Biostatistics (2012), **13**, 2, *pp.* 355–368
doi:10.1093/biostatistics/kxr041
Advance Access publication on November 30, 2011

A marginalized conditional linear model for longitudinal binary data when informative dropout occurs in continuous time

LI SU*
MRC Biostatistics Unit, Robinson Way, Cambridge CB2 0SR, UK
li.su@mrc-bsu.cam.ac.uk

SUMMARY

Within the pattern-mixture modeling framework for informative dropout, conditional linear models (CLMs) are a useful approach to deal with dropout that can occur at any point in continuous time (not just at observation times). However, in contrast with selection models, inferences about marginal covariate effects in CLMs are not readily available if nonidentity links are used in the mean structures. In this article, we propose a CLM for long series of longitudinal binary data with marginal covariate effects directly specified. The association between the binary responses and the dropout time is taken into account by modeling the conditional mean of the binary response as well as the dependence between the binary responses given the dropout time. Specifically, parameters in both the conditional mean and dependence models are assumed to be linear or quadratic functions of the dropout time; and the continuous dropout time distribution is left completely unspecified. Inference is fully Bayesian. We illustrate the proposed model using data from a longitudinal study of depression in HIV-infected women, where the strategy of sensitivity analysis based on the extrapolation method is also demonstrated.

Keywords: Bayesian analysis; HIV/AIDS; Marginal model; Missing data; Sensitivity analysis.

1. INTRODUCTION

Dropout occurs commonly in longitudinal studies. For example, in the HIV Epidemiology Research Study (HERS), a HIV cohort study of 1310 women from 1993 to 2000, it was of interest to examine the time course of depression (defined as whether the Center for Epidemiologic Studies Depression Scale is equal to or greater than 16) in HIV-infected women and other associated factors (Smith *and others*, 1997; Ickovics *and others*, 2001; Su and Hogan, 2010). At baseline, the HERS women were scheduled to be followed up every 6 months for 12 visits. However, the dropout rate in the HERS was appreciable and only 173 women had a depression observation at the 12th visit among the 753 women who were HIV-infected at baseline and did not die with HIV-related reasons during the study period. Moreover, previous studies have suggested that the dropout could be related to the disease progression and associated depressive symptoms (Ickovics *and others*, 2001; Roy and Daniels, 2008; Su and Hogan, 2010). As the

*To whom correspondence should be addressed.

actual measurement times correspond to assessment dates and vary across women (see Figure 1 of Su and Hogan, 2010), following Su and Hogan (2010), in this article the dropout in the HERS is considered to occur in continuous time.

When dropout depends on the unobserved response at the time of dropout, or at future times, even after conditioning on the observed data, it is called "informative" or "nonignorable." To deal with informative dropout, a variety of model-based approaches, including "selection" models (SMs), "pattern mixture" models (PMMs), and "shared parameter" models have been proposed for the joint modeling of the response and dropout processes (Wu and Carroll, 1988; Diggle and Kenward, 1994; Follman and Wu, 1995; Ten Have *and others*, 1998; Wu and Bailey, 1989; Little, 1993, 1994; Hogan and Laird, 1997; Wulfsohn and Tsiatis, 1997; Henderson *and others*, 2000; Tsiatis and Davidian, 2004; Ibrahim and Molenberghs, 2009). Semiparametric approaches were also proposed to adjust for the dependence of the dropout time on the unobserved responses (Rotnitzky *and others*, 1998; Scharfstein *and others*, 1999; Lin and Ying, 2003; Wilkins and Fitzmaurice, 2007).

Within the PMMs framework, conditional linear models (CLMs) by Wu and Bailey (1989) are a useful approach to deal with dropout that can occur at any point in continuous time (not just at observation times). However, one disadvantage of CLMs and PMMs compared with SMs is that their parameters usually lack a direct interpretation in terms of marginal covariate effects if nonidentity link functions are used in the mean structures (Wilkins and Fitzmaurice, 2007; Roy and Daniels, 2008; Su and Hogan, 2010). For some scenarios with only treatment groups and measurement times as the covariates, we can obtain the marginal summaries for covariate strata by averaging the response distributions over the dropout patterns (Fitzmaurice and Laird, 2000; Su and Hogan, 2010). When a number of confounders or quantitative covariates are present, a simple summary of the marginal covariate effects might not be immediately available in a CLM or PMM.

To overcome this limitation, several PMMs have been proposed. Building upon log-linear models, Wilkins and Fitzmaurice (2006) developed a marginalized PMM for short sequences of binary data, where the conditional dependencies among the responses and between the responses and dropout patterns are specified separately in addition to the marginal mean model. To avoid the proliferation of nuisance parameters in full likelihood approaches, Wilkins and Fitzmaurice (2007) proposed a PMM using the semiparametric moment-based approach. Focusing on the scenarios with many unique dropout patterns, Roy and Daniels (2008) developed a PMM where the marginal mean follows a generalized linear model and the mean conditional on the latent class and random effects is specified separately. However, mainly because of the concerns about sample size per dropout pattern and model parsimony, these models may not be directly applicable to the situation where measurement times are irregular across individuals and dropout can occur at any point in continuous time.

In this article, within the Bayesian paradigm, we propose a marginalized conditional linear model (MCLM) to deal with continuous-time informative dropout for long sequences of binary data when the target of inference is the marginal covariate effects. Given the dropout time, models for the mean and dependence (including serial dependence and nondiminishing dependence) structures of the binary responses are specified separately (Heagerty, 2002; Schildcrout and Heagerty, 2007; Roy and Daniels, 2008), while parameters in both models are allowed to depend on the dropout time through linear or quadratic formulations similarly as in the original CLMs. With marginal covariates effects directly specified, we then marginalize the conditional mean over the unspecified dropout time distribution through Rubin's Bayesian bootstrap (Rubin, 1981). Following Su and Hogan (2010), we choose to build the MCLM within the Bayesian paradigm in order to avoid extra bootstrapping of the continuous dropout time for standard error estimation when the delta method fails in nonparametric frequentist approaches (Hogan *and others*, 2004).

