UCLA

Department of Statistics Papers

Title

Longitudinal Models for AIDS Marker Data

Permalink

<https://escholarship.org/uc/item/6t617999>

Authors

W. John Boscardin
Jeremy M.G. Taylor
Ngayee Law

Publication Date

2011-10-25

Longitudinal Models for AIDS Marker Data

W. John Boscardin *

Jeremy M. G. Taylor

Ngayee Law

Department of Biostatistics

University of California, Los Angeles, USA

September 4, 1997

Abstract

Over the past decade, researchers have put a great amount of effort into developing suitable models for the analysis of longitudinal CD4 data and other markers of AIDS progression. These models must be general enough to allow for different patterns of change in the marker data. In this paper, we review the existing literature including our preferred models which involve mixed effects, stochastic terms and independent measurement error. Adding stochastic terms to standard mixed effects models gives an interpretable and parsimonious method for generalizing the covariance structure of the measurement error and short-term variability. We focus on univariate and bivariate models with Integrated Ornstein-Uhlenbeck (IOU) stochastic terms. The IOU

*Address for correspondence: Jeremy M. G. Taylor, Department of Biostatistics, UCLA, Los Angeles, California 90095-1772, USA.

process allows for a range of biologically plausible derivative tracking that encompasses both random trajectory and Brownian motion behavior. We illustrate these modeling techniques on longitudinal CD4 and viral RNA data.

1 Introduction

There is considerable literature on finding suitable models for clinical markers of AIDS. The most widely used marker is the CD4 lymphocyte count, but more recently measures of HIV viral load (especially measures of viral RNA or DNA) have become important.

The available data on such markers are usually characterized by unequal numbers of unevenly spaced observations for each individual. The data are typically both intermittently missing and right censored and this censoring is often informative. We will treat only continuous measures; most of the available data can be regarded as such since the resolution of the recording instruments is high relative to the range of the measurements.

The data have many components of variability. Within an individual, we see measurement error, diurnal variation, short term variability and long-term trends. In addition, the amount of short-term variability and long-term trend vary (often greatly) between individuals. Hughes *et al.*¹ quantified within subject variability for CD4 counts. For example, in their study, mea-

measurements taken 8 weeks apart typically differed by 20 to 30 percent. This reflects both laboratory error and short-term variation. Zeger and Diggle² suggested (at least for the CD4 data they considered) that the two components of variation are of similar size. Viral load data is also extremely variable: Cowles³ reported that intra-patient standard deviations for measurements taken several days apart average $0.3 \log_{10}$ units. See also Raboud *et al.*,⁴ Paxton *et al.*,⁵ and Brambilla.⁶

Understanding the variability in these markers is extremely important from a practical standpoint. For example, it is important to determine a level of increase in CD4 count between measurements taken several months apart that is not likely to be due to random fluctuation. More generally, we need methods to extract the signal from these noisy data (with informative missingness). Furthermore, understanding of the variability is also important for joint modeling of marker and survival data. As discussed in Jewell and Nielsen,⁷ Jewell and Kalbfleisch,⁸ and Fusaro, Nielsen and Scheike,⁹ effective models for the markers are necessary for accurate and efficient survival analysis. We will focus on models which attempt to partition the variability of the data into interpretable components.

Section 2 discusses the application of mixed effects linear models to AIDS marker data. These models can be enhanced by the addition of a stochastic process term: we review such work in section 3. Section 4 describes research extending the analysis to two or more markers. Both of these sections concentrate on models that use an Integrated Ornstein-Uhlenbeck (IOU) stochas-

tic term. This stochastic process is a generalization of random trajectories (perfect derivative tracking) and Brownian motion (no derivative tracking) that allows for intermediate amounts of derivative tracking. The concept of derivative tracking is discussed in Taylor, Cumberland, and Sy.¹⁰ A model is said to have strong derivative tracking if an individual tends to stay on the same fixed path for an extended period of time and the future path is largely determined by the previous path. A model is said to have weak derivative tracking if the slope of an individual's path tends to be continually changing. Since biologically plausible models could fall anywhere in the range from perfect derivative tracking to no derivative tracking, it is desirable to let the data determine the degree of tracking.

The use of various models to study joint marker and survival data is discussed in section 5. Finally, we demonstrate our preferred modeling techniques on viral RNA and CD4 data from a Veterans Administration study in section 6.

2 Mixed effects models

The simplest effective models for marker data assume that an individual's observed values consist of a piecewise polynomial underlying trajectory (with random coefficients) plus independent error, which is usually assumed to incorporate both measurement error and short-term variability. These models

can be written in the general form:

$$Y_i(t_{ij}) = X_i(t_{ij})\boldsymbol{\alpha} + Z_i(t_{ij})\boldsymbol{\beta}_i + \epsilon_{ij} \quad (2.1)$$

where $Y_i(t_{ij})$ represents the (appropriately transformed) j th ($j = 1, \dots, J_i$) measurement made on the i th person ($i = 1, \dots, n$) at time t_{ij} , $X_i(t_{ij})\boldsymbol{\alpha}$ represent the fixed effects, $Z_i(t_{ij})\boldsymbol{\beta}_i$ the random effects, and ϵ_{ij} the error (assumed independent, but not necessarily identically distributed). The $\boldsymbol{\beta}_i$ s are assumed to be independently and identically distributed according to some distribution (usually with unknown hyperparameters). The measurements are usually transformed to make plausible an assumption of normally distributed errors. Notice that the times t_{ij} of measurement are indexed by person to indicate that the measurement times can vary between individuals. These times might represent time since infection, time since entry into a study, time since randomization to an arm of a clinical trial, or calendar date.

