

# Mixtures of Varying Coefficient Models for Longitudinal Data with Discrete or Continuous Non-ignorable Dropout

Joseph W. Hogan

*Center for Statistical Sciences & Department of Community Health, Brown University.*

Xihong Lin

*Department of Biostatistics, University of Michigan*

Benjamin Herman

*Center for Statistical Sciences, Brown University.*

March 21, 2003

## Abstract

The analysis of longitudinal repeated measures data is frequently complicated by missing data due to informative dropout. We describe a mixture model for joint distribution for longitudinal repeated measures, where the dropout distribution may be continuous and the dependence between response and dropout is semiparametric. Specifically, we assume that responses follow a varying coefficient random effects model conditional on dropout time, where the regression coefficients depend on dropout time through unspecified nonparametric functions that are estimated using step functions when dropout time is discrete (e.g., for panel data) and using smoothing splines when dropout time is continuous. Inference under the proposed semiparametric model is hence more robust than the parametric conditional linear model. The unconditional distribution of the repeated measures is a mixture over the dropout distribution. We show that estimation in the semiparametric varying coefficient mixture model can proceed by fitting a parametric mixed-effects model and can be carried out on standard software platforms such as SAS. The model is used to analyze data from a recent AIDS clinical trial and its performance is evaluated using simulations.

*Key words:* smoothing splines, linear mixed model, missing data, non-ignorable dropout, pattern-mixture model, pediatric AIDS, equivalence trial, selection bias, clinical trials.

*Correspondence:* Joseph W. Hogan, Center for Statistical Sciences, Box G-H, Brown University, Providence, RI 02912. Phone 401-863-9243. Fax 401-863-9182. Email [jhogan@stat.brown.edu](mailto:jhogan@stat.brown.edu)

# 1 Introduction

## 1.1 Informative dropout

Dropout and other types of missing data are common in long-term longitudinal studies; in many cases, dropout induces a missingness process that is *nonignorable* in the sense that missingness depends probabilistically on unobserved outcomes, even after conditioning on observable information. Most approaches to handling informative dropout in longitudinal data can be viewed as extensions of standard approaches such as multilevel modeling (Laird and Ware, 1982; Diggle, 1988; Breslow and Clayton, 1993) and marginal modeling (Liang and Zeger, 1986; Zeger and Liang, 1988). Likelihood-based approaches include selection models (Wu and Carroll, 1988; Diggle and Kenward, 1994; Follman and Wu, 1995; Ten Have et al, 1998) and mixture models (Little, 1993, 1994; Hogan and Laird, 1997a; Wu and Bailey, 1989). Recent comprehensive surveys of parametric and likelihood-based approaches to handling dropout in longitudinal data can be found in Little (1995), Hogan and Laird (1997b), and Kenward and Molenberghs (1999). Moment-based methods also have been generalized to handle informative dropout under the selection modeling framework. See Robins, Rotnitzky and Zhao (1995), Rotnitzky, Robins and Scharfstein (1998), and Scharfstein, Robins and Rotnitzky (1999).

In this paper, we develop a general mixture modeling approach for continuous longitudinal repeated measures data where measurement times may be irregular across subjects and where dropout might be at continuous times and potentially non-ignorable. The conditional distribution of repeated measures given dropout follows a varying coefficient model (Zhang et al, 1998) where regression coefficients such as intercepts and slopes depend on dropout through unspecified nonparametric functions. The shapes of the functions are estimated using step functions when dropout time is discrete (e.g., for panel data), and using natural cubic smoothing splines (Silverman and Green, 1994) when dropout is continuous. We show that estimation in the proposed varying coefficient mixture model can proceed by fitting an augmented parametric mixed effect model. The complete data distribution is a mixture of the varying coefficient models over the dropout distribution, and the dropout distribution can be left completely unspecified.

This class of models can be viewed as a nonparametric extension of pattern mixture models (Little, 1993, 1994) and conditional linear models (Wu and Bailey, 1989; Hogan and Laird, 1997a) described above. The former author mainly considered panel data, while the latter authors allowed for the dropout time to be continuous but assumed the regression coefficients to be parametric functions of the dropout time. Estimation can therefore be biased if the parametric functions are misspecified. The proposed varying coefficient mixture model relaxes the parametric assumption by providing a unified framework to allow for flexible dependence of the covariate effects on dropout patterns by assuming regression coefficients to be nonparametric functions of dropout times. Hence estimation of the covariate effects is more robust to misspecification of the dependence between longitudinal responses and dropout.

## 1.2 Motivating example

Protocol 128 of the ACTG was a randomized double-blind equivalency trial of high-dose (180mg per square meter body surface area, six times daily) versus low-dose (90mg) zidovudine (ZDV) for HIV-infected children (Brady et al, 1996). The study enrolled 424 children, randomized them to receive one of the two doses, and followed the children on a number of endpoints for up to five years. In this paper, we are concerned with comparing longitudinal trajectory of CD4 cell counts. Children were scheduled for measurement of CD4 count every 12 weeks, but actual measurement times varied considerably. In addition, only about half of the participants completed three years of follow-up (113/216 [52%] on low dose, 93/208 [45%] on high dose).

A simple but reasonable approach to analyzing these data is to estimate treatment-group-specific CD4 trajectories using a linear random effects model (Laird and Ware, 1982). This model provides valid inference under ignorable dropout (Laird, 1988; Diggle and Kenward, 1994). Basic exploratory analysis suggests that for the observed data, mean square root of CD4 is well described by a linear time trend in both treatment arms (Figure 1). Using the random effects model, estimated change from baseline to week 200 is  $-12.7$  (s.e. 0.8) in the low dose arm,  $-18.2$  (s.e. 1.4) in the high dose arm, for a difference of  $-5.5$  (s.e. 1.6), favoring low dose.

To explore the potential for bias due to outcome-related dropout, we plotted estimated individual least-squares slopes versus dropout time (Figure 2). A clear pattern is evident in both treatment arms, namely that lower slopes are associated with early dropout. Moreover, the bottom panels on Figure 2 indicate that dropout is substantial (completing the protocol is defined as having 156 weeks or more of follow-up). This casts some doubt on the ignorable dropout assumption and hence on the validity of estimates from the linear random effects model, suggesting the need to utilize more elaborate models for addressing potential effects of informative dropout.

The remainder of our paper is organized as follows. The model is described in Section 2, and estimation procedures are detailed in Section 3; this includes estimation for discrete and continuous dropout times, and for settings with censored dropout times. In Section 4, we apply the proposed model to the clinical trial described in above. Section 5 presents a simulation study to evaluate the bias of the proposed method under departures from underlying assumptions. Summary and discussion follow in Section 6.

## 2 Mixtures of varying coefficient models for handling informative dropout

Suppose that the data consist of  $m$  subjects with the  $i$ th subject having  $n_i$  observations over time. For the  $i$ th subject, let  $\mathbf{Y}_i$  be an  $n_i \times 1$  observed outcome vector,  $\mathbf{X}_i$  be an  $n_i \times p$  covariate matrix associated with fixed effects,  $\mathbf{Z}_i$  be an  $n_i \times q$  covariate matrix associated with random effects, and  $U_i$  be the dropout time. The complete data distribution for the response vector is the mixture obtained by integrating the joint distribution  $f(\mathbf{y}, u)$  over  $u$ . Mixture model approaches, therefore, require specification of  $f(\mathbf{y} \mid u)$  and  $f(u)$ . When dropout times are discrete, it is usual to leave  $f(u)$  unspecified and estimate it nonparametrically; for continuous  $u$ , it is possible but not always desirable to use a parametric model such as lognormal (DeGruttola and Tu, 1994; Schluchter, 1992). Our approach is to leave this marginal distribution unspecified.

To capture the dependence between  $\mathbf{Y}$  and  $U$ , we assume that repeated measurements

$\mathbf{Y}_i$  for those who drop out at  $u_i$  follow the varying coefficient random effects model

$$(\mathbf{Y}_i | U = u_i) = \mathbf{X}_i \boldsymbol{\beta}(u_i) + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad (1)$$

where  $\boldsymbol{\beta}(u) = \{\beta_1(u), \dots, \beta_p(u)\}^T$  is a  $p \times 1$  vector of unknown regression coefficient functions of the dropout time  $u$ ,  $\mathbf{b}_i$  is a  $q \times 1$  vector of random effects following  $N\{\mathbf{0}, \mathbf{D}(\boldsymbol{\theta}, u_i)\}$ ,  $\boldsymbol{\epsilon}_i$  is an  $n_i \times 1$  vector of residuals following  $N\{\mathbf{0}, \mathbf{R}_i(\boldsymbol{\theta}, u_i)\}$ , and  $\boldsymbol{\theta}$  is a  $c \times 1$  vector of variance components. Note that we allow the covariance matrices  $\mathbf{D}$  and  $\mathbf{R}$  to depend on the dropout time  $u_i$ . To better understand model (1), we write  $\mathbf{X}_i = (\mathbf{X}_{i1}, \dots, \mathbf{X}_{ip})$ , where  $\mathbf{X}_{ij}$  is an  $n_i \times 1$  vector of the values of the  $j$ th covariate measured over time for the  $i$ th subject. Equation (1) can be written as

$$(\mathbf{Y}_i | U = u_i) = \sum_{j=1}^p \mathbf{X}_{ij} \beta_j(u_i) + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad (2)$$

where  $\beta_j(u)$  represents the  $j$ th covariate effect for those who drop out at time  $u$ .