One advantage of PMMs and CLMs over others is that the unidentifiable part of the model for extrapolating missing data can be distinguished from those identifiable from the observed data, which facilitates

substantive critique and empirical sensitivity analysis (Little and Wang, 1996; Daniels and Hogan, 2000, 2008; Rotnitzky *and others*, 2001). In this article, we will illustrate the unverifiable assumptions in the proposed MCLM and demonstrate sensitivity analysis strategies based on the extrapolation method (Rizopoulos *and others*, 2007) using the HERS depression data.

The remainder of this article is organized as follows. In Section 2, we introduce the model. Computational details are provided in Section 3. In Section 4, we apply our methods to the HERS depression data and conduct a sensitivity analysis to assess the impact of unverifiable assumptions on the scientific conclusions. Conclusions and discussion follow in Section 5.

2. MODEL

Let D_i denote the dropout time for the ith individual ($i = 1, \ldots, N$). At continuous-time points $t_{i1}, \ldots, t_{in_i}$ ($t_{in_i} \leqslant D_i$), we observe the binary responses $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^{\mathrm{T}}$ and the $n_i \times p$ exogenous covariate matrix $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})^{\mathrm{T}}$ (e.g. external or fixed by study design). When the dropout is informative in the sense that it is related to the unobserved responses given the observed data, we need to jointly model $(\mathbf{Y}_i, \mathbf{X}_i, D_i)$. Specifically, building on the marginalized transition and latent variable model (mTLV) by Schildcrout and Heagerty (2007) for long series of binary data, we develop an MCLM by allowing the conditional mean and dependence given the dropout time as well as the marginal mean to be separately specified. Basically, our model formulation involves 4 components:

(a) Marginal model for the mean of the jth response, $\mu_{ij}^{\mathrm{M}} = \mathrm{E}(Y_{ij}|\mathbf{x}_{ij})$.

(b) Conditional model for the mean of the jth response given the dropout time (pattern) D_i, $\mu_{ij}^{\mathrm{C}} = \mathrm{E}(Y_{ij}|\mathbf{x}_{ij}, D_i)$.

(c) Dependence model for the responses given the dropout time D_i, $\mathrm{E}(Y_{ij}|Y_{ij-1}, \ldots, Y_{i1}, b_i, \mathbf{x}_{ij}, D_i)$, where b_i is an individual-level random intercept.

(d) Marginal model for the dropout time distribution, $f(D_i|\mathbf{X}_i)$.

To specify (a), we assume that

$$g(\mu_{ij}^{\mathrm{M}}) = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}, \tag{2.1}$$

where $g(\cdot)$ is a link function, $j = 1, \ldots, n_i$, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of marginal regression coefficients. Both (b) and (c) capture the association between binary responses and the dropout time. In particular, we assume that

$$g(\mu_{ij}^{\mathrm{C}}) = \delta_{ij} + \mathbf{z}_{ij}^{\mathrm{T}}\boldsymbol{\alpha}(D_i), \tag{2.2}$$

where $\mathbf{z}_{ij}$ is a subset of $\mathbf{x}_{ij}$, $\boldsymbol{\alpha}(\cdot)$ is a $q \times 1$ vector of linear or quadratic functions of the dropout time D_i. For identifiability, we use a constraint on $\boldsymbol{\alpha}(\cdot)$ such that $\boldsymbol{\alpha}(T) = 0$, where T indicates the time for study end or the maximum follow-up in the study. Because of the following relationship between (2.1) and (2.2)

$$\mathrm{E}(Y_{ij}|\mathbf{x}_{ij}) = \sum_{D_i} \mathrm{E}(Y_{ij}|\mathbf{x}_{ij}, D_i)\, f(D_i|\mathbf{X}_i),$$

the δ_{ij} term is implicitly a function of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}(\cdot)$, the parameters for (d) and the covariates $\mathbf{x}_{ij}$.

Basically, the model in (2.1) is chosen to obtain the desired target of inference: marginal covariate effects. The conditional mean model in (2.2) specifies how the response mean for individuals differ by their dropout times D_i and this is consistent with the specification in the original CLM by Wu and Bailey (1989). In other words, we allow the response mean to depend on the dropout process using a parametric formulation (e.g. linear or quadratic functions) as in a CLM. It must be recognized that unverifiable

assumptions in (b) influence the inferences about the parameters in (a). For example, in the HERS example, if $\mathbf{z}_{ij}$ includes the time variable t_{ij} and its corresponding coefficient is $\alpha(D_i) = \theta_0 + \theta_1 D_i$, then early dropouts were allowed to have different time slopes of depression compared to later dropouts. However, here we assume that the time slope before dropout at D_i can be extrapolated to characterize the time slope after dropout, where no data after dropout were available to assess the validity of assumption. Therefore, sensitivity analysis is required, and we will demonstrate the corresponding strategies using the HERS example in Section 4.

The purpose of (c) is to account for the dependence between binary responses within individuals and allow full likelihood-based inference for long series of binary data. Following Schildcrout and Heagerty (2007), we consider both serial dependence with a Markov component and nondiminishing dependence with a random intercept. Specifically, the mean of Y_{ij}, conditional on its history $Y_{i1}, \ldots, Y_{ij-1}$, the random intercept b_i, the covariates $\mathbf{x}_{ij}$ as well as the dropout time D_i is $\mu_{ij}^{\mathrm{S}} = \mathrm{E}(Y_{ij}|Y_{ij-1}, \ldots, Y_{i1}, b_i, \mathbf{x}_{ij}, D_i) = \mathrm{E}(Y_{ij}|Y_{ij-1}, b_i, \mathbf{x}_{ij}, D_i)$ and

$$\mathrm{logit}(\mu_{ij}^{\mathrm{S}}) = \Delta_{ij} + \gamma_{ij}(D_i) \cdot Y_{ij-1} + b_i, \quad b_i \sim N\{0, \sigma^2(D_i)\}. \tag{2.3}$$