Each group of J_i measurements on an individual i has dispersion matrix $\text{var}(Y_i) = Z_i R_i Z_i' + \Sigma_i$ where $R_i = \text{var}(\boldsymbol{\beta}_i)$ and $\Sigma_i = \text{var}(\boldsymbol{\epsilon}_i)$. Typically each of these matrices would be parametrized by a small number of unknown parameters which are common to all individuals. As a simple example, for the model

$$Y_i(t_{ij}) = a_i + b_i t_{ij} + \epsilon_{ij}$$

where (a_i, b_i) are distributed as a bivariate normal and the ϵ_{ij} are assumed

to be independent and identically distributed (iid) normal random variables, there are a total of 6 unknown parameters in the model — 5 for the bivariate normal distribution and 1 for the measurement error variance. This type of model was popularized by Laird and Ware¹¹ and then applied to AIDS marker data in several papers.^{12–17}

DeGruttola, Lange, and Dafni¹³ used this model with one additional complexity: since the times since HIV infection are unknown, they modeled $t_{ij} = c_{ij} - \tau_i$ where c_{ij} is the chronological time of the measurement and τ_i is the unknown calendar date of infection for the i th person. Their model also allowed for a level change in the marker value at seroconversion. They used an EM-type algorithm to fit this model to a data set of CD4 values from 201 individuals in the San Francisco Men’s Health Study (SFMHS) all of whom were HIV positive at entry time and had a complete set of five CD4 measurements. They made use of an independent estimate of infection time distribution¹⁸ to impute the missing dates of infection.

Lange, Carlin, and Gelfand¹⁴ built on this model by assuming that the random slopes and intercept depend on covariates (age and presence or absence of herpes). A fully-Bayesian specification is developed by incorporating prior information where it exists and using suitably vague priors where it does not. They also allowed for heterogeneous errors and missing data. They used the Gibbs sampler to fit the model to data from 327 individuals in the SFMHS who had as many as eight measurements on CD4 levels (spaced about 6 months apart, generally). Their machinery allowed for a more gen-

eral piecewise linear trajectory with random knot points, however their data set was fit best by the simple linear growth curve assumption.

Carlin¹⁵ studied a clinical trial in 467 HIV-infected individuals with CD4 counts of below 300 (or with an AIDS diagnosis) for whom AZT was either ineffective or not tolerated. They were randomized to take either didanosine (ddI) or zalcitabine (ddC) and their CD4 counts were measured at 0, 2, 6, 12 and 18 months after entry. About 40 percent of the data were missing especially in the later visits. Clinicians expected to see a slowing in CD4 drops after two months, so a piecewise linear growth model with a change point at 2 months was used. Random effects for population trajectory, difference in population trajectory for ddC vs ddI and difference in population trajectory for AIDS cases versus non-AIDS cases were also included. Markov chain Monte Carlo (MCMC) methods were used to generate simulations from the posterior distribution of the parameters.

Kiuchi *et al.*¹⁶ focused on the Laird-Ware model with homogeneous errors and a piecewise linear growth curve with one change point. Their model attempted to capture the hypothesized rapid decline of CD4 at a time just prior to AIDS diagnosis. Thus, each individual had their own random change point. They studied 131 of the 850 men from the New York blood center cohort study. These 131 men were diagnosed with AIDS sometime during the six years of the study. Each individual had between 1 and 17 observations of CD4 count. They fit their model with both the EM algorithm and MCMC sampling.

Shi, Weiss, and Taylor¹⁷ put the model in equation (2.1) into a semi-parametric framework to model the CD4 counts of 92 infants born to HIV infected mothers in Los Angeles. They modeled both the fixed and random effects parts of the model with cubic B -splines:

$$Y_i(t_{ij}) = B(t_{ij})\boldsymbol{\alpha} + C(t_{ij})\boldsymbol{\beta}_i + \epsilon_{ij}$$

where $B(t)$ is a cubic B -spline basis of dimension K evaluated at time t and $C(t)$ is a transformation of this basis of reduced dimension L . Cross validation was used to choose the number of knots for the fixed and random effects portions of the model. They found K between 5 and 7 and L of 2 or 3 to be appropriate for fitting the data. Given K and L , the model is a mixed effects model as in equation (2.1).

McNeil¹⁹ used a detailed fully Bayesian random effects model to analyze CD4 counts in the context of a cohort study on AZT effect. They looked at ways to validate such models relative to others. They also gave a clear explanation of the need for cautious interpretations of effects in observational studies.