In settings such as a panel design, where subjects are observed at prespecified finite time points (panels) and the underlying dropout times are discrete, we assume the  $\beta_j(u)$  to be step functions (Little and Wang, 1996; Hogan and Laird, 1997). If instead subjects are observed at different irregular time points and the underlying dropout times are continuous, we assume the  $\beta_j(u)$  to be unspecified smooth functions. It follows that model (1) allows the covariate effects vary with dropout times nonparametrically and estimation of the covariate effects can be made more robust.

The varying coefficient specification (1) includes the pattern mixture model (Little, 1993, 1994), random effects pattern mixture model (Little, 1995; Hogan and Laird, 1997), and conditional linear model (Wu and Bailey, 1989; Schluchter, 1992) as special cases. For example, if  $U$  has a discrete distribution with finite support, then the  $\beta_j(u)$  are step functions and (1) is a pattern-mixture model. For continuous dropout times, consider the case where  $\mathbf{X}_i = (\mathbf{1}, \mathbf{T}_i)$  and  $\mathbf{T}_i = (t_{i1}, \dots, t_{in_i})^T$ , where  $t_{ik}$  is the  $k$ th followup time for the  $i$ th subject. If  $\beta_1(u)$  and  $\beta_2(u)$  are polynomial functions of  $u$ , model (1) reduces to a conditional linear model. If the  $\beta_j(u)$  are constant in  $u$ , then the mixture distribution  $f(\mathbf{y})$  has only one component and (1) reduces to a standard random effects model (Laird and Ware, 1982; Diggle, 1984).

Comparisons to selection models (e.g. Wu and Carroll, 1987; Diggle and Kenward, 1994) can be made by deriving the associated probability of dropout at time  $u$  as a function of  $\mathbf{y}$ . Let  $\mathbf{u}^0 = (u_1^0, \dots, u_{r-1}^0)^T$  denote the set of ordered, unique dropout times, and define an arbitrary time  $u_r^0 > u_{r-1}^0$  for the completers. The selection function can be written in closed form as a logistic regression using baseline-category logits, where completers ( $u = u_r^0$ ) define the baseline category. Let  $h_s(\mathbf{y}_i) = \text{pr}(U = u_s^0 | \mathbf{y}_i) / \text{pr}(U = u_r^0 | \mathbf{y}_i)$ , which characterizes the odds of dropping out at time  $u_s^0$  relative to completing the study, given repeated measures  $\mathbf{y}_i$ . Taking logs,

$$\log h_s(\mathbf{y}_i) = \log(\pi_r / \pi_s) + \log f(\mathbf{y}_i | U = u_s^0) - \log f(\mathbf{y}_i | U = u_r^0), \quad (3)$$

where  $\pi_s = \text{Pr}(U_i = u_s^0)$ , and

$$f(\mathbf{y}_i | U_i = u) = |\mathbf{V}_i(\boldsymbol{\theta}, u)|^{-n_i/2} \exp \left[ -\{\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}(u)\}^T \mathbf{V}_i(\boldsymbol{\theta}, u)^{-1} \{\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}(u)\} / 2 \right],$$

with  $\mathbf{V}_i(\boldsymbol{\theta}, u) = \mathbf{Z}_i \mathbf{D}_i(\boldsymbol{\theta}, u) \mathbf{Z}_i^T + \mathbf{R}_i(\boldsymbol{\theta}, u)$ . In general, the selection model (3) is quadratic in  $\mathbf{y}_i$ . A more familiar version obtains when variance is constant across dropout times, i.e. when  $\mathbf{V}_i(\boldsymbol{\theta}, u) = \mathbf{V}_i(\boldsymbol{\theta})$ . Then (3) simplifies to

$$\log h_s(\mathbf{y}_i) = w \left\{ \pi_s, \pi_r, \boldsymbol{\beta}(u_s^0), \boldsymbol{\beta}(u_r^0), \mathbf{V}_i(\boldsymbol{\theta}) \right\} + \left\{ \boldsymbol{\beta}(u_s^0) - \boldsymbol{\beta}(u_r^0) \right\}^T \mathbf{X}_i^T \mathbf{V}_i(\boldsymbol{\theta})^{-1} \mathbf{y}_i, \quad (4)$$

where  $w(\cdot)$  comprises terms that do not depend on  $\mathbf{y}_i$ . Under this formulation, the log relative probability of dropout at  $u_s^0$  is linear in  $\mathbf{y}_i$ , with coefficients that depend on the difference  $\boldsymbol{\beta}(u_s^0) - \boldsymbol{\beta}(u_r^0)$ , increasing in magnitude as  $\boldsymbol{\beta}(u)$  depends more strongly on  $u$ . For binary  $U$ , (4) is consistent with the selection model restrictions used by Little (1994) and Little and Wang (1996) to identify parameters in a two-component pattern mixture model.

### 3 Estimation Procedures

In this section, we discuss estimation procedures when dropout times  $U_i$  are observed for all subjects. Under this circumstance, the likelihood of  $(\mathbf{Y}_i, U_i)$  for the  $i$ th subject can be partitioned as

$$L_i(\mathbf{Y}_i, U_i; \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\pi}) = L_i(\mathbf{Y}_i | U_i; \boldsymbol{\beta}, \boldsymbol{\theta}) L_i(U_i; \boldsymbol{\pi}).$$

It follows that  $(\boldsymbol{\beta}, \boldsymbol{\theta})$  can be estimated by maximizing the conditional likelihood  $\prod_{i=1}^m L_i(\mathbf{Y}_i \mid U_i; \boldsymbol{\beta}, \boldsymbol{\theta})$ , and  $\boldsymbol{\pi}$  can be estimated by maximizing the marginal likelihood  $\prod_{i=1}^m L_i(U_i; \boldsymbol{\pi})$ . We consider estimation procedures first for the situation where subjects are observed at a common set of time points, and then where the set of observation times may be misaligned across subjects.

### 3.1 Estimation procedures for fixed and common observation times

For a fixed design, such as a panel design, subjects are observed at  $n$  (often small) prespecified time points  $(t_1, \dots, t_n)$  and the number of possible dropout times is small. Thus, the nonparametric functions  $\beta_j(u)$  are assumed to be step functions.

Let  $r$  be the number of observed distinct values of  $u$ , with  $r \leq n$ . As indicated in Section 2, let  $\mathbf{u}^0 = (u_1^0, \dots, u_r^0)^T$  be an  $r \times 1$  vector of ordered distinct dropout times, where we assume completers take  $u_r^0 = t_{n+1}$  for some value  $t_{n+1}$ . Hence  $\beta_j(u)$  is fully determined by  $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{jr})^T$ , where  $\beta_{jk}$  represents the  $j$ th covariate effect for those dropping out at  $u_k^0$ . Denote the incidence matrix by  $\mathbf{N}_i = (N_{i1}, \dots, N_{ir})^T$  where  $N_{ik} = 1$  if  $u_i = u_k^0$  and 0 otherwise. Some calculations show that model (2) can then be written as a linear mixed model

$$\mathbf{Y}_i = \sum_{j=1}^p \mathbf{X}_{ij} \mathbf{N}_i^T \boldsymbol{\beta}_j + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i.$$

Equivalently, we have

$$\mathbf{Y} = \sum_{j=1}^p \widetilde{\mathbf{X}}_j \boldsymbol{\beta}_j + \mathbf{Z} \mathbf{b} + \boldsymbol{\epsilon}, \quad (5)$$

where  $\widetilde{\mathbf{X}}_j = \mathbf{X}_j \otimes \mathbf{N}$ ,  $\mathbf{X}_j = (\mathbf{X}_{1j}^T, \dots, \mathbf{X}_{mj}^T)^T$ ,  $\mathbf{N} = (\mathbf{N}_1^T, \dots, \mathbf{N}_m^T)^T$ , and  $\mathbf{Y}$ ,  $\mathbf{Z}$ ,  $\mathbf{b}$ , and  $\boldsymbol{\epsilon}$  are defined similarly to  $\mathbf{N}$ . Here  $\mathbf{A} \otimes \mathbf{B}$  denotes a direct product of matrices  $\mathbf{A}$  and  $\mathbf{B}$ .