Although a logit link function is used here, note that any valid link function can be adopted (Heagerty, 2002). For simplicity, the dependence of Δ_{ij}, $\gamma_{ij}(D_i)$, and $\sigma^2(D_i)$ on $\mathbf{x}_{ij}$ is suppressed for now. Given b_i, the log odds ratio $\gamma_{ij}(D_i)$ measures the serial dependence between Y_{ij} and the immediate previous response Y_{ij-1} among those who drop out at D_i; b_i introduces the nondiminishing (long-range) dependence between responses within individuals. The intercept Δ_{ij} is determined such that the conditional mean model in (2.2) and the dependence model in (2.3) are simultaneously satisfied (Schildcrout and Heagerty, 2007). In other words, Δ_{ij} is the solution to

$$\mathrm{E}(Y_{ij}|\mathbf{x}_{ij}, D_i) = \mathrm{E}_{b_i}[\mathrm{E}_{Y_{i,j-1}|b_i}\{\mathrm{logit}^{-1}(\Delta_{ij} + \gamma_{ij}(D_i) \cdot Y_{ij-1} + b_i)\}].$$

Further, the serial dependence measure $\gamma_{ij}(D_i)$ and random intercept variance $\sigma^2(D_i)$ can be modeled via

$$\gamma_{ij}(D_i) = \mathbf{w}_{ij}^{\mathrm{T}} \boldsymbol{\phi}(D_i), \tag{2.4}$$

$$\log\{\sigma^2(D_i)\} = \mathbf{v}_i^{\mathrm{T}} \boldsymbol{\psi}(D_i), \tag{2.5}$$

where $\mathbf{w}_{ij}$ and $\mathbf{v}_i$ are subsets of $\mathbf{x}_{ij}$, $\boldsymbol{\phi}(\cdot)$, and $\boldsymbol{\psi}(\cdot)$ are vectors of linear or quadratic functions of the dropout time D_i. For example, $\mathbf{w}_{ij}$ can include the gap time between 2 consecutive visits, which accommodates irregular spacing of measurement times. $\mathbf{v}_i$ can include treatment group membership such that the random intercept variance differs by treatment groups, but this treatment effect will vary by the dropout time.

By allowing the dependence parameters to vary by D_i in (2.3), our MCLM has a different within-individual dependence structure from a CLM that only allows the mean parameters, e.g. in (2.2), to vary by D_i. It is well known that with complete data and likelihood-based approaches, properly modeling the within-individual dependence structure can affect the variability estimates more than the point estimates of the mean parameters (Diggle *and others*, 2002). However, with missing data, even point estimates can be biased if the dependence structure is not carefully modeled (Kurland and Heagerty, 2004; Daniels and Hogan, 2008). By including covariates and allowing the dependence on the dropout time in the dependence model, we are trying to minimize these biases in our approach.

Finally, component (d) needs to be specified to complete the joint distribution for $(\mathbf{Y}_i, \mathbf{X}_i, D_i)$. Basically, this can be modeled using any event time distribution, where the dependence on $\mathbf{X}_i$ can be checked

by standard event time regression analysis methods. Here, we adopt a nonparametric approach and allow $f(D_i|\mathbf{X}_i)$ to be completely unspecified within the strata of $\mathbf{X}_i$. Following Su and Hogan (2010), we use Rubin's Bayesian bootstrap (Rubin, 1981) to obtain the posterior of $f(D_i|\mathbf{X}_i)$ for the observed dropout times (see details in the Supplementary material available at *Biostatistics* online).

3. Computational details

We let $\boldsymbol{\theta}$ denote the set of parameters that characterize the functions $\boldsymbol{\alpha}(\cdot)$ in the conditional mean model in (2.2), let $\boldsymbol{\lambda}$ denote the set of parameters that characterize the dependence model in (2.3–2.5), and let $\boldsymbol{\pi}$ index the dropout time distribution $f(D_i|\mathbf{X}_i;\boldsymbol{\pi})$. The likelihood contribution from the response data of the ith individual is

$$\begin{aligned} & f(\mathbf{y}_i|b_i,\mathbf{X}_i,D_i;\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\lambda}) \\ = \; & f(y_{i1}|b_i,\mathbf{x}_{i1},D_i;\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\lambda})\, f(y_{i2}|y_{i1},b_i,\mathbf{x}_{i2},D_i;\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\lambda}),\ldots,f(y_{in_i}|y_{in_i-1},b_i,\mathbf{x}_{in_i},D_i;\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\lambda}) \\ = \; & \prod_{j=1}^{n_i}(\mu_{ij}^{S})^{y_{ij}}(1-\mu_{ij}^{S})^{(1-y_{ij})}. \end{aligned}$$

The posterior distribution for the parameters in an MCLM is proportional to

$$\prod_{i=1}^{N}\{f(\mathbf{y}_i|b_i,\mathbf{X}_i,D_i;\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\lambda})\, f(b_i|\boldsymbol{\lambda})\, f(D_i|\mathbf{X}_i;\boldsymbol{\pi})\}\, p(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\lambda})\, p(\boldsymbol{\pi}),$$

where $p(\cdot)$ is a prior density function. We follow the specification of the original PMMs in the Bayesian paradigm (Daniels and Hogan, 2008) and assume that the priors for $\boldsymbol{\pi}$ are independent of the priors for $(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\lambda})$. It follows that $\boldsymbol{\pi}$ is not a part of the posterior for $(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\lambda})$ and the inference for $\boldsymbol{\pi}$ can be based on the marginal likelihood $\prod_{i=1}^{N} f(D_i|\mathbf{X}_i;\boldsymbol{\pi})$.

We standardize the continuous covariates to have mean 0 and standard deviation 0.5 as recommended by Gelman (2008) and assign independent t priors with 7 degrees of freedom and scale 2.5 (Gelman *and others*, 2008) to the elements of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ as well as those serial dependence parameters within $\boldsymbol{\lambda}$ in (2.4). Independent $N(0,7)$ priors are used for random intercept variance parameters (at log scale) within $\boldsymbol{\lambda}$ in (2.5). The Markov Chain Monte Carlo (MCMC) for posterior sampling is implemented in MATLAB (version 7.1) and more details can be found in the Supplementary material available at *Biostatistics* online.

