Research Article

Statistics in Medicine

Received 15 May 2014, Accepted 27 October 2014 Published online 14 November 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6368

Two-stage model for time-varying effects of discrete longitudinal covariates with applications in analysis of daily process data

Hanyu Yang,[a] James A. Cranford,[b] Runze Li[c] and Anne Buu[d*†]

This study proposes a generalized time-varying effect model that can be used to characterize a discrete longitudinal covariate process and its time-varying effect on a later outcome that may be discrete. The proposed method can be applied to examine two important research questions for daily process data: measurement reactivity and predictive validity. We demonstrate these applications using health risk behavior data collected from alcoholic couples through an interactive voice response system. The statistical analysis results show that the effect of measurement reactivity may only be evident in the first week of interactive voice response assessment. Moreover, the level of urge to drink before measurement reactivity takes effect may be more predictive of a later depression outcome. Our simulation study shows that the performance of the proposed method improves with larger sample sizes, more time points, and smaller proportions of zeros in the binary longitudinal covariate.

Keywords: functional data; alcohol use; measurement error; mixed model; daily process data

1. Introduction

Daily patterns of health risk behaviors such as substance use can be used to assess the risk of developing health problems [1] and examine the dynamics of intervention effects over time [2]. Because of the high cost and heavy participant burden associated with prospective daily data collection, these types of data have been collected primarily using retrospective methods. Thanks to the advancement of new telecommunication technology such as the interactive voice response (IVR), such cost and burden have been reduced. IVR allows humans to interact with a computer via a telephone keypad or by speech recognition. IVR systems can instruct and respond to users with prerecorded audio and also record their responses into databases. They have been adopted as powerful research tools in recent years.

Prospective daily data collection using the IVR has the advantages of cutting costs of staff time as well as minimizing recall bias and tendency to underreport socially undesirable behaviors [3]. Yet, it unavoidably involves self-monitoring of the target behavior, which is an active component of some cognitive-behavioral interventions for substance use disorders [4]. The potential *measurement reactivity* (defined as reducing the target behavior due to self-awareness) is undesirable for those studies that aim to investigate the association between the target behavior and its precursor or consequence. On the other hand, for those applications aiming to facilitate behavior changes, such an effect can be used to boost or extend intervention effects [5]. Thus, verifying measurement reactivity is an important research question, especially given that existing empirical investigations are few and have produced mixed results [6, 7].

[a]*Department of Statistics, Pennsylvania State University, University Park, PA 16802, U.S.A.*
[b]*Department of Psychiatry, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[c]*Department of Statistics and The Methodology Center, Pennsylvania State University, University Park, PA 16802, U.S.A.*
[d]*Department of Epidemiology and Biostatistics, Indiana University, Bloomington, IN 47405, U.S.A.*
**Correspondence to: Anne Buu, Department of Epidemiology and Biostatistics, Indiana University, 1025 East 7th St., Bloomington, IN 47405, U.S.A.*
[†]*E-mail: yabuu@indiana.edu*

In this study, we provide a statistical model that can be used to characterize the trajectory of behavior changes during the IVR assessment.

Another important measurement issue associated with the IVR assessment is its *predictive validity* referring to the utility of the pattern of changes in these repeated measures for predicting a short-term or long-term health outcome. For example, an alcohol study may examine whether the pattern of daily alcohol consumption over a period of time is predictive of alcohol-related problems or symptoms at a later time point. In order to evaluate the effect of longitudinal patterns of health risk behaviors on a health outcome, we have to overcome some methodological challenges. First, the covariate is measured at many time points but the outcome is collected at one future time point so standard longitudinal methods that were designed for longitudinal outcomes are not applicable in this setting. Second, the effect of longitudinal patterns of health risk behaviors on a short-term or long-term health outcome may be a complex function of time. For example, those periods with frequent occurrence of binge drinking (defined as consuming more than five standard alcohol drinks in one episode) tend to have higher negative effects on alcohol-related problems. Third, health-risk behaviors are most of the time self-reported and thus are subject to measurement errors due to recall bias or embarrassment [8]. Fourth, the between-subject and within-subject variability tends to be large in this kind of data, especially for studies on high risk youth who have not yet developed regular patterns of health-risk behaviors [9]. In this study, we provide a statistical model that can address all of these methodological challenges.

Zhang *et al*. [10] developed a two-stage functional mixed model that was motivated by the need to investigate the effect of the follicle stimulating hormone time profile during a menstrual cycle on total hip bone mineral density measured at a single time point. The model consists of two stages: the first stage estimates the periodic subject-specific follicle stimulating hormone profiles using a nonparametric measurement error model; the second stage plugs the estimated subject-specific profiles into a functional linear model and estimates the functional covariate effect. This model was designed for a longitudinal covariate process with *continuous* values and a *continuous* scalar outcome. Yet, in many practical settings, daily process data are discrete. Particularly in substance abuse research, measures of alcohol use often yield count variables (e.g., 'How many drinks containing alcohol did you have yesterday?'); measures of drug use often yield binary variables (e.g., 'Did you use prescription stimulants like Adderall or Ritalin to get high yesterday?'). Thus, in order to make the model more applicable to our research questions, measurement reactivity and predictive validity, we extend it to a more general setting in which both the longitudinal covariate process and outcome could be either *discrete* or *continuous*.

This paper is organized as follows. In Section 2, we present a generalized time-varying effect model that can be used to characterize a discrete longitudinal covariate process and its time-varying effect on a later outcome. In Section 3, the 14-day daily process data of an alcohol study are presented as a motivating example. Section 4 delineates the design and results of the simulation study. We present discussion and concluding remarks in Section 5.

2. The statistical model

Suppose that from the i-th subject ($i = 1, \ldots, n$), we collect an observed longitudinal covariate process $\mathbf{W}_i = (W_{i1}, \ldots, W_{in_i})^T$, a vector of time-invariant covariates $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^T$, and an outcome Y_i. In the context of our research questions, $\mathbf{W}_i$ is daily IVR data of self-reported health-risk behaviors, which tend to contain missing values and are subject to measurement errors. Thus, we assume that there exists a smooth latent individual trajectory $x_i(\cdot)$ in $[T_1, T_2]$, and the observed $\mathbf{W}_i$ is related to this latent $x_i(\cdot)$ through the following nonparametric measurement error model:

$$W_{ij} \sim f_W(\cdot; \mu_{ij}), \quad h(\mu_{ij}) = x_i(t_{ij}), \tag{1}$$

where $f_W(\cdot)$ is a distribution in the exponential family, and $h(\cdot)$ is a known link function. This model would allow us to characterize the individual trajectory of the health-risk behaviors collected through the IVR and examine the hypothesis of *measurement reactivity*.