It follows that estimation of  $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_p^T)^T$  proceeds by solving the linear mixed model normal equation

$$(\widetilde{\mathbf{X}}^T \mathbf{V}^{-1} \widetilde{\mathbf{X}}) \boldsymbol{\beta} = \widetilde{\mathbf{X}}^T \mathbf{V}^{-1} \mathbf{Y},$$

where  $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{X}}_1, \dots, \widetilde{\mathbf{X}}_p)$ ,  $\mathbf{V} = \text{diag}(\mathbf{V}_i)$  and  $\mathbf{V}_i = \text{cov}(\mathbf{Y}_i) = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \mathbf{R}_i$ . The covariance matrix of the estimator of  $\boldsymbol{\beta}$  is  $(\widetilde{\mathbf{X}}^T \mathbf{V}^{-1} \widetilde{\mathbf{X}})^{-1}$ . Estimation of  $\boldsymbol{\theta}$  can be obtained using the

REML estimating equation

$$-\frac{1}{2}\text{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta_j}\right) + \frac{1}{2}(\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\beta})^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_j}\mathbf{V}^{-1}(\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}) = 0,$$

where  $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^T\mathbf{V}^{-1}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{V}^{-1}$ . The  $(j, k)$ th component of the information matrix of the estimator of  $\boldsymbol{\theta}$  is  $(1/2)\text{tr}\{\mathbf{P}(\partial\mathbf{V}/\partial\theta_j)\mathbf{P}(\partial\mathbf{V}/\partial\theta_k)\}$ . If  $\mathbf{D}$  and  $\mathbf{R}$  depend on  $u$ , the identifiability of some components of  $\boldsymbol{\theta}$  based on the observed data might require some constraints on  $\boldsymbol{\theta}$  or sensitivity analysis (Little, 1993). For example, consider the case with two time points ( $n = 2$ ); if dropout is informative, there is no information in the observed data to fit a random intercept and slope model for those who drop out at time 2, and sensitivity analysis is needed.

The necessary and sufficient condition for the  $\boldsymbol{\beta}_j$  to be identifiable from model (5) can be stated as follows:

Let  $\mathcal{S}_k = \{i_{k1}, \dots, i_{km_k}\}$  denote the indexes of those  $m_k$  subjects who drop out at  $u_k^0$ , and  $\mathcal{X}_k = (\mathbf{X}_{i_{k1}}^T, \dots, \mathbf{X}_{i_{km_k}}^T)$ . If  $\text{rank}(\mathcal{X}_k) = p$  for all  $k = 1, \dots, r$ , then the  $\boldsymbol{\beta}_j$  ( $j = 1, \dots, p$ ) are identifiable.

This condition states that when dropout times are discrete and finite, regression coefficients must be separately estimable for each dropout pattern. **If some components of  $\boldsymbol{\beta}_j$  are not identifiable for some dropout patterns, then the observed data do not have information about these components and either parameter constraints or sensitivity analysis is needed. For example, consider the case with two time points ( $n = 2$ ), if dropout is informative, there is no information in the observed data to estimate the mean of the outcome at time 2 for those who drop out at time 2, and sensitivity analysis is hence needed. See Little and Wang (1996) and Daniels and Hogan (200?) for detailed discussion.**

Note that the  $j$ th regression coefficient vector  $\boldsymbol{\beta}_j$  measures the  $j$ th covariate effect conditional on dropout times. Primary interest is in the marginal  $j$ th covariate effect  $\tilde{\boldsymbol{\beta}}_j = E_U\{\boldsymbol{\beta}_j(u)\} = \boldsymbol{\pi}^T\boldsymbol{\beta}_j$ , where  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_r)^T$  and  $\pi_j = P(U = u_k^0)$  is the probability that a subject drops out at  $u_k^0$ . It follows that one can estimate  $\tilde{\boldsymbol{\beta}}_j$  by  $\hat{\boldsymbol{\pi}}^T\hat{\boldsymbol{\beta}}_j$ , where  $\hat{\boldsymbol{\pi}}_j = m_j/m$ ,  $m_j$



is the number of subjects who drop out at  $u_k^0$ , and  $\hat{\boldsymbol{\beta}}$  is the maximum likelihood estimator from fitting model (5).

### 3.2 Estimation for random or misaligned measurement times

For situations when the measurement times are random or misaligned across subjects, repeated measures of each subject are observed at different time points and the underlying dropout times are often continuous. We therefore assume that the regression coefficient functions  $\beta_j(u)$  are twice-differentiable smooth functions and estimate them using cubic smoothing splines. A key feature of cubic smoothing spline estimation is that we can estimate all model components — including the nonparametric regression coefficient functions  $\beta_j(u)$ , variance components  $\boldsymbol{\theta}$  and smoothing parameters — within arguedmented parametric linear mixed model framework.

Following the notation in Section 3.1, let  $r$  be the number of ordered distinct values of the  $U_i$ ,  $\mathbf{u}^0 = (u_1^0, \dots, u_r^0)^T$  be the ordered distinct values of the  $U_i$ , and  $\boldsymbol{\beta}_j = \{\beta_j(u_1^0), \dots, \beta_j(u_r^0)\}^T$  be the values of  $\beta_j(u)$  evaluated at  $\mathbf{u}^0$ . Given the variance components  $\boldsymbol{\theta}$ , the conditional loglikelihood of  $\boldsymbol{\beta}_j$  given the  $u_i$  is

$$\begin{aligned} \ell(\mathbf{Y}|\mathbf{u}; \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p) &= \sum_{i=1}^m -\frac{1}{2} \ln |\mathbf{V}_i| - \frac{1}{2} \left\{ \mathbf{Y}_i - \sum_{j=1}^p \mathbf{X}_{ij} \beta_j(u_i) \right\}^T \mathbf{V}_i^{-1} \left\{ \mathbf{Y}_i - \sum_{j=1}^p \mathbf{X}_{ij} \beta_j(u_i) \right\} \\ &= -\frac{1}{2} \ln |\mathbf{V}| - \frac{1}{2} \left( \mathbf{Y} - \sum_{j=1}^p \widetilde{\mathbf{X}}_j \boldsymbol{\beta}_j \right)^T \mathbf{V}^{-1} \left( \mathbf{Y} - \sum_{j=1}^p \widetilde{\mathbf{X}}_j \boldsymbol{\beta}_j \right) \end{aligned}$$

where  $\widetilde{\mathbf{X}}_j$ ,  $\mathbf{V}_i$ ,  $\mathbf{V}$ , and  $\mathbf{Y}$  were defined in Section 3.1.

Following O'Sullivan, Yandell and Raynor (1986), one can show that the natural cubic smoothing spline estimators of the  $\beta_j(u)$  maximize the following penalized conditional loglikelihood

$$\ell(\mathbf{Y}|\mathbf{u}; \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p) - \frac{1}{2} \sum_{j=1}^p \lambda_j \int_{A_1}^{A_2} [\beta_j''(u)]^2 du = \ell(\mathbf{Y}|\mathbf{u}; \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p) - \frac{1}{2} \sum_{j=1}^p \lambda_j \boldsymbol{\beta}_j^T \mathbf{K} \boldsymbol{\beta}_j, \quad (6)$$

where the  $\lambda_j$  are smoothing parameters controlling the balance between the goodness-of-fit and the smoothness of the estimated  $\boldsymbol{\beta}_j(u)$ ,  $A_1$  and  $A_2$  specify the range of  $u$ , and  $\mathbf{K}$  is the nonnegative definite natural cubic smoothing spline smoothing matrix constructed using  $\mathbf{u}_0$  and defined in equation (2.3) of Green and Silverman (1994).