4. Example

As briefly described in Section 1, our goal is to characterize the depression time course for the 753 HERS women. We exclude those women who died due to HIV-related reasons during the study period because we consider that response-related death mixed with dropout (Kurland and Heagerty, 2005) is another problem that needs further research and is beyond the scope of this article. Depression was measured using the Center for Epidemiologic Studies Depression Scale (CES-D), which ranges from 0 to 60 with larger scores indicating the presence of more symptoms. Following Su and Hogan (2010), we focus on the dichotomized CES-D data that commonly define clinically significant depression in HIV research (Radloff, 1977; Ickovics *and others*, 2001; Cook *and others*, 2004; Leserman, 2008). The analysis of the continuous and binary HERS CES-D data using the original PMM approach (i.e. the marginal covariate effects are not directly specified) can be found in Sections 4.1 and 4.2 of Su and Hogan (2010).

The covariates of interest include baseline characteristics, such as race (Black/White/Latina and others) and initial disease stage (defined as whether the baseline CD4 count is >200), as well as the time variable (in the unit of days). Following Gelman (2008), the time variable is standardized to have mean 0 and standard deviation 0.5.

4.1 *Models under comparison*

We fit an mTLV (Schildcrout and Heagerty, 2007) and an MCLM to the HERS depression data. Assuming "missingness at random" (MAR) and the prior independence of the parameters in the response model and the dropout time distribution, the missingness is ignorable in the mTLV (Little and Rubin, 2002). In both models, the marginal mean of depression follows:

$$\begin{aligned}\operatorname{logit}(\mu_{ij}^{\mathrm{M}}) &= \beta_0 + \beta_1 I(\text{Black}) + \beta_2 I(\text{Latina}) + \beta_3 I(\text{baseline CD4} > 200)\\ &\quad + \beta_4 I(\text{baseline CD4} \leqslant 200) t_{ij} + \beta_5 I(\text{baseline CD4} \leqslant 200) t_{ij}^2\\ &\quad + \beta_6 I(\text{baseline CD4} > 200) t_{ij} + \beta_7 I(\text{baseline CD4} > 200) t_{ij}^2, \end{aligned} \tag{4.1}$$

where $I(\cdot)$ is the indicator function. The quadratic term of the time variable is included to allow more flexibility to characterize the depression time course.

In the mTLV, no conditional mean model given the dropout time is needed, while the dependence structure includes constant first-order serial dependence and a random intercept for nondiminishing dependence:

$$\operatorname{logit}(\mu_{ij}^{\mathrm{S}}) = \Delta_{ij} + \gamma \cdot Y_{ij-1} + b_i, \quad b_i \sim N(0, \sigma^2), \ \log(\sigma^2) = \psi.$$

The conditional mean model in the MCLM is specified as follows:

$$\begin{aligned}\operatorname{logit}(\mu_{ij}^{\mathrm{C}}) &= \delta_{ij} + \theta_1 D_i^* I(\text{baseline CD4} > 200)\\ &\quad + \theta_2 D_i^* I(\text{baseline CD4} \leqslant 200) t_{ij}\\ &\quad + \theta_3 D_i^* I(\text{baseline CD4} > 200) t_{ij}, \end{aligned} \tag{4.2}$$

where the standardized dropout time $D_i^* = (D_i - T)/T$ is within $[-1, 0]$, and $T = 2093$ corresponds to the maximum follow-up days in the HERS. The choice for covariates here is based on the analysis reported in Su and Hogan (2010), where regression coefficients for races were found to be relatively constant over the dropout time. Basically, we allow the regression coefficients in (4.2) to vary as linear functions of the dropout time, and if women reached maximum follow-up in the HERS, their regression coefficients are assumed to be 0 for identifiability purpose because we have specified a separate model (4.1) for the marginal mean of depression. Further, both the first-order serial dependence and the nondiminishing dependence are assumed to be linearly related to the dropout time as follows:

$$\begin{aligned}\operatorname{logit}(\mu_{ij}^{\mathrm{S}}) &= \Delta_{ij} + \gamma_{ij}(D_i) \cdot Y_{ij-1} + b_i, \quad b_i \sim N\{0, \sigma^2(D_i)\},\\ \gamma_{ij}(D_i) &= \lambda_0 + \lambda_1 D_i/T,\\ \log\{\sigma^2(D_i)\} &= \lambda_2 + \lambda_3 D_i/T. \end{aligned} \tag{4.3}$$

Note that if $\theta_1 = \theta_2 = \theta_3 = \lambda_1 = \lambda_3 = 0$, the MCLM is reduced to the mTLV under MAR.

For calculation of the intercept δ_{ij}, we need to obtain the posterior samples of $f(D_i|\mathbf{X}_i)$. Initially, we use Cox regression analysis methods to check the relationship between the discrete covariates (race, baseline CD4 count) and the dropout time distribution. The Whites and Blacks were less likely to drop out than the Latinas and other races; the patients with baseline CD4 count > 200 were also less likely to drop out. Therefore, we have $f(D_i|\mathbf{X}_i) \neq f(D_i)$ in the HERS data and the Bayesian bootstrapping for the observed dropout times is conducted within the race and baseline CD4 groups.

The priors assigned for $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \theta_1, \theta_2, \theta_3$ and γ, λ_0, λ_1 are t priors with 7 degrees of freedom and scale 2.5. The $N(0, 7)$ priors are used for ψ, λ_2, and λ_3. For both models, we run 2 MCMC chains and check the convergence after 5000-iteration burn-in period using history plots. The computing time for the mTLV and MCLM fits of the HERS example (6505 observations) is approximately 2 and 8 h per 1000 iterations, respectively, on our machine (2.59 GHz CPU, 32 GB RAM). Pooled posterior samples of size 10 000 are used for inference.