3 Stochastic modeling

Diggle²⁰ suggested the addition of a stochastic process term such as Brownian motion (BM) to each individual's model equation as a general approach for

modeling longitudinal data. A general form of such a model can be written:

$$Y_i(t_{ij}) = X_i(t_{ij})\boldsymbol{\alpha} + Z_i(t_{ij})\boldsymbol{\beta}_i + W_i(t_{ij}) + \epsilon_{ij} \quad (3.1)$$

where $W_i(t_{ij})$ is the value of a mean zero stochastic process at time t_{ij} and $Y_i(t_{ij})$ is the observed data or some transformation of it. We usually assume that the process is Gaussian, but not necessarily stationary. This model partitions the observations into four parts. $X_i\boldsymbol{\alpha}$ represents the mean pattern for the population. $Z_i\boldsymbol{\beta}_i$ gives an individual level deterministic model of change (for a given $\boldsymbol{\beta}_i$). W_i gives a (usually continuous) stochastic variation about this deterministic path. The inclusion of W is conceptually important, because without W we would be assuming that an individual's path, over a possibly long time interval, would be described by a simple curve determined by a small number of parameters. Finally, the ϵ_{ij} are independent measurement errors (which are indistinguishable from very short time-scale biological variation). It is important to have these independent errors in the model to capture the fact that 2 measurements taken a short time apart are not necessarily similar.

The covariance structure of the J_i measurements on the i th individual can be written as $\text{var}(Y_i) = Z_i R_i Z_i' + M_i + \Sigma_i$ where $M_i = \text{var}(W_i)$. These three components of variance provide a very rich class of models. It is therefore quite important to use parsimonious parametrizations. For example, $Z\boldsymbol{\beta}_i$

may just represent a random intercept. A large amount of data may be required to be able to allocate the variability and heterogeneity to the three components since, in some sense, the three terms are all competing for this variability. One technique to avoid these identifiability problems might be to use available prior information on the size of variance components.

There are several specific applications of these models to AIDS data.^{2,10,21-27} Berman²¹ proposed modeling natural log of CD4 by a stationary Gaussian process plus unknown negative drift. This model has the form:

$$Y_i(t_{ij}) = \mu - \beta t_{ij} + W_i(t_{ij}). \quad (3.2)$$

This model is fit to a data set of CD4 counts from intravenous drug users in New York City. Kanazawa²² allowed the Gaussian process to be non-stationary; they modeled log CD4 as an unknown negative drift plus Brownian motion. This is the same model as in equation (3.2) except that W is restricted to be a Brownian motion. Galai, Muñoz, and Chen²³ proposed a stochastic model with damped exponential correlations. Specifically, they used:

$$Y_i(t_{ij}) = X_i(t_{ij})\boldsymbol{\alpha} + W_i(t_{ij}).$$

where

$$\text{corr}(W_i(s), W_i(t)) = \gamma^{|t-s|^\theta}.$$

Special cases of this model include an AR(1) stochastic term ($\theta = 1$) and

constant autocorrelation ($\theta = 0$). They fit this model (using a piecewise linear mean structure) to marker measurements on 328 seroconverters in the Multicenter AIDS Cohort Study (MACS). They looked at univariate models for CD4 (natural scale), log of β_2 -microglobulin and log of neopterin. We note that for all three of these papers, the model does not contain an independent measurement error term.

Zeger and Diggle² included a term for a smooth trend function (which is estimated through locally adaptive kernel smoothing methods — this smoothing estimate was shown to be nearly equivalent to a spline estimate by Silverman.²⁸ This smooth curve is combined with a fixed effect term and a stationary Gaussian process term:

$$Y_i(t_{ij}) = X_i(t_{ij})\boldsymbol{\alpha} + \mu(t_{ij}) + W_i(t_{ij}) + \epsilon_{ij}.$$

They used the covariance function:

$$\text{cov}(W(s), W(s + u)) = \sigma_w^2 \{\lambda + (1 - \lambda) \exp(-\theta u)\}$$

The model was fit to square roots of CD4 counts from the seroconverters in the MACS study (369 individuals at the time this paper was written).

Vittinghoff, Malani, and Jewell²⁴ also used local linear smoothing methods in a model with a stochastic term component to look at square roots of CD4 counts in the San Francisco Men's Health Study. They introduced two different methods for estimating the underlying curve patterns.

Taylor, Cumberland, and Sy¹⁰ modeled the fourth root of CD4 counts using an Integrated Ornstein-Uhlenbeck (IOU) process. An IOU process is a zero-mean non-stationary Gaussian process with covariance function:

$$\text{cov}(W(s), W(s+u)) = \frac{\sigma_w^2}{2\lambda^3} \left[2\lambda s + e^{-\lambda(s+u)} + e^{-\lambda s} - 1 - e^{-\lambda u} \right].$$

The two parameters σ_w^2 and λ control the variability of the process and the degree to which the derivatives “track”. A small λ and σ_w^2 give a process which tends to continue on similar trajectories for long periods of time. Conversely, a large λ and σ_w^2 result in a process which has very little “memory” for past derivatives. By letting $\lambda \rightarrow \infty$ while holding σ^2/λ^2 equal to a constant (v , say) scaled Brownian motion is obtained, i.e. $\text{cov}(W(s), W(s+u)) = vs$.