For fixed smoothing parameters  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p)^T$  and the variance components  $\boldsymbol{\theta}$ , differentiation of (6) with respect to  $(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p)$  gives their estimating equations as

$$\begin{bmatrix} \widetilde{\mathbf{X}}_1^T \mathbf{V}^{-1} \widetilde{\mathbf{X}}_1 + \lambda_1 \mathbf{K} & \widetilde{\mathbf{X}}_1^T \mathbf{V}^{-1} \widetilde{\mathbf{X}}_2 & \cdots & \widetilde{\mathbf{X}}_1^T \mathbf{V}^{-1} \widetilde{\mathbf{X}}_p \\ \vdots & \vdots & \vdots & \vdots \\ \widetilde{\mathbf{X}}_p^T \mathbf{V}^{-1} \widetilde{\mathbf{X}}_1 & \widetilde{\mathbf{X}}_p^T \mathbf{V}^{-1} \widetilde{\mathbf{X}}_2 & \cdots & \widetilde{\mathbf{X}}_p^T \mathbf{V}^{-1} \widetilde{\mathbf{X}}_p + \lambda_p \mathbf{K} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_p \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{X}}_1^T \mathbf{V}^{-1} \mathbf{Y} \\ \vdots \\ \widetilde{\mathbf{X}}_p^T \mathbf{V}^{-1} \mathbf{Y} \end{bmatrix}. \quad (7)$$

One can solve equation (7) using a backfitting algorithm as follows

$$\widehat{\boldsymbol{\beta}}_j = (\widetilde{\mathbf{X}}_j^T \mathbf{V}^{-1} \widetilde{\mathbf{X}}_j + \lambda_j \mathbf{K})^{-1} \widetilde{\mathbf{X}}_j^T \mathbf{V}^{-1} (\mathbf{Y} - \sum_{k \neq j} \widetilde{\mathbf{X}}_k \boldsymbol{\beta}_k)$$

for  $j = 1, \dots, p$ .

Following Zhang et al (1998), we show that the above cubic smoothing spline estimators of the  $\boldsymbol{\beta}_j$  can in fact be easily obtained by fitting an argumented linear mixed model. Specifically,  $\boldsymbol{\beta}_j$  can be written via a one-to-one transformation as

$$\boldsymbol{\beta}_j = \mathbf{U} \boldsymbol{\gamma}_j + \mathbf{B} \mathbf{a}_j,$$

where  $\mathbf{U} = (\mathbf{1}, \mathbf{u}^0)$ ,  $\mathbf{B} = \mathbf{L}(\mathbf{L}^T \mathbf{L})^{-1}$ ,  $\mathbf{L}$  is an  $r \times (r-2)$  full rank matrix satisfying  $\mathbf{K} = \mathbf{L} \mathbf{L}^T$  and  $\mathbf{L}^T \mathbf{U} = \mathbf{0}$ ,  $\boldsymbol{\gamma}_j$  is a  $2 \times 1$  unknown vector and  $\mathbf{a}_j$  is a  $(r-2) \times 1$  unknown vector. It can be easily shown that  $\boldsymbol{\beta}_j^T \mathbf{K} \boldsymbol{\beta}_j = \mathbf{a}_j^T \mathbf{a}_j$ . The penalized loglikelihood (6) becomes

$$\ell(\mathbf{Y} | \mathbf{U}; \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_p) - \frac{1}{2} \sum_{j=1}^p \lambda_j \mathbf{a}_j^T \mathbf{a}_j.$$

It follows that the nonparametric natural cubic spline estimators of the  $\boldsymbol{\beta}_j$  can be obtained by fitting the parametric linear mixed model

$$\mathbf{Y} = \sum_{j=1}^p (\widetilde{\mathbf{X}}_j \mathbf{U}) \boldsymbol{\gamma}_j + \sum_{j=1}^p (\widetilde{\mathbf{X}}_j \mathbf{B}) \mathbf{a}_j + \mathbf{Z} \mathbf{b} + \boldsymbol{\epsilon}, \quad (8)$$

where  $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_p^T)^T$  is a vector of regression coefficients,  $\mathbf{a} = (\mathbf{a}_1^T, \dots, \mathbf{a}_p^T)^T$  and  $\mathbf{b}$  are independent random effects with  $\mathbf{a} \sim N(\mathbf{0}, \boldsymbol{\Lambda}(\boldsymbol{\tau}))$ ,  $\mathbf{b} \sim N(\mathbf{0}, \text{diag}\{\mathbf{D}(\boldsymbol{\theta})\})$ ,  $\boldsymbol{\Lambda}(\boldsymbol{\tau}) = \text{diag}(\tau_j \mathbf{I})$ ,  $\tau_j = 1/\lambda_j$  and  $\boldsymbol{\tau} = (\tau_1, \dots, \tau_p)^T$ , and  $\boldsymbol{\epsilon}_i \sim N\{\mathbf{0}, \mathbf{R}_i(\boldsymbol{\theta}, u_i)\}$ .

Estimation of  $\boldsymbol{\gamma}$  and  $\mathbf{a}$  can proceed using the BLUP estimator by solving the normal equation

$$\begin{bmatrix} \mathbf{H}^T \mathbf{V}^{-1} \mathbf{H} & \mathbf{H}^T \mathbf{V}^{-1} \mathbf{G} \\ \mathbf{G}^T \mathbf{V}^{-1} \mathbf{H} & \mathbf{G}^T \mathbf{V}^{-1} \mathbf{G} + \boldsymbol{\Lambda}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{H}^T \mathbf{V}^{-1} \mathbf{Y} \\ \mathbf{G}^T \mathbf{V}^{-1} \mathbf{Y} \end{bmatrix}, \quad (9)$$

where  $\mathbf{H} = (\widetilde{\mathbf{X}}_1\mathbf{U}, \dots, \widetilde{\mathbf{X}}_p\mathbf{U})$  and  $\mathbf{G} = (\widetilde{\mathbf{X}}_1\mathbf{B}, \dots, \widetilde{\mathbf{X}}_p\mathbf{B})$ . Denoting by  $\widehat{\boldsymbol{\gamma}}$  and  $\widehat{\mathbf{a}}$  the solution of (9), the natural cubic smoothing spline estimator of  $\boldsymbol{\beta}_j$  is  $\widehat{\boldsymbol{\beta}}_j = \mathbf{U}\widehat{\boldsymbol{\gamma}}_j + \mathbf{B}\widehat{\mathbf{a}}_j$ . One can easily show that the  $\widehat{\boldsymbol{\beta}}_j$  from (9) are identical to those obtained from solving (7). The natural cubic spline estimators  $\widehat{\boldsymbol{\beta}}_j$  are unique when  $\mathbf{H}$  is of full rank.

We have so far assumed that the smoothing parameters  $\boldsymbol{\lambda}$  and the variance components  $\boldsymbol{\theta}$  are known when estimating the  $\boldsymbol{\beta}_j$ . They are usually unknown in practice and need to be estimated from the data. Examination of the modified linear mixed model (8) suggests that  $\boldsymbol{\tau}$  behaves like variance components; therefore, following Zhang, et al (1998), we estimate the smoothing parameters  $\boldsymbol{\tau}$  and the variance components  $\boldsymbol{\theta}$  simultaneously using restricted maximum likelihood (REML) by treating  $\boldsymbol{\tau}$  as extra variance components in addition to  $\boldsymbol{\theta}$  in (8). Some calculations show that the REML estimating equations for  $\boldsymbol{\theta}$  and  $\boldsymbol{\tau}$  are

$$\begin{aligned} -\frac{1}{2}\text{tr}\left(\widetilde{\mathbf{P}}\frac{\partial\mathbf{V}}{\partial\theta_k}\right) + \frac{1}{2}(\mathbf{Y} - \sum_{j=1}^p \mathbf{X}_j\boldsymbol{\beta}_j)^T \mathbf{V}^{-1} \frac{\partial\mathbf{V}}{\partial\theta_k} \mathbf{V}^{-1} (\mathbf{Y} - \sum_{j=1}^p \mathbf{X}_j\boldsymbol{\beta}_j) &= 0 \\ -\frac{1}{2}\text{tr}\left(\widetilde{\mathbf{P}}\widetilde{\mathbf{X}}_j\mathbf{B}\mathbf{B}^T\widetilde{\mathbf{X}}_j^T\right) + \frac{1}{2}(\mathbf{Y} - \sum_{j=1}^p \mathbf{X}_j\boldsymbol{\beta}_j)^T \mathbf{V}^{-1} \widetilde{\mathbf{X}}_j\mathbf{B}\mathbf{B}^T\widetilde{\mathbf{X}}_j^T \mathbf{V}^{-1} (\mathbf{Y} - \sum_{j=1}^p \mathbf{X}_j\boldsymbol{\beta}_j) &= 0, \end{aligned}$$

where  $\widetilde{\mathbf{P}} = \mathbf{V}^{-1} - \mathbf{V}^{-1}(\mathbf{H}, \mathbf{G})\mathbf{C}^{-1}(\mathbf{H}, \mathbf{G})^T\mathbf{V}^{-1}$ , and  $\mathbf{C}$  is the coefficient matrix on the left hand side of (9). Parameters from the varying coefficient mixture model (1) can therefore be obtained by fitting the parametric linear mixed model (8) by SAS Proc Mixed. The smoothing matrix  $\mathbf{B}$  needs to be computed in advance.