4.2 *Results*

Table 1 presents the results from both the mTLV and the MCLM. In the MCLM, both the conditional mean regression coefficients and the dependence parameters indicate some associations with the dropout time. Specifically, earlier dropouts are shown to have larger main effect of baseline CD4 count ($\hat{\theta}_1$ [posterior mean] $= -0.22$, 95% credible interval (CI) $= [-0.77; 0.34]$). If their baseline CD4 counts are $\leqslant 200$, earlier dropouts had larger time slopes than later dropouts ($\hat{\theta}_2 = -0.46$, 95% CI $= [-1.67; 0.95]$), while if their baseline CD4 counts are >200, later dropouts had larger time slopes than earlier dropouts ($\hat{\theta}_3 = 0.20$, 95% CI $= [-0.64; 0.93]$). In other words, those early dropouts who had severe immunosuppression at

Table 1. *Results from the HERS analysis. The posterior means, standard deviations (SD), and the* 95% *CI are reported for the marginal regression coefficients, conditional mean, and dependence parameters from the fitted MCLM and mTLV*

	MCLM				mTLV			
Parameter	Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%
β_0	0.28	0.22	−0.15	0.77	0.32	0.18	−0.06	0.63
β_1	−0.19	0.13	−0.45	0.05	−0.26	0.11	−0.47	−0.04
β_2	0.37	0.16	0.05	0.71	0.24	0.14	−0.03	0.53
β_3	0.00	0.21	−0.37	0.39	0.02	0.18	−0.29	0.40
β_4	−0.17	0.21	−0.62	0.18	−0.25	0.18	−0.57	0.09
β_5	−0.59	0.28	−1.12	0.01	−0.66	0.29	−1.18	−0.05
β_6	−0.29	0.08	−0.45	−0.12	−0.28	0.04	−0.37	−0.20
β_7	0.19	0.10	0.00	0.39	0.24	0.10	0.02	0.40
θ_1	−0.22	0.28	−0.77	0.34				
θ_2	−0.46	0.68	−1.67	0.95				
θ_3	0.20	0.39	−0.64	0.93				
λ_0	0.63	0.45	−0.26	1.53				
λ_1	0.67	0.52	−0.36	1.70				
γ					1.19	0.09	1.02	1.36
λ_2	0.26	0.21	−0.17	0.64				
λ_3	0.36	0.25	−0.10	0.87				
ψ					0.55	0.05	0.46	0.66
σ^2					1.74	0.09	1.58	1.93

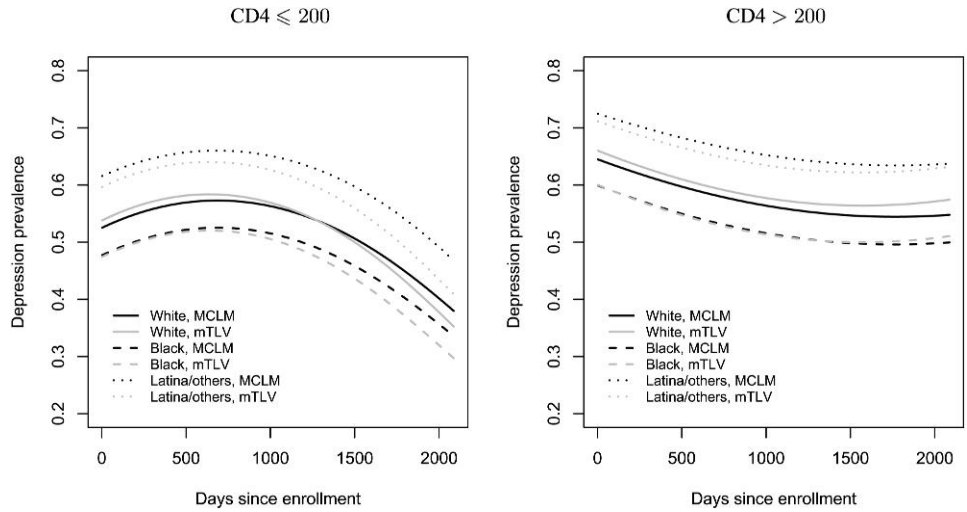

Fig. 1. Posterior mean estimates of depression prevalence by race and baseline CD4 groups from the mTLV and MCLM fits of the HERS depression data.

baseline (CD4 $\leqslant$ 200) tended to have higher change rates of depression than later dropouts, but for patients who had baseline CD4 counts over 200, this pattern was reversed. However, given the fact that women with baseline CD4 $>$ 200 were less likely to drop out, the influence of dropout on the binary responses is relatively small for them. Finally, the first-order serial dependence and nondiminishing dependence are also shown to vary positively with the dropout time ($\hat{\lambda}_1 = 0.67$, 95% CI $= [-0.36; 1.70]$; $\hat{\lambda}_3 = 0.36$, 95% CI $= [-0.10; 0.87]$). Overall, compared with the mTLV fit, the MCLM adjusted the marginal depression prevalence profiles upward at the later period of followup for the group with baseline CD4 $\leqslant$ 200 and the largest adjustment occurred for the Latina/others group (left panel of Figure 1). On the other hand, the marginal depression prevalence profiles for both the White and Latina/others groups were shifted slightly if their baseline CD4 counts are $>$200, but the general time trends remain stable (right panel of Figure 1).

Recall that when $\theta_1 = \theta_2 = \theta_3 = \lambda_1 = \lambda_3 = 0$, the MCLM is reduced to the mTLV under MAR. Therefore, if we assume that MAR is violated, the parameters θ_1, θ_2, θ_3, λ_1, and λ_3 will quantify the degree to which MAR fails to hold. Since the estimated 95% CIs for all these parameters cover zero, there is no strong evidence from the HERS data that the MCLM fit is preferred to the mTLV fit under MAR. The goodness of fit of the MCLM was further assessed by posterior predictive checks based on completed-data plots obtained by multiple imputation of the missing responses (Gelman *and others*, 2005; see details in the Supplementary material available at *Biostatistics* online.).

In summary, we observed that, regardless of their baseline CD4 counts, Latinas and other race groups had higher depression prevalence over time as compared with Blacks and Whites. Given their races, women with different baseline CD4 counts all had downward trends in depression prevalence over time. There is no sufficient evidence from the data to show that these trends differ (see Figure 3).