The IOU process was put into a model along with random linear trend and measurement error:

$$Y_i(t_{ij}) = a_i + b_i t_{ij} + W_i(t_{ij}) + \epsilon_{ij}$$

where (a_i, b_i) has a bivariate normal distribution and the ϵ_{ij} were assumed to be iid normal with common variance σ_e^2 and $Y_i(t_{ij})$ is the fourth root of the CD4 count at time t_{ij} . This model, therefore, has 8 unknown parameters — 5 for the distribution of (a, b) , 2 for the covariance structure of W and 1 for the measurement error variance. In practice, there was confounding between the random slope and the stochastic process parameters. A slightly reduced

model with a non-random slope, i.e.:

$$Y_i(t_{ij}) = a_i + bt_{ij} + W_i(t_{ij}) + \epsilon_{ij}$$

was used instead – this reduced model has only 6 parameters. This model covers a wide range of biologically plausible behavior. It was found that the same model but with W a scaled Brownian motion process fit almost equally as well. In other words, a model with no derivative tracking was found to be consistent with the CD4 count data in this study.

Wang and Taylor²⁵ analyzed data from an AIDS clinical trial with a model of the form:

$$Y_i(t_{ij}) = \mu_g(t_{ij}) + W_i(t_{ij}) + \epsilon_{ij}$$

where $\mu_g(t)$ is a smooth mean curve for group g . In particular, each μ_g was taken to be a cubic polynomial spline. To avoid overfitting, the splines were fit through penalized likelihood, with roughness penalty λ (which is selected through cross-validation techniques). For a given value of λ , though, this can be put under the framework of equation (3.1) if prior information is allowed on the fixed effect parameters α in that model. Several choices for $W(t)$ were considered including a generalized AR(1) type process and a random effects model. The model was applied to data from AIDS Clinical Trial Group 016 (ACTG 016) which followed 61 HIV positive patients, 34 randomized to AZT and 27 to placebo.

LaValley and DeGruttola²⁷ combined ideas from Taylor, Cumberland, and Sy¹⁰ and Tsiatis, DeGruttola, and Wulfsohn²⁹ in the analysis of a three arm AIDS clinical trial, with particular focus on evaluating whether CD4 is a good surrogate endpoint.

Taylor and Law²⁶ compared the importance of the particular choice of covariance structure for predicting future CD4 values. The main finding of this paper was that the choice of covariance structure is important for individual level predictions. The IOU and Brownian motion models appeared not only to give a better fit (based on likelihood) to CD4 counts from the MACS study, but also to be more efficient and accurate in making future predictions (based on the MACS data and a simulation study). Random effects models in the cases examined, appeared to fit the MACS data reasonably well, but did not perform as well in predictions for individuals. CD4 counts were transformed by square root, cube root and fourth root with no difference in results.

4 Bivariate marker models

Sy, Taylor, and Cumberland³⁰ extended the model in Taylor, Cumberland, and Sy¹⁰ (section 3) to joint modeling of both CD4 and beta-2-microglobulin measurements using a bivariate IOU process (see e.g. Jones³¹). Let B and C be $k \times k$ real matrices subject to certain conditions including C symmetric

(see Cox,³² p. 226). The k -variate IOU process has covariance function:

$$\text{cov}(\mathbf{W}(s), \mathbf{W}(s+u)) = \left(e^{B's} - I - B's \right) B'^{-2}C + CB^{-2} \left(e^{B(s+u)} - e^{B(u)} - Bs \right), \quad (4.1)$$

for s and u non-negative. If the eigenvalues of B tend to minus infinity while holding CB^{-1} constant, we obtain $\text{Cov}(\mathbf{W}(s), \mathbf{W}(s+u)) = \mathbf{V}s$, where $\mathbf{V}$ is positive-definite and symmetric. This is the covariance function of a bivariate scaled Brownian motion process with variances v_{11} and v_{22} and inter-variable covariance v_{12} . The bivariate version of the IOU process has 4 parameters for B and 3 for C (since it is symmetric), for a total of 7 parameters; the bivariate Brownian motion process has only 3 parameters for $\mathbf{V}$. The off-diagonal elements of B have an interesting possible interpretation. If, for example, B is lower triangular, then a change in the derivative of the first process is affected by the derivatives of both processes. However, a change in the derivative of the second process is only affected by the derivative of the second process. That is, the model allows for asymmetry in how Y^1 and Y^2 influence each other.

The full model considered was

$$\begin{aligned} Y_i^1(t_{ij}) &= a_i + b_i t_{ij} + W_i^1(t_{ij}) + \epsilon_{ij}^1 \\ Y_i^2(t_{ij}) &= c_i + d_i t_{ij} + W_i^2(t_{ij}) + \epsilon_{ij}^2 \end{aligned}$$

where Y_i^1 and Y_i^2 are the two transformed marker variables of interest,

(a_i, b_i, c_i, d_i) have a multivariate normal distribution, (W_i^1, W_i^2) is a bivariate IOU process, and $(\epsilon_i^1, \epsilon_i^2)$ are independent and normally distributed with variances $\sigma_{\epsilon_1}^2$ and $\sigma_{\epsilon_2}^2$, respectively. This model was fit to data from the Los Angeles portion of the MACS study. CD4 was transformed by a fourth root and β_2 -microglobulin by logarithm. As in the univariate case, it was difficult to differentiate between the stochastic process parameters and the random slopes. Therefore, b and d were constrained to be fixed effects in conjunction with the IOU process (or the Brownian motion special case); random b_i and d_i were still used for the random effects model.