In the context of our research question, *predictive validity*, we are interested in studying the effect of longitudinal patterns of health-risk behaviors on a later health outcome. Assume that the outcome Y_i is related to the individual trajectory $x_i(\cdot)$ and the time-invariant covariate $\mathbf{Z}_i$ through a partial functional

generalized linear model [11]

$$Y_i \sim f_Y(\cdot;\eta_i), \quad g(\eta_i) = \mathbf{Z}_i^T\boldsymbol{\delta} + \int_{T_1}^{T_2} x_i(t)\gamma(t)dt, \tag{2}$$

where $f_Y(\cdot)$ is a distribution belonging to an exponential family, $g(\cdot)$ is a known link function, $\boldsymbol{\delta}$ is a vector of regression coefficients of $\mathbf{Z}_i$, and $\gamma(\cdot)$ is a smooth function for the time-varying effect of longitudinal patterns of health-risk behaviors. In particular, if Y_i's are Gaussian, it becomes a partial functional linear model [12] of

$$Y_i = \mathbf{Z}_i^T\boldsymbol{\delta} + \int_{T_1}^{T_2} x_i(t)\gamma(t)dt + \epsilon_i.$$

The calibration regression method is adopted with the estimation procedure involving two stages: Stage-I estimates $x_i(t)$ in Model (1) based on the observed data W_{ij}; Stage-II estimates $\boldsymbol{\delta}$ and $\gamma(t)$ in Model (2), given the estimate $\hat{x}_i(t)$ from Stage-I and the observed data $\mathbf{Z}_i$ and Y_i. Here, we only provide a general description of the estimation procedure. Interested readers are referred to the Appendix for the technical details. In Stage-I, the subject-specific latent profile $x_i(\cdot)$ is decomposed into a population profile and a random individual deviation: $x_i(\cdot) = x_0(\cdot) + d_i(\cdot)$. When the natural cubic spline (NCS) technique is applied, Model (1) becomes a generalized additive mixed model [13]. In Stage-II, the predictors $\hat{x}_i(\cdot)$'s from Stage-I are then plugged into Model (2). With the NCS technique, the calibration model yields a semiparametric model, and thus similar procedures can be employed for the estimation.

3. The motivating example

3.1. The study on alcoholic couples

To demonstrate applications of the proposed model, we conducted statistical analysis on the real data of a study on the feasibility of using IVR technology to collect daily diary data from alcoholic couples for 14 consecutive days [14]. Fifty-four alcoholic married couples (either spouse met DSM-IV diagnosis [15] of past year alcohol use disorder) were recruited from the University of Michigan Addiction Treatment Services (37%) and the local community. At baseline, couples completed questionnaires about their past month moods, marital interactions, and drinking behaviors and received an extensive IVR training session. Participants were instructed to call a toll-free telephone number separately during a designated time window (5–9 PM) when they had 15 min of privacy to report their daily moods, marital interactions, and alcohol involvement. Responses were automatically entered into an online database. In order to increase compliance, the participants were informed that they would receive an automated reminder call if they had not called the IVR system by 8 PM; the amount of the incentive they received would depend on their level of compliance. Participants completed a total of 1418 out of a possible 1512 daily reports, for an overall compliance rate of 94%. About half of the sample completed all 14 daily IVR reports.

In this study, we applied the proposed method to characterize the overall change in self-reported urge to drink (a binary covariate) during the 14 days of IVR assessment. Urge to drink refers to a broad range of thoughts, physical sensations, or emotions that tempt someone to drink, even though he/she has at least some desire not to. A decreasing trajectory of urge to drink would be an evidence to support the theory of measurement reactivity. We also modeled the time-varying effect of urge to drink on a continuous scale of depression (the Beck Depression Inventory [16]) that was measured in 6 months after the IVR assessment. This set of analysis allows us to investigate the predictive validity of the daily patterns of urge to drink.

3.2. Statistical analysis on real data

Because our real data were collected from married couples, we modified the notations in Section 2 slightly by denoting $i = 1, \ldots, m$ as the family, and k as the family member nested in i.

We modeled the binary longitudinal covariate W_{ikj}, urge to drink, in a certain day (1 for 'yes' and 0 for 'no'), as

$$W_{ikj} \sim \text{Bernoulli}(\mu_{ikj}), \quad \text{logit}(\mu_{ikj}) = x_{ik}(t_{ikj}).$$

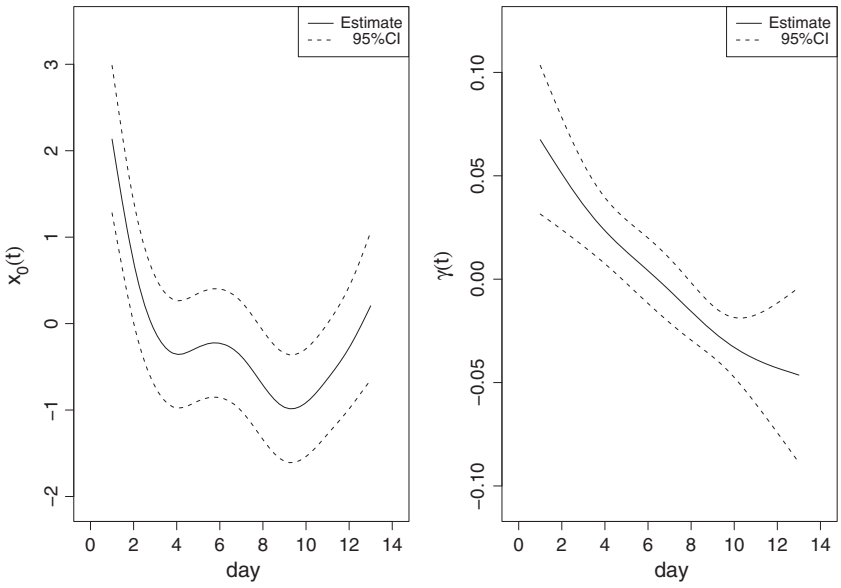

Figure 1. Real data analysis results: trajectories of the longitudinal covariate $x_0(t)$ and its time-varying effect $\gamma(t)$.