4.3 *Sensitivity analysis*

In previous section, the mTLV and MCLM appeared to have similar fits to the observed HERS CES-D data. However, the assumptions for extrapolating the missing responses given the observed data are different in these models. In the mTLV, MAR is assumed such that the conditional distribution of missing

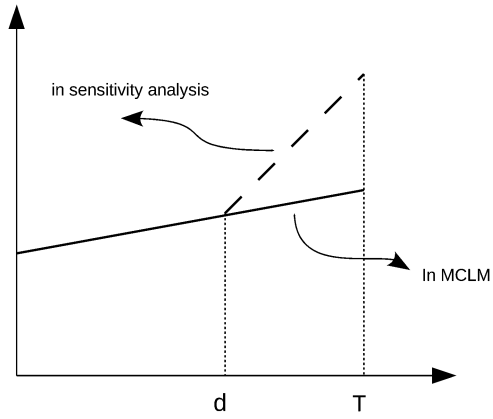

Fig. 2. Illustration of the unverifiable assumption made in the MCLM: the horizontal axis represents time since enrollment, the vertical axis represents the conditional mean of depression at the logit scale, and T represents the study end or maximum follow-up. At time d, some participants dropped out of the HERS. Therefore, the depression time slope after d is not estimable from the observed data. In the MCLM, the depression time slope before dropout is extrapolated to the time slope after dropout (the solid line). In the corresponding sensitivity analysis, we allow the time slope after dropout to follow a piecewise linear model (the dashed line). That is, the time slope before dropout is not necessarily equal to the time slope after dropout.

depression responses given the observed data for those who remained in the study at d is the same as the corresponding conditional distribution for those who left the study at d (Molenberghs *and others*, 1998), i.e.

$$f(y_{ij}|y_{i1},\ldots,y_{ij-1},\mathbf{X}_i,t_{ij-1}\leqslant d<t_{ij})=f(y_{ij}|y_{i1},\ldots,y_{ij-1},\mathbf{X}_i,t_{ij}\leqslant d).$$

In the MCLM, we assume that given the dropout time d and the covariates, missing data after dropout share the same parameters as observed data before dropout. For example, in the HERS example, it is assumed that given their baseline CD4 counts, women with observed dropout at d had the same time slope for $t_{ij}>d$ as for $t_{ij}\leqslant d$. This is clear from the illustration in Figure 2. The time slope after dropout cannot be obtained from the observed data and has to be extrapolated in the MCLM. Both assumptions in the mTLV and MCLM cannot be verified from the observed data and sensitivity analysis is required (Little and Wang, 1996; Daniels and Hogan, 2000; Rotnitzky *and others*, 2001; Daniels and Hogan, 2008).

We demonstrate an example of sensitivity analysis regarding the abovementioned assumption in the MCLM. The strategy of sensitivity analysis for the MCLM can be based on the extrapolation method (Rizopoulos *and others*, 2007). Basically, we assume a different time slope for $t_{ij}>d$, i.e. assume a continuous piecewise linear model with a change point at d (see Figure 2) . For the group with baseline CD4 $\leqslant$ 200, we assume the conditional mean model as follows:

$$\text{logit}(\mu_{ij}^{\text{C}})=\delta_{ij}+\theta_2 D_i^* t_{ij}+\omega_0(D_i^*)(t_{ij}-\tilde{D}_i)_+$$

where $(x)_+=x$ if $x>0$ and 0 otherwise, $\tilde{D}_i$ is the observed dropout time standardized to have the same scale of t_{ij} and $\omega_0(D_i^*)$ is the change of the slope after dropout that is different across specific dropout times. The model for baseline CD4 $>$ 200 is similar but with $\omega_1(D_i^*)$ representing the slope change after dropout:

$$\text{logit}(\mu_{ij}^{\text{C}})=\delta_{ij}+\theta_1 D_i^*+\theta_3 D_i^* t_{ij}+\omega_1(D_i^*)(t_{ij}-\tilde{D}_i)_+.$$

In principle, sensitivity analysis should be based on the parameters that cannot be identified by the observed data, such as $\omega_0(D_i^*)$ and $\omega_1(D_i^*)$. We assume a simple functional form for $\omega_0(\cdot)$ and $\omega_1(\cdot)$:

$$\omega_0(D_i^*) = -a_0 D_i^* = -a_0(D_i - T)/T,$$

$$\omega_1(D_i^*) = -a_1 D_i^* = -a_1(D_i - T)/T.$$

Thus, when $D_i = T$ is the maximum follow-up, no adjustment is made about the slope after dropout (i.e. for study completers), while the slope is adjusted upward by a_0 or a_1 when $D_i = 0$, that is, when the participants dropped out after the enrolment visit. For example, when $a_0 = 2$ and some HERS women with baseline CD4 $\leqslant 200$ dropped out the study at 1 year ($d = 365$), we assume that before dropout their time slopes are $\hat{\theta}_2(d - T)/T = -0.46(365 - 2093)/2093 = 0.38$, but their time slopes after dropouts are $(\hat{\theta}_2 - a_0)(d - T)/T = (-0.46 - 2)(365 - 2093)/2093 = 2.03$.

In Figure 3, we fix the nonidentifiable parameters a_0 and a_1 at various combinations of their values and compare the estimated prevalence differences of depression between baseline CD4 groups for White women to check their sensitivity to a_0 and a_1. The results for Latinas and Blacks are similar. Estimates for the early time period after enrollment are close across all model fits, including the original mTLV and MCLM fits. Depending on specific combination of a_0 and a_1, the baseline CD4 group difference in depression prevalence is adjusted downward or upward at the later follow-up period. However, the pointwise 95% credible bands from the MCLM fit cover all these estimated depression prevalence profiles even when we choose a_0 and a_1 at relatively large values (i.e. large changes in time slopes after dropout are assumed). In practice, caution needs to be taken about how to choose values or assign priors for sensitivity parameters. In this particular example, we only showed a simple case by setting them as constants (i.e. assign 1–0 point mass prior). Informative priors on sensitivity parameters can also be used based on expert opinions and prior elicitation from previous studies (Daniels and Hogan, 2008).