5 Joint models for markers and survival

More recently, focus has centered around joint modeling of marker and survival data. Intuitively, the longitudinal pattern of the marker data should give more information about survival than simple fixed covariate models.

Jewell and Kalbfleisch⁸ stressed the importance of joint modeling of marker and survival data for AIDS research. They discuss informative censoring and gave a framework for doing analysis.

Tsiatis, DeGruttola, and Wulfsohn²⁹ developed a joint model for CD4 marker data and time to death for 281 patients in a clinical trial of AZT. The model for logarithm of CD4 was random piecewise linear (with a knot to allow for AZT effect). The parameters in this model were put into a proportional hazards model. Their results suggested that CD4 is not entirely

satisfactory as a surrogate marker for clinical endpoints such as progression to AIDS or death.

DeGruttola and Tu³³ studied 511 patients enrolled between 1986 and 1987 in protocol 002 of the ACTG. 421 of these patients had died by 1990. They used a random piecewise quadratic model for logarithm of CD4 together with an accelerated failure time model (AFT) for survival times. Faucett and Thomas³⁴ looked at 109 HIV positive individuals who had contracted the virus through blood transfusion (and for whom the exact date of infection was known). Log CD4 and square root CD4 were both modeled by random linear trajectories. The parameters from this model were incorporated into a proportional hazards model for time to AIDS progression. Their model was fully Bayesian and was fit using Gibbs sampling.

Pawitan and Self³⁵ combined a random piecewise linear marker model with a Weibull survival model to study time to AIDS progression for 159 patients in the Toronto AIDS cohort study. They considered both log of CD4 and log of the CD4 to CD8 ratio. They found that the latter marker was a better predictor of time to AIDS.

Cowles³ used a random effects model for logarithm of viral RNA measurements to look at the 37 patients in ACTG 116B/117 who received continuing AZT treatment. The parameters for the marker model were put into an accelerated failure time model with time varying covariates to look at time to AIDS progression. A fully Bayesian model was specified and fit using MCMC methods.

Doksum and Normand³⁶ combined stochastic process models (similar to those of Berman^{21,37} as discussed in section 3) with survival data from the San Francisco Mens Health Study. This work built on the ideas in Doksum.³⁸

6 Analysis of viral load data

We illustrate the methods described in this paper using some data from an AIDS clinical trial. The trial we consider is VA Cooperative Group trial 298 of early versus late AZT.^{39,40} The stochastic longitudinal models described in sections 3 and 4 will be fit to the CD4 and viral load observations in the late treatment arm. We consider the three year follow-up period prior to the termination of the double blinded portion of the study, and we exclude observations from patients after they switched to AZT. We further restrict analysis to the subset of patients with at least one viral load measurement and exclude the baseline data (the patients were enrolled into the study subject to a certain range of CD4 count — thus, the variance at baseline was much smaller than the model assumes). This restricted data set is essentially an untreated cohort of 141 HIV infected subjects. The protocol called for measurement at 1, 2, 4, 8, 12, 16, 20 and 24 months after randomization, but most subjects had either fewer or more measurements than this. Many subjects had an additional viral load measurement at 7 days past randomization which is included in the analysis. The mean follow-up time was 15.4 months. The total number of CD4 count measurements was 817. Ninety percent of

the patients had between one and ten CD4 counts. The total number of viral load measurements was 659. The number of viral loads measurements on an individual was between one and twelve for ninety percent of the patients.

We first fit a univariate model to the base 10 logarithm of the viral load data (denoted Y^H , for number of copies of HIV) with bivariate normal random effects for the intercept and slope, i.e.

$$Y_i^H(t_{ij}) = a_i + b_i t_{ij} + \epsilon_{ij}^H.$$

where the measurement error is assumed iid normal with variance σ_H^2 . We refer to this model as RE(a, b) + ME. We next fit the same model with the addition of a scaled Brownian motion term. This term is parameterized by v_H : the variance of the stochastic process at time 1. As explained in section 3, we found that we needed to restrict the slope parameter to be a fixed effect giving us a total of 5 parameters. This model will be denoted RE(a) + FE(b) + BM + ME. We also fit an IOU model (which has one more parameter than the Brownian motion model), but it is not shown as the parameter estimates converged towards the Brownian motion estimates, with no significant improvement in the log-likelihood. The results for fitting the two univariate models to the viral load data are given in table 1.