The depression score Y_{ik} is related to the recruitment setting Z_{ik} (1 for the treatment sample from the University of Michigan Addiction Treatment Services; and 0 for the community sample recruited from the local community) and the latent individual trajectory $x_{ik}(\cdot)$ through

$$Y_{ik} = Z_{ik}\delta + \int_{T_1}^{T_2} x_{ik}(t)\gamma(t)dt + a_i + \epsilon_{ik},$$

where a_i is an additional random effect with $N(0, \sigma_a^2)$ that characterizes the family effect.

We obtain $\hat{x}_{ik}(\cdot)$ following the procedures described in Appendix A.2, when $W_{ikj} \sim \text{Bernoulli}(\mu_{ikj})$ and $h(\cdot) = \text{logit}(\cdot)$. Because a random effect term is introduced, δ and $\gamma(\cdot)$ are then estimated by fitting a calibration model

$$\mathbf{Y} = \mathbf{Z}\delta + \hat{\mathbf{X}}_C\boldsymbol{\gamma} + \mathbf{N}_a\mathbf{a} + \boldsymbol{\epsilon}^*,$$

where $\mathbf{N}_a$ is a $n \times m$ matrix with 1 for the (ik, i)-th element and 0 elsewhere, and $\mathbf{a}$ is a m-dimensional random vector with $N(\mathbf{0}, \sigma_a^2\mathbf{I})$. Following [17], we have $\hat{\delta}$ and $\hat{\boldsymbol{\gamma}}$ in the forms of Equation (A.3) in Appendix A.3, with $\mathbf{W}_z = \mathbf{W} - \mathbf{W}\hat{\mathbf{X}}_C\left(\hat{\mathbf{X}}_C^T\mathbf{W}\hat{\mathbf{X}}_C + \lambda_\gamma\mathbf{K}\right)^{-1}\hat{\mathbf{X}}_C^T\mathbf{W}$ and $\mathbf{W}_x = \mathbf{W} - \mathbf{W}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{W}$, where $\mathbf{W} = \left(\sigma_{\epsilon^*}^2\mathbf{I} + \sigma_a^2\mathbf{N}_a\mathbf{N}_a^T\right)^{-1}$. Moreover, both σ_a^2 and $\sigma_{\epsilon^*}^2$ are estimated through restricted maximum likelihood; the smoothing parameter λ_γ is selected by generalized cross validation (GCV), which is defined as Equation (A.1) in the Appendix.

Figure 1 characterizes the overall covariate trajectory $x_0(\cdot)$ and the time-varying effect $\gamma(\cdot)$. The left panel indicates that although participants' urge to drink showed a systematic decrease in the first week of IVR assessment, it rebounded during the second week. This implies that the effect of measurement reactivity was only short-term. The right panel indicates that the initial level of urge to drink (i.e. before measurement reactivity took effect) was more predictive of the depression level 6 months later, because the time-varying effect of urge to drink on depression was only significantly positive in the first few days of the IVR assessment.

Furthermore, our analysis indicates that the treatment sample had a higher level of depression than the community sample ($\hat{\delta} = 0.4105$ with the 95% confidence interval of $[0.2657, 0.5552]$). The variance components in the model were estimated as $\hat{\sigma}_1^2 = 2.60^2$, $\hat{\sigma}_2^2 = 1.02^2$, $\hat{\tau}_d = 3.64^2$, $\hat{\sigma}_a^2 = 0.05^2$, and $\hat{\sigma}_\epsilon^2 = 0.48^2$. Moreover, the smoothing parameters selected by GCV were $\lambda_x = 0.4$ and $\lambda_\gamma = 20$.

4. Simulation study

Our simulation experiment was designed to evaluate the performance of the proposed model under different situations. Particularly, we manipulated three factors: (i) the sample sizes: small $n = 100$, medium $n = 200$ and large $n = 400$; (ii) the number of time points: two weeks $r = 14$, three weeks $r = 21$ and four weeks $r = 28$; and (iii) the proportion of zeros in the longitudinal covariate: 50%, 70%, and 90%, which can be achieved by adjusting the population profile function $x_0(\cdot)$. While we manipulated three factors, the rest of the design of our simulation was based on the data features of the motivating example in Section 3.

We set $\mathbf{t}^0$ to be r equally spaced time points in $[-1.5, 1.5]$, where $r = 14$, 21 or 28. It is worth to note that, observation time points t_{ikj}'s were not necessary to be balanced among all the subjects. Particularly, for each subject ik, the longitudinal covariate process $W_{ik}(t)$ was observed at t_{ikj} for $j = 1, \ldots, n_{ik}$, where n_{ik} was an integer randomly chosen from $\{\lceil 0.8r \rceil, \ldots, r-1, r\}$, and t_{ikj}'s were n_{ik} distinct points among $\mathbf{t}^0$.

The longitudinal covariate data W_{ikj}'s were generated from

$$W_{ikj} \sim \text{Bernoulli}(\mu_{ikj}), \quad \text{logit}(\mu_{ikj}) = x_{ik}(t_{ikj}) = x_0(t_{ikj}) + d_{ik}(t_{ikj}).$$

