then tried adjusting our SE estimates via an application of Rubin's rule and the von Hippel approach. Our estimation approach does not fit into standard scenarios addressed by the Rubin's rule and so this approach resulted occasional under-coverage; we found

that the bootstrap-based von Hippel approach was slightly better at accurately capturing the uncertainty in our point estimates. Through simulation studies, we found that coverage of confidence intervals is not sensitive to the choice of R. Development of an alternative method for SE estimation could be a useful direction for future work, as it would avoid the added computation time of bootstrapping.

In our formulation of the longitudinal dynamic factor model, we assume that the locations of structural zeros within the loadings matrix are known. Behavioral science theories regarding how emotions relate to each other can be used to justify the locations of these zeros in our setting. While our work could be extended to allow for estimation of an entirely non-null loadings matrix, this extension would present additional challenges related to model identifiability and interpretation; however, it would likely be useful in other settings in which scientific evidence for informing the structure of the loadings matrix is limited. In addition to specifying the locations of the structural zeros, we also require the number of latent factors (p) to be pre-specified. In our motivating application, use of two latent factors is a natural choice. However, if selecting the optimal number of latent factors for summarizing the observed longitudinal data is of interest, Akaike information criterion (AIC) and Bayesian information criterion (BIC) could be used; we suggest using BIC as it tends to error on the side of selecting a slightly simpler model. For more discussion on using AIC and BIC to compare longitudinal submodels with different numbers of latent factors, please see Abbott et al.[28] A further extension could integrate a data-driven approach for selecting the number of latent factors into the estimation algorithm.

Importantly, our proposed method does not allow us to make any inference about the order of the values of latent affect states and occurrence of adverse health events; the method only provides information about associations. This limitation results, in part, from the modeling decision that we made to account for the partially unobservable nature of both the events of interest and longitudinal predictors recorded in the data. The observational nature of the data limits the ability to make causal conclusions about how affect impacts smoking, regardless of how the models are fitted (ie, jointly or separately). Nonetheless, this method enables investigators to model multiple dynamic states of risk and their connection with adverse health events in a way that can advance behavioral theory and generate hypotheses that can be tested in future research. For example, the results can guide the development of just-in-time interventions[35,36] that explicitly target the newly identified states of risk, and inform the design of micro-randomized trials that investigate the utility of delivering (vs. not delivering) these interventions.

ACKNOWLEDGEMENTS

Funding for this manuscript was provided by the National Center on Minority Health and Health Disparities (NIMHD) under award number R01MD010362; by the National Institute on Drug Abuse (NIDA) under award numbers R01DA039901, P50DA054039, F31DA057048; and by the National Cancer Institute (NCI) under award numbers U01CA229437 and P30CA042014. Additional support was provided by the Huntsman Cancer Foundation. The NIH had no role in the study design, collection, analysis or interpretation of the data, writing the manuscript, or the decision to submit the paper for publication. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Huntsman Cancer Foundation.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data from the smoking cessation study are not available due to privacy concerns. Simulated example data and code are available at: https://github.com/madelineabbott/OUF_CR_twostage.

ORCID

Madeline R. Abbott https://orcid.org/0000-0002-5344-3732
Walter H. Dempsey https://orcid.org/0000-0002-7852-2269

REFERENCES

1. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. *Stat Sin*. 2004;14:809-834.
2. Rustand D, van Niekerk J, Teixeira Krainski E, Rue H, Proust-Lima C. Fast and flexible inference for joint models of multivariate longitudinal and survival data using integrated nested Laplace approximations. *Biostatistics*. 2024;25(2):429-448.