The log likelihood for the Brownian motion model is a bit higher. The estimates of the population intercepts and slope are very similar for both models, as we had observed before with CD4 data.¹⁰ The models indicate

Model	No. of Params.	Log- Likelihood	Intercept $\hat{a}$ (SE)	Slope $\hat{b}$ (SE)	Covariance Parameter	ME σ_H
1. RE(a, b) + ME	6	-684.0	4.06 (.07)	0.04 (.05)	$\sigma_{ab} = -0.23$	0.55
2. RE(a) + FE(b) + BM + ME	5	-676.0	4.06 (.07)	0.08 (.07)	$v_H = 0.29$	0.48

Table 1: Parameter estimates for fitting univariate models to the viral load data.

a baseline level of viral load of around 4.05 base-10 logs, or about 11,000 copies. The slope coefficient indicates a slow increase of around 0.05 per year (increase of about 10% in natural units), although this estimate is not significantly different from zero for either model. In comparison, this effect is much smaller than the observed effect of AZT or other current treatments. Model 2 gives us plausible parameter estimates, one less parameter and a greater log-likelihood. Based on our previous work, we feel that this model quite likely will give more accurate predictions of future marker values on an individual.²⁶ Furthermore, the IOU model (which is capable of exhibiting either random trajectory or Brownian motion-type behavior) suggests that no derivative tracking is most in agreement with the data.

Next, we combine the viral load data with the CD4 data. The CD4 counts are transformed by fourth roots¹⁰ and denoted as Y^C . We consider a variety of random effects and stochastic models for the bivariate data. The random effects models have the form:

$$\begin{aligned}
Y_i^H(t_{ij}) &= a_i + b_i t_{ij} + \epsilon_{ij}^H \\
Y_i^C(t_{ij}) &= c_i + d_i t_{ij} + \epsilon_{ij}^C.
\end{aligned}$$

The first model assumes a general four-dimensional normal distribution for (a_i, b_i, c_i, d_i) . This model therefore has 16 parameters: 14 for the multivariate normal and 2 for the measurement error variances. We will abuse the shorthand introduced in the univariate model and refer to this model as $\text{RE}(a, b, c, d) + \text{ME}$. This is model 1a in tables 2 and 3. Model 1b assumes that (a_i, b_i) and (c_i, d_i) have independent bivariate normal distributions. This model has 12 parameters — we will write it as $\text{RE}(a, b) + \text{RE}(c, d) + \text{ME}$.

Next, we looked at adding a stochastic term to the bivariate random effects models. As for the univariate models, we need to restrict the slope parameters to be fixed effects, with the intercepts remaining as random effect terms, possibly correlated. We considered three different models with Brownian motion. Model 2a adds a bivariate Brownian motion term for a total of 12 parameters. The shorthand for this model is $\text{RE}(a, c) + \text{FE}(b, d) + \text{BVBM} + \text{ME}$. Model 2b assumes independent Brownian motion terms (IBM) and has 11 parameters. Finally, model 2c assumes completely independent univariate models, $\text{RE}(a) + \text{RE}(c) + \text{FE}(b, d) + \text{IBM} + \text{ME}$, for 10 parameters. We also fit several types of bivariate IOU models by placing different restrictions on the covariance parameters in the B matrix (see equation (4.1)). However, we found that the restricted models either fit poorly or else converged towards the bivariate Brownian motion model. The unrestricted bivariate IOU model (which has 16 parameters) fit only slightly better than the bivariate Brownian motion, had very large negative eigenvalues for the B matrix, and had similar parameter estimates to the bivariate Brownian motion. Therefore,

we restrict attention in this paper to the Brownian motion models since the data seem to support very little derivative tracking.

Tables 2 and 3 show the results of fitting these models to the viral load and CD4 counts. The Brownian motion models fit better as measured by the likelihood than their random effects counterparts (i.e. model 2a, bivariate Brownian motion, fits better than model 1a, bivariate random effects; model 2c, independent Brownian motions, fit better than model 1b, independent random effects), and, in each case, have fewer parameters. Within the Brownian motion models, the bivariate model fits significantly better than either of the independent stochastic models. All the models estimate a starting level for fourth root of CD4 of around 4.3 (around 350 on the natural scale) with a decrease of -0.25 per year. We remark again, that this is small compared to the immediate increase seen after treatment by AZT or other drug regimens.

For random effects model 1a, the intervariable covariance parameters in table 3 have sensible interpretations: higher initial CD4 counts are associated with lower initial viral loads, as reflected in a negative correlation σ_{ac} ; steeper CD4 decreases are associated with steeper viral load increases, so σ_{bd} is negative. For the cross-parameter covariances, steeper CD4 declines tend to be seen in patients with high initial viral load (σ_{ad} is negative), and higher initial CD4 levels are associated with faster viral load increases (σ_{bc} is positive). This last observation probably reflects the fact that the virus needs CD4 cells to replicate.