Here, we consider three choices of $x_0(\cdot)$: $x_0^{(1)}(t) = 0.8t^4 - t^2 - 0.5t - 0.1$, $x_0^{(2)}(t) = 0.8t^4 - t^2 - 0.5t - 1.3$, and $x_0^{(3)}(t) = 0.8t^4 - t^2 - 0.5t - 3$, which correspond to the proportions of zeros in longitudinal covariates at 50%, 70%, and 90%, respectively. In this way, we are able to examine our hypothesis that the proposed method may perform better in the setting with a ratio of 0s to 1s being 50 : 50 than in the other settings with the corresponding ratios of 70 : 30 or 90 : 10, because the former tends to have greater Fisher information. Moreover, the random process $d_{ik}(\cdot)$ was determined by $\mathbf{d}_{ik} = \mathbf{B}_* \mathbf{b}_{ik}$ with $\mathbf{b}_{ik}$'s being a random sample from $\text{N}(\mathbf{0}, \text{diag}\{1^2, 0.6^2, 1.5^2 \mathbf{I}_{(r-2)\times(r-2)}\})$. The response data Y_{ik}'s were then generated from

$$Y_{ik} = Z_{ik}\delta + \int_{-1.5}^{1.5} x_{ik}(t)\gamma(t)dt + a_i + e_{ik},$$

where $\delta = 0.4$; $a_i \sim \text{N}(0, 0.2^2)$; $e_{ik} \sim \text{N}(0, 0.6^2)$; and $\gamma(t) = -0.6\arctan(0.8t)$ that simulates the corresponding function estimated from the real data in the motivating example. In addition, to generate the time-invariant covariate, Z_{ik}'s were randomly drawn from $\{0, 1\}$ with $\text{P}(Z_{ik} = 1) = 0.5$ for each subject ik.

In summary, we manipulated three factors: the sample size (n), the number of time points (r), and the population profile function ($x_0(\cdot)$) and considered $3 \times 3 \times 3 = 27$ situations in total. Under each situation, we generated $N = 1000$ data sets and applied the proposed two-stage method to estimate the parameters. For each parameter, the mean squared error (MSE) and its empirical standard error (se) were calculated from N replications. In terms of the nonparametric functions $x_0(\cdot)$ and $\gamma(\cdot)$, the mean integrated squared error (MISE) and its empirical standard error were used to summarize the results.

The results of the simulation study are summarized in Tables 1–3. Table 1 shows the MSEs/MISEs and the associated standard errors for the three different sample sizes, holding the other two factors constant. When the sample size becomes larger, both the MSE of the coefficient parameter δ and the MISEs of the population profile function $x_0(t)$ and the time-varying effect $\gamma(t)$ become smaller. These results indicate that the performance of the proposed method improves as the sample size increases. The estimation for variance components (i.e., σ_1^2, σ_2^2, τ_d in the longitudinal covariate model, and σ_a^2 and σ_ϵ^2 in the response model) also benefits from a larger sample size. Such benefit, in turn, enhances the accuracy of estimation for δ, $x_0(t)$, and $\gamma(t)$, which are of primary interest. Table 2 demonstrates that the performance of the proposed method is better when we collect data from more time points. However, such improvement is not as salient as the effect of a larger sample size. Table 3 supports our hypothesis that the proposed method performs worse when the binary longitudinal covariates contain a higher proportion of 0's, because the amount of information contained in the data is reduced. Such an effect is particularly evident as the proportion of zeros increases from 50% to 90%.

In addition to the effects of single factors demonstrated in Tables 1–3, we have conducted simulations to investigate potential interactions between each pair of factors. When the sample size and the number of time points are both varied (holding the proportion of zeros constant at 50%), the performance of the proposed model improves with the number of time points when the sample size is relatively small ($n = 100$), but such an effect is not apparent when the sample size is large ($n = 400$). A similar pattern is found

Table 1. Simulation results with varied sample sizes n ($r = 21$, $x_0^{(1)}(t)$).

	$n = 100$		$n = 200$		$n = 400$	
	MSE/MISE	SE	MSE/MISE	SE	MSE/MISE	SE
δ	0.0181	0.0252	0.0101	0.0137	0.0049	0.0066
σ_1^2	0.0498	0.0576	0.0405	0.0384	0.0348	0.0272
σ_2^2	0.0091	0.0114	0.0050	0.0059	0.0031	0.0037
τ_d	0.7122	1.1583	0.4181	0.6587	0.2376	0.3273
σ_a^2	0.0041	0.0075	0.0022	0.0036	0.0013	0.0017
σ_ϵ^2	0.0194	0.0219	0.0195	0.0160	0.0210	0.0129
$x_0(\cdot)$	0.2313	0.1385	0.1448	0.0752	0.0983	0.0458
$\gamma(\cdot)$	0.0259	0.0340	0.0145	0.0142	0.0104	0.0082

MSE, mean squared error; MISE, mean integrated squared error; SE, standard error.

Table 2. Simulation results with varied numbers of time points r ($n = 200$, $x_0^{(1)}(t)$).

	$r = 14$		$r = 21$		$r = 28$	
	MSE/MISE	SE	MSE/MISE	SE	MSE/MISE	SE
δ	0.0096	0.0130	0.0101	0.0137	0.0090	0.0125
σ_1^2	0.0498	0.0481	0.0405	0.0384	0.0340	0.0343
σ_2^2	0.0064	0.0085	0.0050	0.0059	0.0042	0.0054
τ_d	0.6676	1.0266	0.4181	0.6587	0.2915	0.4464
σ_a^2	0.0021	0.0034	0.0022	0.0036	0.0020	0.0033
σ_ϵ^2	0.0256	0.0215	0.0195	0.0160	0.0156	0.0143
$x_0(\cdot)$	0.1565	0.0855	0.1448	0.0752	0.1420	0.0777
$\gamma(\cdot)$	0.0165	0.0187	0.0145	0.0142	0.0134	0.0145

MSE, mean squared error; MISE, mean integrated squared error; SE, standard error.

Table 3. Simulation results with varied population profile $x_0(t)$ ($n = 200$, $r = 21$).

	$x_0^{(1)}(t)$: 50% zeros		$x_0^{(2)}(t)$: 70% zeros		$x_0^{(3)}(t)$: 90% zeros	
	MSE/MISE	SE	MSE/MISE	SE	MSE/MISE	SE
δ	0.0101	0.0137	0.0105	0.0149	0.0121	0.0163
σ_1^2	0.0405	0.0384	0.0496	0.0453	0.0735	0.0714
σ_2^2	0.0050	0.0059	0.0054	0.0069	0.0092	0.0120
τ_d	0.4181	0.6587	0.4918	0.7761	0.8911	1.5323
σ_a^2	0.0022	0.0036	0.0023	0.0041	0.0026	0.0046
σ_ϵ^2	0.0195	0.0160	0.0210	0.0170	0.0310	0.0224
$x_0(\cdot)$	0.1448	0.0752	0.2098	0.1331	0.3462	0.2186
$\gamma(\cdot)$	0.0145	0.0142	0.0168	0.0217	0.0424	0.0803

MSE, mean squared error; MISE, mean integrated squared error; SE, standard error.

as we examine the effects of the sample size and the proportion of zero simultaneously. Specifically, the negative effect of a larger proportion of zeros diminishes as the sample size increases. By the same token, the positive effect of a larger sample size is more noticeable when the proportion of zeros becomes larger. However, when the number of time points and the proportion of zeros are manipulated simultaneously, we do not observe an interaction effect like those in the other two pairs. The two-way tables of this set of simulations are available upon request.