5. DISCUSSION

We have proposed a new model for dealing with informative dropout that occurs in continuous time. The marginal covariate effects of interest are directly modeled and the relationship between the binary responses and the dropout process is specified using linear or quadratic formulations in both conditional mean and dependence models. In our Bayesian approach, the continuous dropout time distribution is not modeled and its uncertainty is properly taken into account by Bayesian bootstrapping when obtaining marginal covariate effects.

In this article, we focused on the scenario with dropouts only. There were 173 HERS women who actually finished 12 scheduled visits. Su and Hogan (2010) distinguished these administratively censored patients from dropouts and allowed them to form a separate pattern in their varying coefficient modeling approach to these data. They found that the parameter estimates for responses from these patients were similar to those from later dropouts (e.g. those who finished 11 visits). Therefore, for simplicity, in the analysis reported in Section 4, we treated the follow-up times of administratively censored patients (ranged from 1952 to 2093 days) the same as the dropout times. In practice, distinguishing administrative censoring from dropouts might be more important when patients have staggered entry and informative dropout is present (Li and Schluchter, 2004). The proposed MCLM can be extended by allowing the parameters to depend on administrative censoring times through linear or quadratic functions, but these functions are distinct from those for dropout times.

We have assumed that the relationship between the dropout time and binary responses follows the linear or quadratic formulations. Unspecified smooth functions modeled by penalized splines (Ruppert *and others*, 2003) can be used to allow more flexibility for this relationship (Hogan *and others*, 2004; Su and Hogan, 2010). However, we found that the estimation of the dependence parameters is usually less

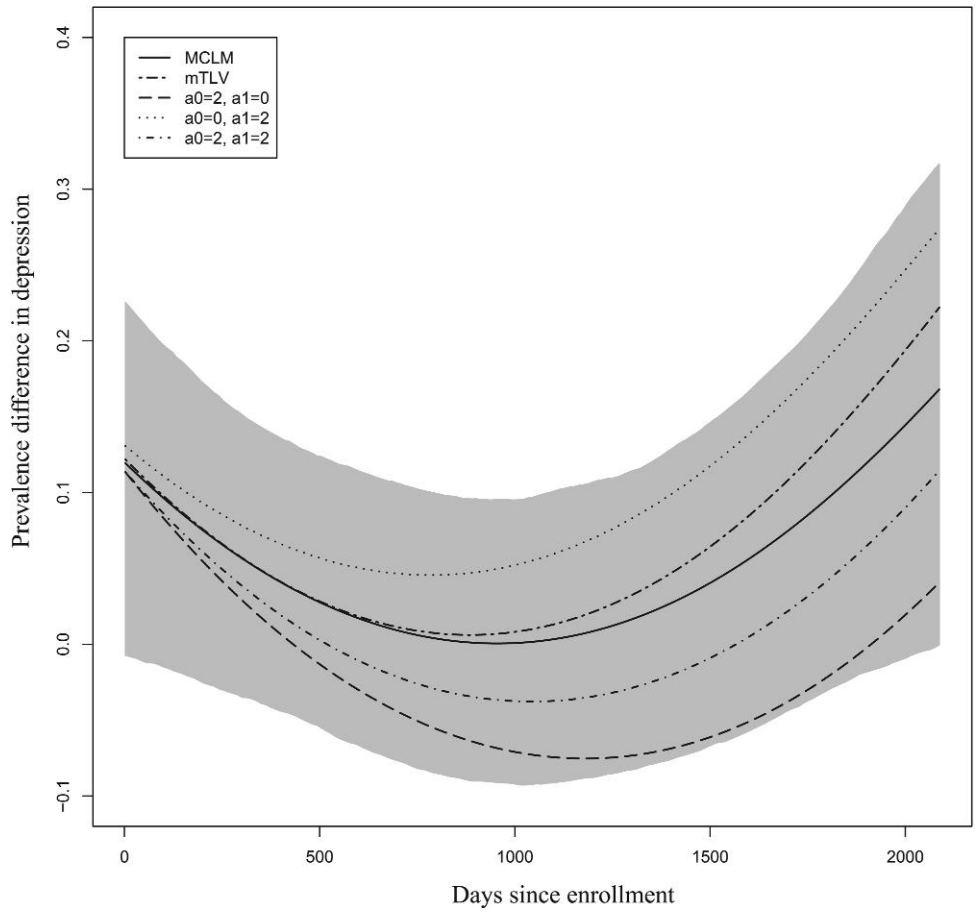

Fig. 3. Sensitivity analysis for the MCLM of the HERS depression data: posterior mean estimates of the prevalence difference of depression between baseline CD4 groups (CD4 $>$ 200 vs. CD4 $\leqslant$ 200) for White women with fixed values for sensitivity parameters a_0 and a_1 compared with the results from the mTLV and MCLM (the results for Latinas and Blacks are similar); gray shades represent corresponding pointwise 95% credible bands from the MCLM fit.

stable than for the mean parameters due to the sparsity nature of the binary data. Therefore, incorporating unspecified smooth functions in the mean structure of the MLCM is a more practical extension and the same penalized spline approach described in Su and Hogan (2010) can be applied straightforwardly.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The author would like to thank the associate editor and 2 referees for their constructive comments that considerably improved this article. Data from the HER study were collected under grant U64-CCU10675 from the US Centers for Disease Control and Prevention. *Conflict of Interest:* None declared.

FUNDING

The Medical Research Council (UK) (unit programme number U105261167).

REFERENCES

COOK, J. A., GREY, D., BURKE, J., COHEN, M.H., GURTMAN, A. C., RICHARDSON, J. L., WILSON, T. E., YOUNG, M. A. AND HESSOL, N. A. (2004). Depressive symptoms and AIDS-related mortality among a multisite cohort of HIV-positive women. *American Journal of Public Health* **94**, 1133–1140.

DANIELS, M. J. AND HOGAN, J. W. (2000). Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout. *Biometrics* **56**, 1241–1248.

DANIELS, M. J. AND HOGAN, J. W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*, Volume 101, Monographs on Statistics and Applied Probability. Boca Ration, FL: CRC Press.