### 3.3 Inference for marginal regression coefficients

The marginal  $j$ th covariate effect  $\widetilde{\beta}_j$  is  $\widetilde{\beta}_j = E_U\{\beta_j(u)\} = \int \beta_j(u)dF(u)$ , where  $F(u)$  is the c.d.f. of  $u$ . Using the estimated cubic smoothing spline  $\widehat{\beta}_j(u)$  and the empirical c.d.f.  $\widehat{F}(u)$ , one can estimate  $\widetilde{\beta}_j$  by  $\int \widehat{\beta}_j(u)d\widehat{F}(u) = \widehat{\boldsymbol{\pi}}^T\widehat{\boldsymbol{\beta}}_j$ , where  $\widehat{\boldsymbol{\pi}} = \sum_{i=1}^m \mathbf{N}_i/m$ . The delta method has been used for standard error estimation for mixture models where the support of  $U$  is very small relative to the number of subjects (Hogan and Laird, 1997; Fitzmaurice and Laird, 2000). In our application with continuous dropout times, we found that the delta method performed poorly; the bootstrap was used instead, treating the subject as the basic resampling unit. Details are provided in the application.

## 4 Application to AIDS clinical trial

In this section we analyze the ACTG data using several different mixture model formulations representing different assumptions about the missing data mechanism, provide detailed interpretation of the model parameters, and provide some guidelines for model selection.

### 4.1 Variable transformations and candidate models

As indicated in Section 1.2, observed CD4 data were transformed to the square root scale to reduce positive skewness. We fit three models (for computational tractability, the models were fit separately by dose). The first is a standard random effects model (REM) with subject-specific intercepts and linear time trends. The REM, briefly summarized in Section 1.2, assumes that the complete data in each treatment arm follow the linear mixed model

$$Y_{il} = \beta_1 + \beta_2 t_{il}^* + b_{1i} + b_{2i} t_{il}^* + \epsilon_{il}, \quad (10)$$

where  $Y_{il}$  is square root of CD4 count at time  $t_{il}$ ,  $l = 1, \dots, n_i$ ,  $\mathbf{b}_i = (b_{1i}, b_{2i}) \sim N(\mathbf{0}, \mathbf{D})$  and  $\epsilon_{il} \sim N(0, \sigma^2)$  (i.e.  $\mathbf{R}_i(\boldsymbol{\theta}) = \sigma^2 \mathbf{I}_{n_i}$ ), with independence between  $\epsilon_{il}$  and  $\mathbf{b}_i$ . The time axis is rescaled using  $t_{il}^* = (t_{il} - \bar{U}) / \text{range}(t_{il})$  so that the new time scale has range 1 and is centered at the sample mean of dropout times (the range is computed for the pooled sample from both dose arms). Model 0 is a standard random effects model, but can be viewed as a special case of (1) where  $\beta_j(u)$  is constant in  $u$  for  $j = 1, 2$ . **I think the following sentences will cause confusions and trouble and make the analysis not clear. I think it is better to cut it. I don't understand why this only affects the SEs. the interpretation of  $\beta_0$  is the mean of  $Y$  at  $\bar{U}$ , which might not be very attractive. I am not suggesting to redo the analysis, since I understand bootstrapping is very time-consuming. We probably just leave the way it is for now. After the paper is submitted, do you think whether it is worthwhile if we recenter  $t$  by  $\bar{t}$  and refit the models?**

*Yes I think we should re-fit. The point of the text - which I deleted - is that  $\beta_0$  is also the mean of  $Y$  unconditionally on  $U$  (take expectation over  $U$ ). I think it affects only SE because the SE of the conditional mean (estimated from proc mixed) will be different than the SE of the marginal mean (which must also include variability in estimating  $E(U)$  with  $\bar{U}$ .)*

In addition to the REM, we fit a conditional linear model (CLM) in which individual intercepts and slopes are linear functions of dropout time, and a varying coefficient model (VCM), where intercepts and slopes are unspecified smooth functions of dropout time. The CLM is precisely model (8) under the assumption that  $\boldsymbol{\tau} = \mathbf{0}$  (equivalently,  $\mathbf{a}_j = \mathbf{0}$  for all  $j$ ), and therefore it is just a specialized version of the VCM. Separately by treatment, the CLM elaborates (10) such that  $\beta_1(u) = \gamma_1 + \gamma_2 u^*$  and  $\beta_2(u) = \gamma_3 + \gamma_4 u^*$ , where  $u^* = (u - \bar{U})/\text{range}(t_{il})$ . In our parameterization with the rescaled time axis,  $\gamma_1$  is the mean of  $\sqrt{\text{CD4}}$  at  $u = \bar{U}$ ,  $\gamma_2$  is the ‘main effect’ of dropout, which is the slope of  $\beta_1(u)$ ,  $\gamma_3$  is mean change in  $\sqrt{\text{CD4}}$  from baseline for those with  $U_i = \bar{U}$ , and  $\gamma_4$  is the effect of interaction between dropout and change from baseline, representing the slope of  $\beta_2(u)$  on  $u$ . These parameters are estimated in straightforward fashion to fitting by SAS PROC MIXED a standard linear mixed model  $Y_{ij} = \mathbf{H}_{ij}^T \boldsymbol{\gamma} + \mathbf{Z}_{ij}^T \mathbf{b}_i + \epsilon_{ij}$ , where  $\mathbf{H}_{ij} = (1, U_{ij}^*, t_{ij}^*, t_{ij}^* U_{ij}^*)^T$ , and  $\mathbf{Z}_{ij} = (1, t_{ij}^*)^T$ .

Rescaling the time axis has some practical advantages in terms of model fitting and interpretation: (i) it makes the estimates more stable by increasing the variance of individual slopes away from zero; (ii) for the conditional linear model, the intercept and slope main effect are already averaged over dropout time (cf. Fitzmaurice Stat Med); (iii) for the REM and CLM, the slope parameter corresponds to average total change in square root CD4 from baseline to the longest follow up time for the study; (iv) for the CLM, the parameters  $\gamma_2$  and  $\gamma_4$  contrast mean intercept and slope, respectively, between those who drop out immediately (directly after  $U = 0$ ) and those who complete the study protocol ( $U = \max_{il} t_{il}$ ).

The VCM uses the same design matrix for  $\boldsymbol{\gamma}$  as the CLM. Estimation of the  $\mathbf{a}_j$  is made by specifying the  $r - 2$  columns of  $\widetilde{\mathbf{X}}_1 \mathbf{B}$  and  $\widetilde{\mathbf{X}}_2 \mathbf{B}$  as independent random effects with respective variances  $\tau_1^2$  and  $\tau_2^2$ .

## 4.2 Summary of fitted models

### 4.2.1 Parameter estimates

To get a crude understanding of whether MAR may be a valid assumption, we begin by summarizing regression coefficients for parameterizations of the conditional distribution  $f(\mathbf{y} | u)$  given by the REM and CLM; these appear in Table 1. For both models, the parameter

$\gamma_3$  quantifies average change from baseline to the maximum follow up time. Recall that in the CLM,  $\beta_0(u) = \gamma_1 + \gamma_2(u^* - \bar{u}^*)$  and  $\beta_1(u) = \gamma_3 + \gamma_4(u^* - \bar{u}^*)$ , so that if we assume MAR is violated according to linear dependence between intercept and dropout and/or slope and dropout, the parameters  $\gamma_2$  and  $\gamma_4$  will quantify the degree to which MAR fails to hold. The parameter estimates from Model 1 suggest that intercepts vary considerably by dropout time ( $\hat{\gamma}_2 = 14.6$  and  $13.0$  [s.e. 3.3 and 3.5] for low dose and high dose, respectively); the same pattern is indicated for slopes, where  $\hat{\gamma}_4 = 26.1$  and  $33.7$  (s.e. 4.8 and 6.5) for low and high dose. These parameters are interpreted as the average difference in intercept ( $\gamma_2$ ) and slope ( $\gamma_4$ ) between those who dropped out immediately after enrolling and those who were followed for the maximum time (220 weeks). Hence the CLM indicates that dropouts have lower intercepts and slopes than completers, leading to selection bias for end-of-study comparisons due to missing data on less healthy participants. Because dropout time is centered at its mean in the CLM,  $\hat{\gamma}_3$  is the expected change in square root CD4 from week 0 to 220 if a subject were followed for the whole study period (note however that the standard error is from the conditional distribution of  $\mathbf{Y}$  given  $U$ ). A comparison of the estimates of  $\boldsymbol{\gamma}$  under REM and CLM suggests that the effect of accounting for dropout via CLM is to correct the average change in  $\sqrt{\text{CD4}}$  downward.