The overall covariate trajectory across all subjects $x_0(\cdot)$ is crucial in our research setting, because it characterizes the trajectory of behavior changes during the IVR assessment and, therefore, can be used to examine measurement reactivity. The time-varying effect $\gamma(\cdot)$ is also very important because it can be used to identify the period of time during the IVR assessment that is more predictive of future outcomes. Thus, we evaluate the performance of the proposed method by comparing the estimated curves of $x_0(\cdot)$ and $\gamma(\cdot)$ with the true curves in each setting. Figure 2 shows the estimated curves derived from evaluating

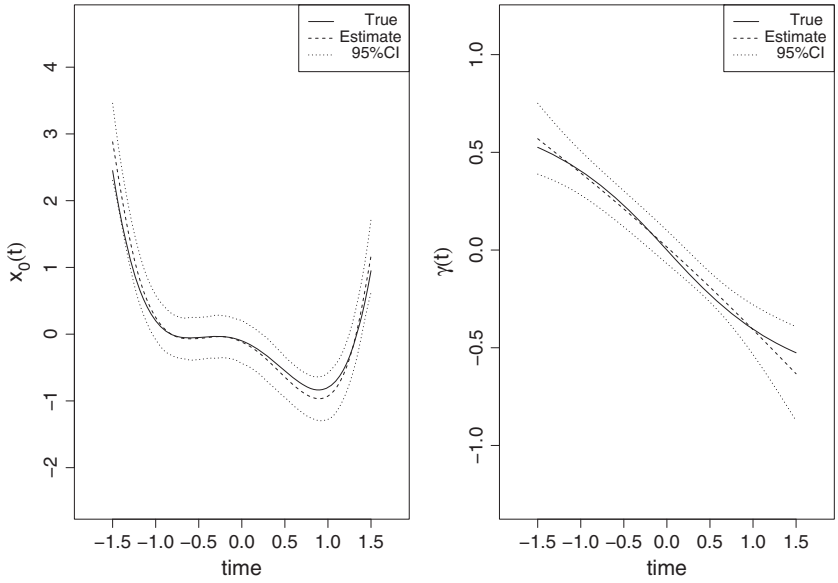

Figure 2. Simulation results: estimates of $x_0(t)$ and $\gamma(t)$ under the setting of $n = 200$, $r = 21$, and $x_0^{(1)}(t)$.

the fitted functions of the N replications at a set of grid points and connecting means at all grid points, for the setting of $n = 200$, $r = 21$, and $x_0^{(1)}(t)$. This figure indicates that our proposed method estimates $x_0(\cdot)$ and $\gamma(\cdot)$ well with relatively small bias, and the true functions are covered by 95% pointwise empirical confidence intervals. The figures for other settings look similar and are available upon request.

5. Discussion

This study proposes a generalized time-varying effect model that can be used to characterize a continuous or discrete longitudinal covariate process and its time-varying effect on a later outcome that could be continuous or discrete. The proposed method can be applied to examine two important research questions for daily process data: measurement reactivity and predictive validity. We demonstrate these applications using health-risk behavior data collected from alcoholic couples through the IVR system. The proposed model can also be used to analyze daily process data that are collected using other modern technology including the mobile phone text messaging, handheld computers, or Web-based data collection. The statistical analysis results show that the effect of measurement reactivity may only be evident in the first week of IVR assessment and fade away afterwards. Moreover, the level of urge to drink before measurement reactivity takes effect may be more predictive of the level of depression 6 months later.

We conduct a simulation study based on the features of real data to evaluate the performance of the proposed method under different situations. The results show that the performance improves with larger sample sizes, more time points, and smaller proportions of zeros in the binary longitudinal covariate. Future research may consider multiple correlated longitudinal covariates because health-risk behaviors such as substance use, violence, and sexual risk behavior tend to co-occur. Furthermore, in some applications, researchers may collect multiple correlated outcomes or longitudinal outcomes at later time. More methodological work is thus needed to extend the proposed model to handle such complex data.

Appendix A: details of the two-stage estimation

A.1. Natural cubic spline

In the estimation procedure, we propose to approximate nonparametric functions $x_i(\cdot)$ and $\gamma(\cdot)$ by employing the NCS [18], with knots $\mathbf{t}^0 = (t_1^0, \ldots, t_r^0)^T$ being an r-dimensional vector of ordered distinct values of all time points $\{t_{ij}\}$, $(i = 1, \ldots, n, j = 1, \ldots, 1, \ldots, n_i)$. According to Green and Silverman [18], there exists a set of r piecewise cubic polynomial basis functions $\mathbf{c}(\cdot) = (c_1(\cdot), \ldots, c_r(\cdot))^T$ such that $x_i(\cdot) = \mathbf{x}_i^T\mathbf{c}(\cdot)$ and $\gamma(\cdot) = \boldsymbol{\gamma}^T\mathbf{c}(\cdot)$, where $\mathbf{x}_i = x_i(\mathbf{t}^0)$ and $\boldsymbol{\gamma} = \gamma(\mathbf{t}^0)$ are evaluated at $\mathbf{t}^0$. Particularly, each $c_l(\cdot)$ itself is an NCS function satisfying $c_l(\mathbf{t}^0) = \mathbf{e}_l$, where $\mathbf{e}_l$ is an r-dimensional vector with 1 in the l-th element

and 0 elsewhere. Hence, by defining $\mathbf{C} = \int_{T_1}^{T_2} \mathbf{c}(t)\mathbf{c}^T(t)dt$, we can write the integral term in Model (2) as $\int_{T_1}^{T_2} x_i(t)\gamma(t)dt = \mathbf{x}_i^T\mathbf{C}\boldsymbol{\gamma}$, for computational purposes.