Model	No. of Params	Log- Likelihood	HIV slope $\hat{b}$ (SE)	CD4 Slope $\hat{d}$ (SE)
1a. RE(a, b, c, d) + ME	16	-953.6	0.07 (.05)	-0.26 (.03)
1b. RE(a, b) + RE(c, d) + ME	12	-971.5	0.04 (.05)	-0.25 (.03)
2a. RE(a, c) + FE(b, d) + BVBM + ME	12	-943.9	0.15 (.07)	-0.24 (.03)
2b. RE(a, c) + FE(b, d) + IBM + ME	11	-951.1	0.09 (.07)	-0.24 (.02)
2c. RE(a) + RE(c) + FE(b, d) + IBM + ME	10	-962.9	0.08 (.07)	-0.24 (.03)

Table 2: Model specifications, log-likelihoods and parameter estimates for fitting bivariate models to the viral load and CD4 data.

Model	Covariance Parameters			ME	
				σ_H	σ_C
1a	$\sigma_{ac} = -0.44$	$\sigma_{bd} = -0.52$	$\sigma_{ad} = -0.30, \sigma_{bc} = 0.34$	0.54	0.27
1b				0.55	0.27
2a	$v_H = 0.34$	$v_C = 0.06$	$v_{H,C} = -0.09$	0.48	0.25
2b	$v_H = 0.30$	$v_C = 0.05$		0.48	0.25
2c	$v_H = 0.29$	$v_C = 0.05$		0.48	0.25

Table 3: Covariance parameter estimates for fitting bivariate models to the viral load and CD4 data.

For the Brownian motion models, table 3 lists the elements of V — the dispersion matrix for the bivariate Brownian motion process as described in section 4. For model 2a, we observe that the off-diagonal element $v_{H,C} = -0.09$. This translates into a correlation of $-0.09/(\sqrt{0.34}\sqrt{0.06}) = -0.63$ between the two components of the stochastic process at a given time. (For two different times, this correlation is weighted by the square root of the ratio of the earlier to the later time). The two stochastic process terms seem to have substantial negative correlation which explains why models 2b and 2c (where the off-diagonal is constrained to be zero) do not fit nearly as well.

One of the more informative ways to compare these models is through

the predictions they make. Consider an individual with a viral load of 4.0 logs at one month (the median value at that time in the data set). Figure 1a shows for the bivariate Brownian motion model (model 2a) the prediction of viral load out to 12 months for this individual conditional on different levels of CD4 at one month. That is:

$$\mathbf{E}(Y^H(t)|Y^H(1) = 4.0, Y^C(1) = x), \text{ for } t = 2, \dots, 12$$

The top line is for a transformed CD4 count equal to 3.9 (a count of 231 on the natural scale), the middle line is the prediction for Y^C equal to 4.3, and the bottom line is for Y^C of 4.7 (these correspond to 345 and 488 respectively on the natural scale). These three CD4 values were chosen to be the 5th, 50th and 95th percentiles of the values in the data set at one month. Figure 1b shows the same plot for the random effects model (model 1a). Notice that the positive covariance σ_{bc} causes the slope of the outer lines to “pinch” in towards the middle line in figure 1b. On the other hand, the Brownian motion model predicts the same slope for all individuals. Also it is worth noting that the pointwise prediction standard errors for all these lines are around 0.6 log units which is larger than the range of the vertical axis — the difference in predictions is small compared to the standard deviation at baseline.

Figures 1 c) and d) are analogous to a) and b) for predicting CD4 values given transformed CD4 counts equal to 4.3 at time one month. Predictions

are shown for the 5th, 50th and 95th percentile values of log viral load at one month (top line corresponds to Y^H equal to 2.2, middle line to 4.0 and bottom line to 5.3). Here, we see for the random effects model, that the negative σ_{ad} causes the upper and lower lines to diverge from the middle line.

These predictions show graphically the extent to which the processes are linked, and indicate that the joint modeling is definitely worth the effort. Separate models would generate identical prediction curves, and this does not seem clinically plausible.

7 Discussion

We have seen that the class of models discussed in sections 3 and 4 is a general class which includes the simple random effects model as a special case. In general, Diggle²⁰ recommends including a stochastic process term in longitudinal modeling. The resulting class of models is much richer and yet retains intuitive interpretability of its parameters.

For the VA data set in section 6, the model that fit best used random intercept, Brownian motion and independent measurement error. We attempted to fit a generalization of this model using IOU processes, but identifiability of the parameters was difficult. However, the highest likelihoods occurred at parameter values consistent with the Brownian motion solution. Certainly, this data set seems to support only a very limited amount of derivative tracking

Figure 1: Predictions of base-10 logarithm of viral load (top two panels) and fourth root of CD4 count (bottom two panels) for different initial values of CD4 count and viral load. Left panels are predictions for model 2a and right panels for model 1a.

in individual marker processes. Based on this and also the likelihood values and parsimony considerations, we prefer the Brownian motion model 2a for this data set.

One solution to the identifiability problem for the IOU models might be to use a Bayesian approach with informative prior distributions on some of the variance components. A Bayesian treatment makes particular sense in this setting as there is a fair amount of prior information on the size of these components. Inference would be done via MCMC simulations from the posterior distribution of the parameters.

Another direction for generalizing the covariance structures, might be antedependence models (see, for example, Zimmerman and Nuñez-Anton⁴¹). These models also have interpretable covariance parameters, and the IOU and Brownian motion models can be seen as special cases. Over-parametrization will certainly be an important issue here, and we would most likely need to use prior information in our model.