In the VCM we allow both  $\beta_1(u)$  and  $\beta_2(u)$  to be completely unspecified for both treatment arms, and estimate them using cubic smoothing splines. Plots of the estimated functionals from the VCM, together with empirical Bayes estimates of individual intercepts and slopes, appear in Figure 3. In both treatment arms, baseline mean and CD4 slope are highly associated with dropout time: those who remain in the study for longer periods have higher values of both. Except for the baseline mean on the low dose arm, the associations appear to be highly nonlinear.

#### 4.2.2 Model selection and treatment comparison

Xihong, I revised this somewhat and agree that we should drop the table. You may feel like you want to drop all references to REML and AIC. I leave this up to you. I bolded places where I made changes.

Because the complete data likelihood factors over  $(\boldsymbol{\beta}, \boldsymbol{\theta})$  and  $\boldsymbol{\pi}$ , and because our three

models differ only in the specification of the conditional factor  $f(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\beta}, \boldsymbol{\theta})$ , model selection can **in principle** be based on criteria for the conditional  $(\mathbf{y} \mid \mathbf{u})$  model. However, formal model selection procedures comparing parametric and nonparametric mixed effects models are not well developed currently and are beyond the scope of this paper; furthermore, there are potential difficulties associated with comparing likelihoods for models that are not properly nested in a traditional way, e.g., due to boundary-value problems. **As a very rough guide, we compared Akaike’s Information Criterion (AIC) — the loglikelihood minus a term that penalizes linearly in the number of parameters — for conditional specifications across the three models. The VCM gives the most favorable value of AIC, suggesting that the benefit gained in terms of fit justifies increased complexity.**

Table 2 lists estimates and associated standard errors for the marginal regression coefficients  $\tilde{\beta}_j = \int \beta_j(u) dF(u)$ , for  $j = 1, 2$ , estimated according to the procedure described in Section 3.3 (note that the marginal coefficients — and not the conditional coefficients reported in Table 1 — are of direct scientific interest). Standard errors were calculated using bootstrap resampling based on 100 replicated datasets sampled with replacement. Not surprisingly, standard errors associated with the VCM are increased relative to the CLM, reflecting uncertainty about the functional form of  $\beta_j(u)$ , but the increase is relatively modest. Comparing slopes in the low-dose arm, for example,  $\text{s.e.}(\hat{\beta}_2) = 1.0$  for the CLM and **X.X** for the VCM. No loss of efficiency is seen in the estimated intercepts.

Table 2 also indicates the degree to which adjusting for selection bias affects the final inferences about treatment. Under the MAR assumption (REM), estimated mean difference in total change in  $\sqrt{\text{CD4}}$  is  $-5.5$ , with  $Z$  – statistic =  $-3.4$ ; adjustment under the CLM gives the same estimated effect, with  $Z$  – statistic =  $-2.8$ . Both lead to the conclusion that low dose is superior to high dose because the decline in CD4 is less steep. Under the VCM, the correction for possible selection biases on low dose changes the slope estimate from  $-12.7$  (REM) to  $-17.1$ ; the correction is less severe on high dose ( $-18.2$  for REM, compared to  $-20.1$  for VCM); the effect is to narrow the gap in treatment effect to  $-3.0$ , with  $Z = -X.X$ , representing a change of **XXX** standard errors relative to the REM, and **XXX** standard errors relative to the CLM.

The VCM also provides an important substantive insight, namely that participants who drop out of low dose (the experimental dose in this trial) tend to have steeper decline in their CD4 counts, compared to those on high dose. The trial was designed to see whether the lower dose, known to be associated with fewer side effects in adults, would have efficacy equal to the high dose. The form of the  $\beta_2(u)$  functions suggests that among the early dropouts, rate of change in CD4 for those on low dose is substantially less than for those on high dose. In an MAR analysis, early dropouts contribute less information to the estimate of population slope because they have fewer observed data points, leading to the selection bias (and potentially erroneous conclusion) seen in the REM.

## 5 Simulation study

Our model gives the analyst considerable flexibility in specifying dependence between outcome and dropout in the context of a mixture model, and avoids biases that are possible if the functional form of this dependence is assumed to be known. The primary innovation of the VCM over CLM is that  $\beta(u)$  can be left unspecified, but this generalization relies on the key assumption that  $\beta(u)$  is a (vector of) smooth, twice differentiable functions of  $u$ . We designed a brief simulation study to investigate the performance of our model under violations of this assumption.

Each simulation uses datasets with  $n = 50$  subjects having up to 15 unique dropout times, with  $\beta(u)$  taking three different forms. We compare estimates of mean change from baseline from a standard random effects model, the conditional linear model with components of  $\beta(u)$  assumed linear in  $u$ , and from the varying coefficient model with  $\beta(u)$  left unspecified. Specifically, we assume

$$y_{il} = \beta_1(u_i) + \beta_2(u_i)t_{il} + b_{1i} + b_{2i}t_{il} + e_{il},$$

where  $(b_{1i}, b_{2i})^T \sim N(\mathbf{0}, \mathbf{D})$ ,  $e_{il} \sim N(0, \sigma^2)$ . There are 15 time points  $\{t_{il}\}$ , equally spaced between 0 and 1. This simulation uses  $d_{11} = 4$ ,  $d_{22} = 0.1$ ,  $d_{12} = -0.1$  (correlation  $\approx -0.15$ ), and  $\sigma^2 = 1$ , which implies that between-subject variation exceeds within-subject variation by a factor of about 4.



Dropout is generated from a beta mixture of binomial distributions as follows:  $p \sim \text{Beta}(1.5, 1.5)$  (mean 0.5);  $U^* \sim \text{Bin}(15, p)$ , and dropout time  $U = U^*/15 \in (0, 1)$ . Finally, we assume  $\beta_1(u) = 0$  and vary the functional form of  $\beta_2(u)$ ; candidate functions are:

(i)  $-\exp(\alpha u)$ ,

(ii)  $\exp(\alpha u)I(u < t^*) + \exp(\alpha t^*)I(u \geq t^*)$

(exponential with plateau effect for dropouts beyond  $t^*$ ),

(iii)  $\alpha_1 I(u < t^*) + \alpha_2 I(u \geq t^*)$

(two-piece step function).

Case (i) actually meets the assumptions for the VCM, and is included for validating our simulation and estimation routines; case (ii) violates the smoothness assumption and case (iii) violates both smoothness and continuity assumptions.

For (i), at  $\alpha = -4$ , completers ( $U = 1$ ) have mean change from baseline  $\beta_2(1) = \exp(-4) \approx 0.02$  and early dropouts ( $U = 0$ ) have mean change  $-1$ , a difference of about 3 SD (because  $d_{22} = 0.1$ ). Under (ii), we keep  $\alpha = -4$  and invoke the plateau effect at  $t^* = 2/3$ , leading to a structure wherein those who complete 2/3 of the study or more have average change from baseline equal to  $\exp(-4 \times 2/3) \approx 0.07$ . For (iii), we keep  $t^* = 2/3$  and set  $\alpha_1 = 0$ ,  $\alpha_2 = 1$ .

Results are reported in Table 3. As expected, the VCM gives virtually unbiased estimation of the true slope for case (i), where  $\beta_2(u)$  is both continuous and smooth, while both the REM and CLM show substantial upward bias. This comparison is not as trivial as it would appear, however, because exploratory plots of OLS slopes versus dropout time (e.g. Figure 2) do not always reveal an obvious functional form for  $\beta(u)$ , particularly in the early part of the time axis. One advantage to the VCM is its effectiveness in finding a signal from noisy data.

The VCM shows only very little bias for estimating the true slope for the continuous but not everywhere-differentiable function from case (ii), but exhibits more bias than the CLM for the discontinuous function in case (iii). In all cases, however, the VCM outperforms the random effects model.

## 6 Discussion

We have proposed a mixture-modeling approach to analyzing longitudinal data with outcome-dependent dropout. Our model assumes that covariate effects depend on dropout time through unspecified functions  $\{\beta_j(u)\}$ , where  $u$  is dropout time. When dropout times are discrete, the  $\beta_j(u)$  are step functions, and when dropout is continuous,  $\beta_j(u)$  are assumed to be unspecified smooth functions of  $u$ . This formulation generalizes pattern-mixture models (Little, 1993, 1994) and random-effects mixture models (Hogan and Laird, 1997a, Wu and Bailey, 1988, 1989, Follman and Albert, 2000) for continuous response data. Using an example from an AIDS clinical trial, we show that the model has the potential to adjust for potential selection biases induced by poor responders dropping out early; moreover, in our example at least, the increase in standard errors associated with leaving  $\beta(u)$  unspecified is less than expected.