DIGGLE, P. J., HEAGERTY, P., LIANG, K.-Y. AND ZEGER, S. L. (2002). *Analysis of Longitudinal Data*. New York: Oxford University Press.

DIGGLE, P. J. AND KENWARD, M. G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics* **43**, 49–93.

FITZMAURICE, G. M. AND LAIRD, N. M. (2000). Generalized linear mixture models for handling nonignorable dropouts in longitudinal studies. *Biostatistics* **1**, 141–156.

FOLLMAN, D. AND WU, M. C. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics* **51**, 151–168.

GELMAN, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* **27**, 2865–2873.

GELMAN, A., JAKULIN, A., PITTAU, M. G. AND SU, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics* **2**, 1360–1383.

GELMAN, A., VAN MECHELEN, I., VERBEKE, G., HEITJAN, D. F. AND MEULDERS, M. (2005). Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics* **61**, 74–85.

HEAGERTY, P. J. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics* **58**, 342–351.

HENDERSON, R., DIGGLE, P. AND DOBSON, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–480.

HOGAN, J. W. AND LAIRD, N. M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* **16**, 239–257.

HOGAN, J. W., LIN, X. AND HERMAN, B. (2004). Mixtures of varying coefficient models for longitudinal data with discrete or continuous non-ignorable dropout. *Biometrics* **60**, 854–864.

IBRAHIM, J. G. AND MOLENBERGHS, G. (2009). Missing data methods in longitudinal studies: a review. *Test* **18**, 1–43.

ICKOVICS, J. R., HAMBURGER, M. E., VLAHOV, D. *and others* FOR THE HIV EPIDEMIOLOGY RESEARCH STUDY GROUP. (2001). Mortality, CD4 cell count decline, and depressive symptoms among HIV-seropositive women. *Journal of the American Medical Association* **285**, 1466–1474.

KURLAND, B. F. AND HEAGERTY, P. J. (2004). Marginalized transition models for longitudinal binary data with ignorable and non-ignorable drop-out. *Statistics in Medicine* **23**, 2673–2695.

KURLAND, B. F. AND HEAGERTY, P. J. (2005). Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics* **6**, 241–258.

LESERMAN, J. (2008). Role of depression, stress, and trauma in HIV disease progression. *Psychosomatic Medicine* **70**, 539–545.

LI, J. AND SCHLUCHTER, M. D. (2004). Conditional mixed models adjusting for non-ignorable drop-out with administrative censoring in longitudinal studies. *Statistics in Medicine* **23**, 3489–3503.

LIN, D. Y. AND YING, Z. (2003). Semiparametric regression analysis of longitudinal data with informative drop-outs. *Biostatistics* **4**, 385–398.

LITTLE, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125–134.

LITTLE, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika* **81**, 471–483.

LITTLE, R. J. A. AND RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition. New York: John Wiley & Sons.

LITTLE, R. J. A. AND WANG, Y. (1996). Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics* **52**, 98–111.

MOLENBERGHS, G., MICHIELS, B., KENWARD, M. G. AND DIGGLE, P. J. (1998). Monotone missing data and pattern-mixture models. *Statistica Neerlandica* **52**, 153–161.

RADLOFF, L. S. (1977). The CES-D scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement* **1**, 385–401.

RIZOPOULOS, D., VERBEKE, G. AND LESAFFRE, E. (2007). Sensitivity analysis in pattern mixture models using the extrapolation method. Unpublished manuscript.

ROTNITZKY, A., ROBINS, J. M. AND SCHARFSTEIN, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321–1339.

ROTNITZKY, A., SCHARFSTEIN, D., SU, T. L. AND ROBINS, J. (2001). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics* **57**, 103–113.

ROY, J. AND DANIELS, M. J. (2008). A general class of pattern mixture models for nonignorable dropout with many possible dropout times. *Biometrics* **64**, 538–545.

RUBIN, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics* **9**, 130–134.

RUPPERT, D., WAND, M. P. AND CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.

SCHARFSTEIN, D., ROBINS, J. AND ROTNITZKY, A. (1999). Adjusting for nonignorable nonresponse using semiparametric nonresponse models with time dependent covariates (with discussion). *Journal of the American Statistical Association* **94**, 1096–1146.

SCHILDCROUT, J. S. AND HEAGERTY, P. J. (2007). Marginalized models for moderate to long series of longitudinal binary response data. *Biometrics* **63**, 322–331.

SMITH, D. K., WARREN, D. L., VLAHOV, D., SCHUMAN, P., STEIN, M. D., GREENBERG, B. L. AND HOLMBERG, S. D. (1997). Design and baseline participant characteristics of the Human Immunodeficiency Virus Epidemiology Research (HER) Study: a prospective cohort study of human immunodeficiency virus infection in US women. *American Journal of Epidemiolology* **146**, 459–469.

SU, L. AND HOGAN, J. W. (2010). Varying-coefficient models for longitudinal processes with continuous-time informative dropout. *Biostatistics* **11**, 93–110.

TEN HAVE, T. R., KUNSELMAN, A. R., PULKSTENIS, E. P. AND LANDIS, J. R. (1998). Mixed effects logistic regression models for longitudinal binary response data with informative drop-out. *Biometrics* **54**, 367–383.

TSIATIS, A. AND DAVIDIAN, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14**, 809–834.

WILKINS, K. J. AND FITZMAURICE, G. M. (2006). A hybird model for nonignorable dropout in longitudinal binary responses. *Biometrics* **62**, 168–176.

WILKINS, K. J. AND FITZMAURICE, G. M. (2007). A marginalized pattern-mixture model for longitudinal binary data when nonresponse depends on unobserved responses. *Biostatistics* **8**, 297–305.

WU, M. C. AND BAILEY, K. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model (corr: volume 46, p 889). *Biometrics* **45**, 939–955.

WU, M. C. AND CARROLL, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **44**, 175–188.

WULFSOHN, M. S. AND TSIATIS, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–339.

[*Received February 18, 2011; revised September 29, 2011; accepted for publication September 29, 2011*]