In order to smooth a function $s(\cdot)$, the smoothing spline technique adds the following quadratic roughness penalty term to the quasi-likelihood of $s(\cdot)$:

$$J_s = \frac{1}{2}\lambda\int_{T_1}^{T_2}(s''(t))^2dt = \frac{1}{2}\lambda\mathbf{s}^T\mathbf{K}\mathbf{s},$$

where $\mathbf{s} = s(\mathbf{t}^0)$, $\lambda \geq 0$ is a smoothing parameter, and $\mathbf{K}$ is an $r \times r$ smoothing matrix specified in Equation (2.3) of [18]. In addition, following [19] and [13], we denote $\mathbf{T} = (\mathbf{1}_r, \mathbf{t}^0)$ as the non-trivial null space of $\mathbf{K}$ and decompose $\mathbf{K}$ as $\mathbf{K} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is an $r \times (r-2)$ full-rank matrix satisfying $\mathbf{L}^T\mathbf{T} = \mathbf{0}$. We also define $\mathbf{B} = \mathbf{L}(\mathbf{L}^T\mathbf{L})^{-1}$ and use it in the Stage-I of estimation.

A.2. Stage-I

The latent individual trajectory $x_i(\cdot)$ in Model (1) can be further decomposed as

$$x_i(\cdot) = x_0(\cdot) + d_i(\cdot),$$

where $x_0(\cdot)$ is the population trajectory, and $d_i(\cdot)$ is the deviation of an individual trajectory from the population trajectory. We assume that $x_0(\cdot)$ and $d_i(\cdot)$ are both NCS functions [18] and $d_i(\cdot)$'s are independent mean-zero Gaussian processes. Model (1) can thus be expressed as

$$\mathbf{W}_i \sim f_W(\cdot;\boldsymbol{\mu}_i), \quad h(\boldsymbol{\mu}_i) = \mathbf{N}_i\mathbf{x}_0 + \mathbf{N}_i\mathbf{d}_i,$$

where $h(\boldsymbol{\mu}_i) = (h(\mu_{i1}), \ldots, h(\mu_{in_i}))^T$, $\mathbf{x}_0 = x_0(\mathbf{t}^0)$, $\mathbf{d}_i = d_i(\mathbf{t}^0)$, and $\mathbf{N}_i$ is an $n_i \times r$ incidence matrix mapping $(t_{i1}, \ldots, t_{in_i})^T$ to $\mathbf{t}^0$ such that the (j,l)-th element is 1 if $t_{ij} = t_l^0$ and 0 otherwise.

In order to account for the smoothness of the random Gaussian processes $d_i(\cdot)$'s, we take a transformation of $\mathbf{d}_i = \mathbf{B}_*\mathbf{b}_i$. Here, $\mathbf{B}_* = (\mathbf{T}, \mathbf{B})$ is an $r \times r$ full-rank matrix, $\mathbf{b}_i$'s are r-dimensional random vectors independently distributed as $\mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{D}(\boldsymbol{\theta}))$, where $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \tau_d)^T$ is a vector of variance components, and $\mathbf{D}(\boldsymbol{\theta}) = \mathrm{diag}\{\sigma_1^2, \sigma_2^2, \tau_d\mathbf{I}_{(r-2)\times(r-2)}\}$ is the covariance matrix. By denoting $\mathbf{Q}_i = \mathbf{N}_i\mathbf{B}_*$, $h(\boldsymbol{\mu}) = (h(\boldsymbol{\mu}_1)^T, \ldots, h(\boldsymbol{\mu}_n)^T)^T$, $\mathbf{N} = (\mathbf{N}_1^T, \ldots, \mathbf{N}_n^T)^T$, $\mathbf{Q} = \mathrm{diag}\{\mathbf{Q}_1, \ldots, \mathbf{Q}_n\}$, and $\mathbf{b} = (\mathbf{b}_1^T, \ldots, \mathbf{b}_n^T)^T \sim \mathrm{N}(\mathbf{0}, \mathcal{D})$ with $\mathcal{D} = \mathrm{diag}\{\mathbf{D}, \ldots, \mathbf{D}\}$, we obtain a special form of the generalized additive mixed model ([13])

$$h(\boldsymbol{\mu}) = \mathbf{N}\mathbf{x}_0 + \mathbf{Q}\mathbf{b}.$$

Following the method in Section 3 of [13], both $\mathbf{x}_0$ and $\mathbf{b}$ can be derived by maximizing the double penalized quasi-likelihood

$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i} d_{ij} - \frac{1}{2}\mathbf{b}^T\mathcal{D}^{-1}\mathbf{b} - \frac{1}{2}\lambda_x\mathbf{x}_0^T\mathbf{K}\mathbf{x}_0,$$

with the conditional deviance

$$d_{ij} = -2\int_{W_{ij}}^{\mu_{ij}} \frac{w_{ij}(W_{ij} - s)}{v_W(s)}ds,$$

where $v_W(\cdot)$ is the variance function determined by $f_W(\cdot)$; and w_{ij}'s are prior weights. Explicitly, $\hat{\mathbf{x}}_0$ and $\hat{\mathbf{b}}$ solve the estimating equations of

$$\mathbf{N}^T\mathcal{W}\boldsymbol{\Delta}(\mathbf{W} - \boldsymbol{\mu}) - \lambda_x\mathbf{K}\mathbf{x}_0 = \mathbf{0}, \quad \mathbf{Q}^T\mathcal{W}\boldsymbol{\Delta}(\mathbf{W} - \boldsymbol{\mu}) - \mathcal{D}^{-1}\mathbf{b} = \mathbf{0},$$

where $\mathbf{W} = (\mathbf{W}_1^T, \ldots, \mathbf{W}_n^T)^T$, $\boldsymbol{\Delta} = \mathrm{diag}\{\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_n\}$ with $\boldsymbol{\Delta}_i = \mathrm{diag}\{h'(\mu_{ij})\}$, and $\mathcal{W} = \mathrm{diag}\{\mathcal{W}_1, \ldots, \mathcal{W}_n\}$ with $\mathcal{W}_i = \mathrm{diag}\{\frac{w_{ij}}{\phi_W v_W(\mu_{ij})(h'(\mu_{ij}))^2}\}$.