The primary innovation in our approach is that the functional dependence between covariate effects can be left unspecified. Our simulation study shows that when the parametric form of the dependence of the regression coefficients on the dropout time is misspecified, conditional linear models yield biased estimates, while varying-coefficient mixture models still yield unbiased estimates and our method is more robust. In many applications, this is a decided advantage over the Wu and Bailey model because the form of  $\beta(u)$  will rarely be known or intuitive. Moreover, it is our experience that using polynomials leads to overfitting and/or extrapolations well outside the range of data, particularly when the polynomial has degree greater than 2. When  $u$  is continuous, our simulation also shows that inferences are unlikely to be sensitive to lack of smoothness in  $\beta(u)$ , but could be affected by discontinuities. In both cases, however, bias is substantially less than under an MAR analysis.

Another advantage to mixture modeling in general is that extrapolations of the missing data are completely transparent, and lend themselves well both to substantive critique and to empirical sensitivity analysis (e.g. Rubin, 1977; Little and Wang, 1996; Daniels and Hogan, 2000). In our application, for example, we assume dropouts at time  $u$  have the same slope for  $t > u$  as for  $t \leq u$ ; on the surface this is a strong assumption but it is relatively easily modified and is a sensible starting point for sensitivity analysis. For example, one could

perform a sensitivity analysis by varying the slope after the dropout time. It should be noted that the observed data do not contain any information about noninformative dropout mechanism. One either needs to make an assumption about what happens after the time of dropout or perform a sensitivity analysis. Because mixture models are specified in terms of the conditional distribution of  $\mathbf{Y}$  given  $U$ , the extrapolation  $E(\mathbf{Y}_{\text{mis}} \mid \mathbf{Y}_{\text{obs}}, U = u)$  is obvious (and in the mixture-of-normals model is linear in  $\mathbf{Y}_{\text{obs}}$ ); by contrast, the same extrapolation for selection models is not obvious in most cases (but see Rotnitzky et al, 2000 for a discussion of this issue for semiparametric selection models). Further work will consider censoring in the dropout distribution, sensitivity analyses, and generalization to discrete distributions.

## References

Breslow and Clayton, 1993

DeGruttola and Tu, 1994

Diggle, 1988

Diggle and Kenward, 1994

Fitzmaurice, Stat Med

Follman and Wu, 1995

Hogan and Laird, 1997a

Hogan and Laird, 1997b

Kenward and Molenberghs, 1999 SMMR

Laird, 1988

Laird and Ware, 1982.

Liang and Zeger, 1986 Biometrika

Little, 1993 JASA

Little, 1994 Biometrika

Little, 1995 JASA

Robins, Rotnitzky and Zhao 1995 JASA

Rotnitzky, 2000 (sensitivity analysis)

Rotnitzky, Robins and Scharfstein, 1998 JASA

Rubin, 1977 (sensitivity analysis - formalizing subjective notions.)

Scharfstein, Robins and Rotnitzky, 1999 JASA

Schluchter, 1992

Ten Have et al., 1998

Wu and Bailey, 1988

Wu and Bailey, 1989

Wu and Carroll, 1988

Zeger and Liang, 1986 Biometrics

Table 1: Parameter estimates from conditional part of joint model, assuming linear random effects structure (REM) and conditional linear model (CLM).

Model	Parameter*	AZT Dose	
		Low (90mg)	High (180mg)
REM	$\gamma_1$	28.6 (0.8)	30.1 (0.9)
	$\gamma_3$	-12.7 (0.8)	-18.2 (1.4)
CLM	$\gamma_1$	28.9 (0.8)	30.3 (0.8)
	$\gamma_2$	14.6 (3.3)	13.0 (3.5)
	$\gamma_3$	-15.9 (0.9)	-21.4 (1.4)
	$\gamma_4$	26.1 (4.8)	33.7 (6.5)

\* See Section 4 for definitions.

Table 2: Estimated intercept and slope characterizing marginal mean of CD4 trajectory under three different specifications for conditional part of joint model, with standard errors estimated via bootstrap.

Model	Parameter	Low Dose	High Dose	Difference (s.e.)	Z
REM	$\beta_1$	28.6 (0.8)	30.1 (0.9)		
	$\beta_2$	-12.7 (0.8)	-18.2 (1.4)	-5.5 (1.6)	3.4
CLM	$\beta_1$	28.9 (0.8)	30.3 (0.9)		
	$\beta_2$	-15.9 (1.0)	-21.4 (1.8)	-5.5 (2.0)	2.8
VCM	$\beta_1$	29.0 (0.7)	29.9 (—)		
	$\beta_2$	-17.1 (1.3)	-20.1 (—)	-3.0 (—)	

Table 3: Results from simulation to characterize bias. REM = linear random effects model; CLM = conditional linear model with  $\beta_2(u)$  linear; VCM = varying coefficient model with  $\beta_2(u)$  unspecified. Each estimate represents a sample average of estimated slopes over 100 replicated datasets, each having 100 subjects with up to 15 repeated measures. Standard errors for simulation-based estimated mean appear in parentheses.

Underlying model*	True Slope ( $\beta_2$ )	Estimated Slope		
		REM	CLM	VCM
(i) Continuous, smooth	-0.159**	-0.073 (0.018)	-0.119 (0.035)	-0.160 (0.028)
(ii) Continuous, not smooth	-0.170**	-0.062 (0.015)	-0.100 (0.027)	-0.166 (0.033)
(iii) Discontinuous	-0.587**	-0.211 (0.018)	-0.622 (0.031)	-0.715 (0.038)

\* See Section 5 for model descriptions

\*\* Computed to nearest 0.001 via Monte-Carlo simulation

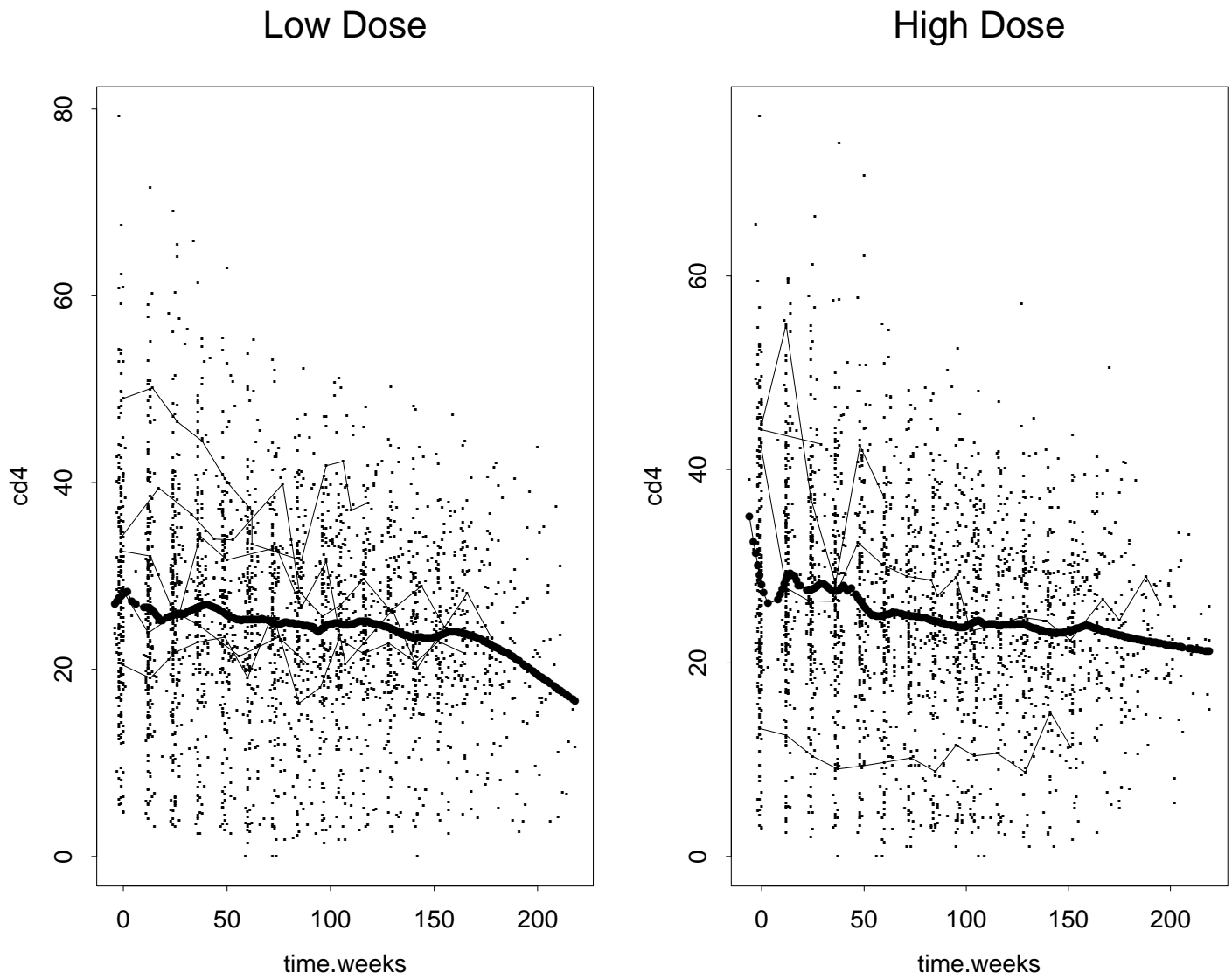


Figure 1: Observed (square root) CD4 counts versus time and stratified by dose, with lowess regression line fit to pooled sample and five individual profiles highlighted.