3. Rathbun SL, Song X, Neustifter B, Shiffman S. Survival analysis with time-varying covariates measured at random times by design. *J R Stat Soc Ser C Appl Stat*. 2013;62(3):419-434.
4. Rizopoulos D. Estimation of joint models. In: Chow S-C, ed. *Joint Models for Longitudinal and Time-to-Event Data, with Applications in R*. New York: CRC Press; 2012:62-74.
5. Mauff K, Steyerberg E, Kardys I, Boersma E, Rizopoulos D. Joint models with multiple longitudinal outcomes and a time-to-event outcome: a corrected two-stage approach. *Stat Comput*. 2020;30(4):999-1014.
6. Stennett A, Krebs NM, Liao J, Richie JP Jr, Muscat JE. Ecological momentary assessment of smoking behaviors in native and converted intermittent smokers. *Am J Addict*. 2018;27(2):131-138.
7. Gunter RG, Szeto EH, Suh S, Kim Y, Jeong SH, Waters AJ. Associations between affect, craving, and smoking in Korean smokers: an ecological momentary assessment study. *Addict Behav Rep*. 2020;12:100301.
8. Berg KC, Crosby RD, Cao L, et al. Facets of negative affect prior to and following binge-only, purge-only, and binge/purge events in women with bulimia nervosa. *J Abnorm Psychol*. 2013;122(1):111-118.
9. Proust-Lima C, Amieva H, Jacqmin-Gadda H. Analysis of multivariate mixed longitudinal data: a flexible latent process approach. *Br J Math Stat Psychol*. 2013;66(3):470-487.
10. Muthén BO. Beyond SEM: general latent variable modeling. *Behaviormetrika*. 2002;29:81-117.
11. Roy J, Lin X. Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics*. 2000;56(4):1047-1054.
12. Dunson DB. Dynamic latent trait models for multidimensional longitudinal data. *JASA*. 2003;98(463):555-563.
13. Proust C, Jacqmin-Gadda H, Taylor JMG, Ganiayre J, Commenges D. A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. *Biometrics*. 2006;62(4):1014-1024.
14. Wang X, Berger JO, Burdick DS. Bayesian analysis of dynamic item response models in educational testing. *Ann Appl Stat*. 2013;7:126-153.
15. Liu S, Ou L, Ferrer E. Dynamic mixture modeling with *dynr*. *Multivariate Behav Res*. 2021;56(6):941-955.
16. Taddé BO, Jacqmin-Gadda H, Dartigues JF, Commenges D, Proust-Lima C. Dynamic modeling of multivariate dimensions and their temporal relationships using latent processes: application to Alzheimer's disease. *Biometrics*. 2020;76(3):886-899.
17. Proust-Lima C, Dartigues JF, Jacqmin-Gadda H. Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: a latent process and latent class approach. *Stat Med*. 2016;35(3):382-398.
18. Muthén LK, Muthén B. Categorical latent variable modeling using Mplus: longitudinal data. Mplus Short Courses Topic 6. 2009 Accessed June 29, 2023. https://www.statmodel.com/download/Topic%206.pdf
19. McCurdy S, Molinaro A, Pachter L. Factor analysis for survival time prediction with informative censoring and diverse covariates. *Stat Med*. 2019;38(20):3719-3732.
20. Liu M, Sun J, Herazo-Maya JD, Kaminski N, Zhao H. Joint models for time-to-event data and longitudinal biomarkers of high dimension. *Stat Biosci*. 2019;11(3):614-629.
21. Larsen K. The Cox proportional hazards model with a continuous latent variable measured by multiple binary indicators. *Biometrics*. 2005;61(4):1049-1055. doi:10.1111/j.1541-0420.2005.00374
22. Muthén LK, Muthén BO. Example 6.22: continuous-time survival analysis using a parametric proportional hazards model with a factor influencing survival. *Muthén & Muthén. Mplus User's Guide*. 8th ed. Los Angeles, CA; 1998-2017 Accessed June 29, 2023. https://www.statmodel.com/HTML_UG/chapter6V8.htm
23. Asparouhov T, Muthén B. Continuous-Time Survival Analysis in Mplus. 2018 Accessed June 29, 2023. https://www.statmodel.com/download/Survival.pdf
24. Saleheen N, Ali AA, Hossain SM, et al. puffMarker: a multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. *Proc ACM Int Conf Ubiquitous Comput*. 2015:999–1010.
25. Hovsepian K, Al'Absi M, Ertin E, Kamarck T, Nakajima M, Kumar S. cStress: towards a gold standard for continuous stress assessment in the mobile environment. *Proc ACM Int Conf Ubiquitous Comput*. 2015;493-504.
26. Fiore MC, Jaén CR, Baker TB, et al. *Treating tobacco use and dependence: 2008 update*. Clinical Practice Guideline. Rockville, MD: US Department of Health and Human Services. Public Health Service; 2008.
27. Potter LN, Yap J, Dempsey W, Wetter DW, Nahum-Shani I. Integrating intensive longitudinal data (ILD) to inform the development of dynamic theories of behavior change and intervention design: a case study of scientific and practical considerations. *Prev Sci*. 2023;24:1-13.
28. Abbott MR, Dempsey WH, Nahum-Shani I, Lam CY, Wetter DW, Taylor JMG. A continuous-time dynamic factor model for intensive longitudinal data arising from mHealth studies. *arXiv*. 2023. doi:10.48550/arXiv.2307.15681
29. Tran TD, Lesaffre E, Verbeke G, Duyck J. Latent Ornstein-Uhlenbeck models for Bayesian analysis of multivariate longitudinal categorical responses. *Biometrics*. 2021;77(2):689-701.
30. Lumley T. survey: analysis of complex survey samples. R Package Version 4.0. 2020.
31. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley; 1987.
32. von Hippel PT, Bartlett JW. Maximum likelihood multiple imputation: faster imputations and consistent standard errors without posterior draws. *Statist Sci*. 2021;36(3):400-420.
33. Bartlett JW, Hughes RA. Bootstrap inference for multiple imputation under uncongeniality and misspecification. *Stat Methods Med Res*. 2020;29(12):3533-3546.
34. Remington NA, Fabrigar LR, Visser PS. Reexamining the circumplex model of affect. *J Pers Soc Psychol*. 2000;79(2):286-300. doi:10.1037/0022-3514.79.2.286
35. Nahum-Shani I, Hekler EB, Spruijt-Metz D. Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework. *Health Psychol*. 2015;34(S):1209-1219.

36. Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Ann Behav Med*. 2018;52(6):446-462.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Abbott MR, Nahum-Shani I, Lam CY, Potter LN, Wetter DW, Dempsey WH. A latent variable approach to jointly modeling longitudinal and cumulative event data using a weighted two-stage method. *Statistics in Medicine*. 2024;43(21):4163-4177. doi: 10.1002/sim.10171