Moreover, the variance component $\boldsymbol{\theta}$ can be estimated by maximizing the marginal log-quasi-likelihood and then bias-corrected, following similar procedures described in Sections 4 and 5 of [13]. Specifically, if $x_0(\cdot)$ is represented as $\mathbf{x}_0 = \mathbf{T}\boldsymbol{\alpha}_x + \mathbf{B}\mathbf{a}_x$, with $\boldsymbol{\alpha}_x$ having a uniform prior distribution and $\mathbf{a}_x \sim \mathrm{N}(\mathbf{0}, \lambda_x^{-1}\mathbf{I})$, the marginal log-quasi-likelihood is approximated as

$$-\frac{1}{2}\log|\mathcal{V}| - \frac{1}{2}\log|(\mathbf{NT})^T\mathcal{V}^{-1}(\mathbf{NT})| - \frac{1}{2}\left\{\tilde{\mathbf{W}} - (\mathbf{NT})\hat{\boldsymbol{\alpha}}_x\right\}^T \mathcal{V}^{-1}\left\{\tilde{\mathbf{W}} - (\mathbf{NT})\hat{\boldsymbol{\alpha}}_x\right\},$$

where $\tilde{\mathbf{W}} = \mathbf{N}\mathbf{x}_0 + \mathbf{Q}\mathbf{b} + \boldsymbol{\Delta}(\mathbf{W} - \boldsymbol{\mu})$, $\mathbf{R} = \mathbf{Q}\mathcal{D}\mathbf{Q}^T + \mathcal{W}^{-1}$, $\mathcal{V} = \lambda_x^{-1}(\mathbf{NB})(\mathbf{NB})^T + \mathbf{R}$, and $\hat{\boldsymbol{\alpha}}_x$ is derived from $\hat{\mathbf{x}}_0$. That is, for each element $\boldsymbol{\theta}_l$ of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_l$ is the solution to

$$-\frac{1}{2}\mathrm{tr}\left(\mathbf{P}\frac{\partial\mathbf{R}}{\partial\boldsymbol{\theta}_l}\right) + \frac{1}{2}\left(\tilde{\mathbf{W}} - \mathbf{N}\hat{\mathbf{x}}_0\right)^T\mathbf{R}^{-1}\frac{\partial\mathbf{R}}{\partial\boldsymbol{\theta}_l}\mathbf{R}^{-1}\left(\tilde{\mathbf{W}} - \mathbf{N}\hat{\mathbf{x}}_0\right) = 0,$$

where $\mathbf{P} = \mathbf{R}^{-1} - \mathbf{R}^{-1}(\mathbf{NB}_*)\mathbf{H}^{-1}(\mathbf{NB}_*)^T\mathbf{R}^{-1}$. A correction on $\hat{\boldsymbol{\theta}}$ is then taken, following [13] and [20]. The smoothing parameter λ_x is selected via the GCV. In particular, the GCV is defined by assuming independence of the observations, that is,

$$GCV(\lambda_x) = \frac{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i} d_{ij}}{N\left(1 - \frac{edf_x}{N}\right)^2}, \tag{A.1}$$

where $N = \sum_{i=1}^{n} n_i$, and edf_x is the effective degree of freedom in estimating $\mathbf{x}_0$. The parameter λ_x is then chosen by a grid search over $GCV(\lambda_x)$'s.

A.3. Stage-II

We plug in the estimate of $\mathbf{x}_i$'s from Stage-I and fit a calibration model

$$g(\boldsymbol{\eta}) = \mathbf{Z}\boldsymbol{\delta} + \hat{\mathbf{X}}_C\boldsymbol{\gamma},$$

where $\hat{\mathbf{X}}_C = \hat{\mathbf{X}}\mathbf{C}$ with $\hat{\mathbf{X}} = (\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_n)^T$. Hence, $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ are estimated by maximizing the penalized pseudo-quasi-likelihood of

$$-\frac{1}{2}\sum_{i=1}^{n} d_i - \frac{1}{2}\lambda_{\gamma}^T\mathbf{K}\boldsymbol{\gamma},$$

with the deviance

$$d_i = -2\int_{Y_i}^{\eta_i}\frac{w_i(Y_i - s)}{v_Y(s)}ds,$$

where $v_Y(\cdot)$ is the variance function determined by $f_Y(\cdot)$; and w_i's are prior weights. Specifically, estimators $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\gamma}}$ can be iteratively solved from the estimating equations of

$$\mathbf{Z}^T\mathcal{U}^*\boldsymbol{\Xi}(\mathbf{Y} - \boldsymbol{\eta}) = \mathbf{0}, \quad \hat{\mathbf{X}}^T\mathcal{U}^*\boldsymbol{\Xi}(\mathbf{Y} - \boldsymbol{\eta}) - \lambda_{\gamma}\mathbf{K}\boldsymbol{\gamma} = \mathbf{0},$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{\Xi} = \mathrm{diag}\{g'(\eta_i)\}$, and $\mathcal{U}^* = \mathrm{diag}\{\frac{w_i}{\phi_{Y}v_Y(\eta_i)(g'(\eta_i))^2}\}$. The smoothing parameter λ_{γ} is selected through GCV, which is similar to that for λ_x in Stage-I.

Particularly, if Y_i is Gaussian, the calibration model is

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\delta} + \hat{\mathbf{X}}_C\boldsymbol{\gamma} + \boldsymbol{\epsilon}^*, \tag{A.2}$$

where ϵ_i^*'s are treated as independent following $\boldsymbol{\epsilon}^* \sim \mathrm{N}(\mathbf{0}, \sigma_{\epsilon^*}^2\mathbf{I})$. We instead estimate $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ by maximizing the penalized pseudo-log-likelihood of

$$-\frac{1}{2\sigma_{\epsilon^*}^2}\left(\mathbf{Y}-\mathbf{Z}\boldsymbol{\delta}-\hat{\mathbf{X}}_C\boldsymbol{\gamma}\right)^T\left(\mathbf{Y}-\mathbf{Z}\boldsymbol{\delta}-\hat{\mathbf{X}}_C\boldsymbol{\gamma}\right)-\frac{1}{2}\lambda_\gamma\boldsymbol{\gamma}^T\mathbf{K}\boldsymbol{\gamma},$$

which yields maximum penalized likelihood estimators [17]:

$$\hat{\boldsymbol{\delta}}=\left(\mathbf{Z}^T\mathbf{W}_z\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{W}_z\mathbf{Y},\quad \hat{\boldsymbol{\gamma}}=\left(\hat{\mathbf{X}}_C^T\mathbf{W}_x\hat{\mathbf{X}}_C+\lambda_\gamma\mathbf{K}\right)^{-1}\hat{\mathbf{X}}_C^T\mathbf{W}_x\mathbf{Y}, \tag{A.3}$$

where $\mathbf{W}_z=\frac{1}{\sigma_{\epsilon^*}^2}\left\{\mathbf{I}-\hat{\mathbf{X}}_C\left(\hat{\mathbf{X}}_C^T\hat{\mathbf{X}}_C+\sigma_{\epsilon^*}^2\lambda_\gamma\mathbf{K}\right)^{-1}\hat{\mathbf{X}}_C^T\right\}$ and $\mathbf{W}_x=\frac{1}{\sigma_{\epsilon^*}^2}\left\{\mathbf{I}-\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\right\}$.

Moreover, the variance $\sigma_{\epsilon^*}^2$ in Model (A.2) is estimated from the restricted maximum likelihood. Explicitly, $\hat{\sigma}_{\epsilon^*}^2$ solves the estimating equation of

$$-\frac{1}{2}\mathrm{tr}(\mathbf{P})+\frac{1}{2\sigma_{\epsilon^*}^4}(\mathbf{Y}-\mathbf{Z}\hat{\boldsymbol{\delta}}-\hat{\mathbf{X}}\hat{\boldsymbol{\gamma}})^T(\mathbf{Y}-\mathbf{Z}\hat{\boldsymbol{\delta}}-\hat{\mathbf{X}}\hat{\boldsymbol{\gamma}})=0,$$

where

$$\mathbf{P}=\frac{1}{\sigma_{\epsilon^*}^2}\left\{\mathbf{I}-(\mathbf{Z},\hat{\mathbf{X}})\begin{pmatrix}\mathbf{Z}^T\mathbf{Z} & \mathbf{Z}^T\hat{\mathbf{X}}\\ \hat{\mathbf{X}}^T\mathbf{Z} & \hat{\mathbf{X}}^T\hat{\mathbf{X}}+\sigma_{\epsilon^*}^2\lambda_\gamma\mathbf{K}\end{pmatrix}^{-1}(\mathbf{Z},\hat{\mathbf{X}})^T\right\}.$$

Similar to Stage-I, the smoothing parameter λ_γ is selected by GCV.

Acknowledgements

Cranford's research was supported by a National Institutes of Health (NIH) grant R21 AA015105; Li's research was supported by NIH grants P50 DA010075, P50 DA036107, and R01 CA168676; and Buu's research was supported by NIH grants, K01 AA016591 and R01 DA035183. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Mundt JC, Perrine MW, Searles JS, Walter D. An application of interactive voice response (IVR) technology to longitudinal studies of daily behavior. *Behavior Research Methods, Instruments, & Computers* 1995; **27**:351–357.
2. Gwaltney CJ, Magill M, Barnett NP, Apodaca TR, Colby SM, Monti PM. Using daily drinking data to characterize the effects of a brief alcohol intervention in an emergency room. *Addictive Behaviors* 2011; **36**:248–250.
3. Bardone AM, Krahn DD, Goodman BM, Searles JS. Using interactive voice response technology and timeline follow-back methodology in studying binge eating and drinking behavior: different answers to different forms of the same question. *Addictive Behaviors* 2000; **25**:1–11.
4. Simpson TL, Kivlahan DR, Bush KR, McFall ME. Telephone self-monitoring among alcohol use disorder patients in early recovery: a randomized study of feasibility and measurement reactivity. *Drug and Alcohol Dependence* 2005; **79**:241–250.
5. Tucker JA, Blum ER, Xie L, Roth DL, Simpson CA. Interactive voice response self-monitoring to assess risk behaviors in rural substance users living with HIV/AIDS. *AIDS Behav* 2012; **16**:432–440.
6. Stritzke WGK, Dandy J, Durkin K, Houghton S. Use of interactive voice response (IVR) technology in health research with children. *Behavior Research Methods* 2005; **37**:119–126.
7. Barta WD, Tennen H, Litt MD. Measurement reactivity in diary research. In *Handbook of research methods for studying daily life*, Mehl MR, Conner TS (eds). Guilford Press: New York, 2012; 108–123.
8. Tourangeau R, Yan T. Sensitive questions in surveys. *Psychological Bulletin* 2007; **133**:859–883.
9. Collins LR, Kashdan TB, Koutsky JR, Morsheimer ET, Vetter CJ. A self-administered timeline followback to measure variations in underage drinkers' alcohol intake and binge drinking. *Addictive Behaviors* 2008; **33**:196–200.
10. Zhang D, Lin X, Sowers MF. Two-stage functional mixed models for evaluating the effect of longitudinal covariate profiles on a scalar outcome. *Biometrics* 2007; **63**:351–362.
11. Muller HG, Stadtmuller U. Generalized functional linear models. *The Annals of Statistics* 2005; **33**:774–805.
12. Ramsay JO, Silverman BW. *Functional Data Analysis* Second Editon. Springer: New York, 2005.
13. Lin X, Zhang D. Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B* 1999; **61**:381–400.
14. Cranford JA, Tennen H, Zucker RA. Feasibility of using interactive voice response to monitor daily drinking, moods, and relationship processes on a daily basis in alcoholic couples. *Alcoholism: Clinical and Experimental Research* 2010; **34**: 499–508.
15. American Psychiatric Association. *Diagnostic and Statistical Manual* 4th edition American Psychiatric Association: Washington DC, 1994.
16. Beck AT, Steer RA, Brown GK. *Manual for Beck Depression Inventory-II*. Psychological Corporation: San Antonio, TX, 1996.

17. Zhang D, Lin X, Raz J, Sowers MF. Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association* 1998; **93**:710–719.
18. Green PJ, Silverman BW. *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall: London, 1994.
19. Green PJ. Penalized likelihood for general semi-parametric regression models. *International Statistical Review* 1987; **55**:245–259.
20. Lin X, Breslow NE. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* 1996; **91**:1007–1016.