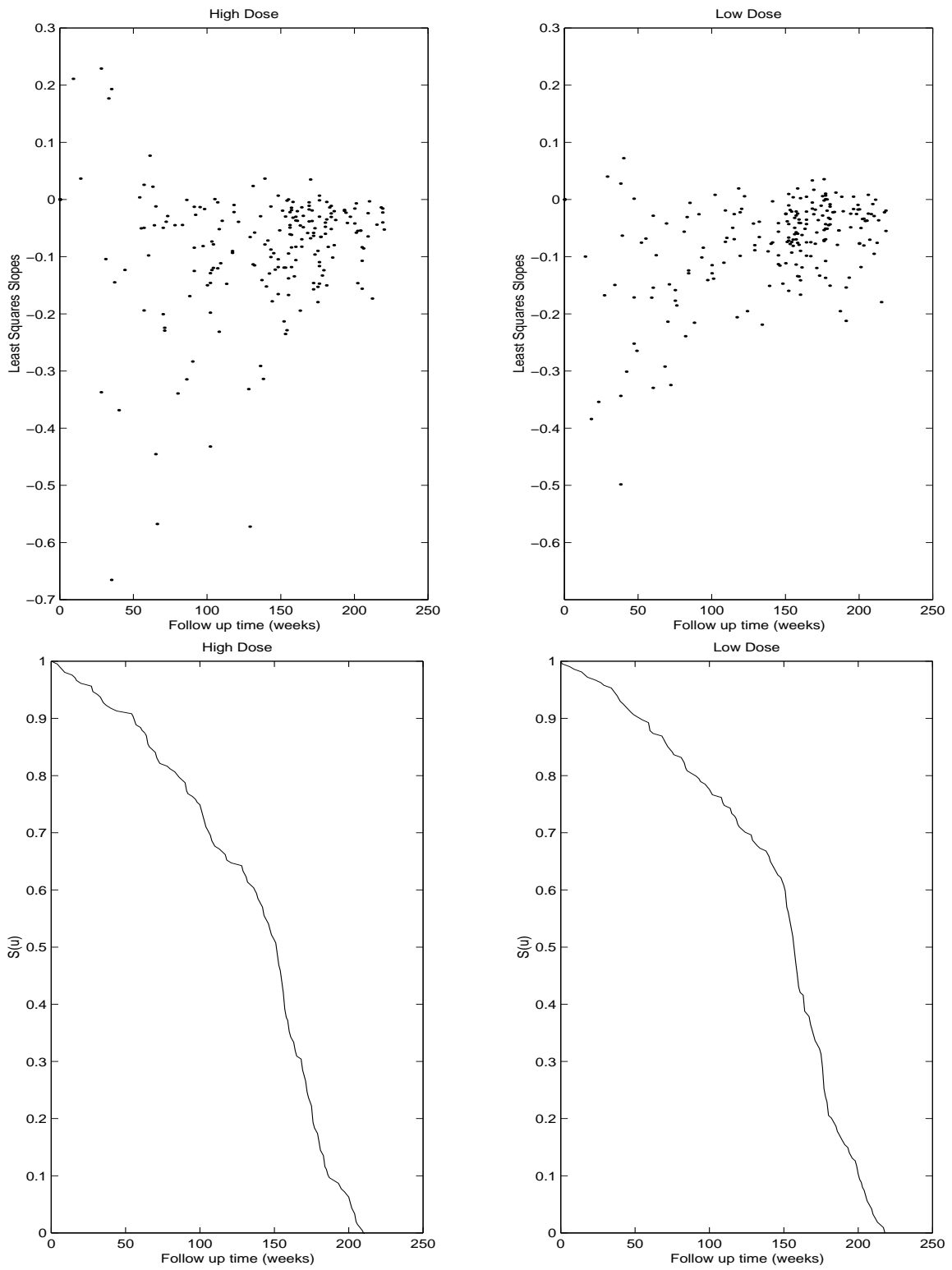


Figure 2: Individual-specific OLS slopes for square-root CD4 versus as a function of follow-up time, and proportion remaining in follow up, stratified by dose.

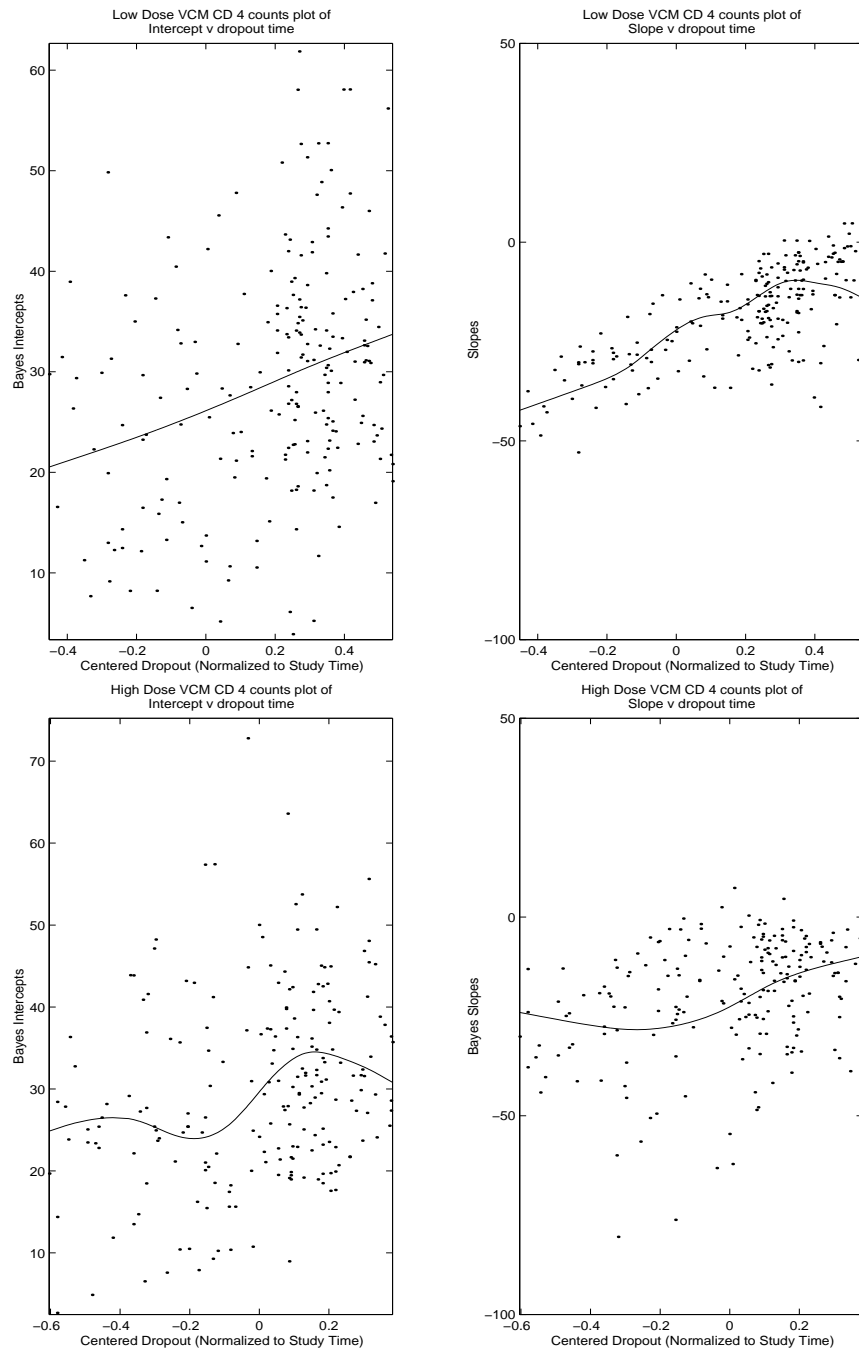


Figure 3: Estimated functions  $\beta_1(u)$  and  $\beta_2(u)$  for low- and high-dose ZDV arms, together with empirical Bayes estimates of individual intercepts and slopes (on square root CD4 scale). Slopes correspond to change from baseline to week 200.