8 Acknowledgements

This work was supported by grants AI 29196 and AI 73070 from the National Institutes of Health and by grant #50011 from the National Institute of General Medical Sciences. The authors thank Pamela Hartigan for providing the data from the VA Cooperative Group trial.

References

- 1 Hughes MD, Stein DS, Gundacker HM, *et al.*. Within-subject variation in CD4 lymphocyte count in asymptomatic human immunodeficiency virus infection: Implications for patient monitoring. *The Journal of Infectious Diseases* 1994; **169**: 28–36.
- 2 Zeger SL, Diggle PJ. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* 1994; **50**: 689–699.
- 3 Cowles MK. Bayesian joint accelerated failure time models for surrogate markers and survival time. Preprint.
- 4 Raboud JM, Montaner JSG, Conway B, *et al.*. Variation in plasma RNA levels, CD4 cell counts, and p24 antigen levels in clinically stable men with Human Immunodeficiency Virus. *Journal of Infectious Diseases* 1996; **174**: 191–194.
- 5 Paxton WB, Coombs RW, McElrath MJ, *et al.*. Longitudinal analysis of quantitative virologic measures in Human Immunodeficiency Virus-infected subjects with ≥ 400 CD4 lymphocytes: Implications for applying measurements to individual patients. *The Journal of Infectious Diseases* 1997; **175**: 247–254.
- 6 Brambilla D. Variability of HIV viral load markers. Presented at ENAR Spring Meetings.

- 7 Jewell N, Nielsen JP. A framework for consistent prediction rules based on markers. *Biometrika* 1993; **80**(1): 153–164.
- 8 Jewell NP, Kalbfleisch JD. Marker processes in survival analysis. In: Jewell N, Dietz K, Farewell V, eds., *AIDS Epidemiology: Methodological Issues*, Boston: Birkhauser-Boston. 1992; pages 211–230, pages 211–230.
- 9 Fusaro RE, Nielsen JP, Scheike TH. Marker-dependent hazard estimation: an application to AIDS. *Statistics in Medicine* 1993; **12**: 843–865.
- 10 Taylor JMG, Cumberland WG, Sy JP. A stochastic model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association* 1994; **89**(427): 727–736.
- 11 Laird NM, Ware JH. Random effects models for longitudinal data. *Biometrics* 1982; **38**: 963–974.
- 12 Muñoz A, Cary V, Saah AJ, *et al.* Predictors of decline in CD4 lymphocytes in a cohort of homosexual men infected with human immunodeficiency virus. *Journal of the Acquired Immune Deficiency Syndrome* 1988; **1**: 396–404.
- 13 DeGruttola V, Lange N, Dafni U. Modeling the progression of HIV infection. *Journal of the American Statistical Association* 1991; **86**(415): 569–577.
- 14 Lange N, Carlin BP, Gelfand AE. Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers. *Journal of the American Statistical Association* 1992; **87**(419): 615–632.

- 15 Carlin BP. Hierarchical longitudinal modeling. In: Gilks W, Richardson S, Spiegelhalter D, eds., *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall. 1996; pages 303–319, pages 303–319.
- 16 Kiuchi AS, Hartigan JA, Holford T, Rubinstein P, Stevens CE. Change points in the series of T4 counts prior to AIDS. *Biometrics* 1995; **51**: 236–248.
- 17 Shi M, Weiss RE, Taylor JMG. An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Applied Statistics* 1996; **45**(2): 151–163.
- 18 Bacchetti P, Moss AR. Incubation period of AIDS in San Francisco. *Nature* 1989; **338**: 251–253.
- 19 McNeil AJ. Bayes estimates for immunological progression rates in HIV disease. Preprint.
- 20 Diggle PJ. An approach to the analysis of repeated measurements. *Biometrics* 1988; **44**: 959–971.
- 21 Berman SM. A stochastic model for the distribution of HIV latency time based on T4 counts. *Biometrika* 1990; **77**(4): 733–741.
- 22 Kanazawa Y. Estimating expected time of infection with HIV. Technical Report 13, SIMS (Statistical Methodology for study of the AIDS Epidemic), 1990.

- 23 Galai N, Muñoz A, Chen K, *et al.*. Tracking of markers and onset of disease among HIV-1 seroconverters. *Statistics in Medicine* 1993; **12**: 2133–2145.
- 24 Vittinghoff E, Malani HM, Jewell NP. Estimating patterns of CD4 lymphocyte decline using data from a prevalent cohort of HIV infected individuals. *Statistics in Medicine* 1994; **13**: 1101–1118.
- 25 Wang Y, Taylor JMG. Inference for smooth curves in longitudinal data with application to an AIDS clinical trial. *Statistics in Medicine* 1995; **14**: 1205–1218.
- 26 Taylor JMG, Law N. Does the covariance structure matter in longitudinal modelling for the prediction of future CD4 counts? Preprint.
- 27 LaValley MP, DeGruttola V. Models for empirical Bayes estimators of longitudinal CD4 counts. *Statistics in Medicine* 1997; .
- 28 Silverman BW. Spline smoothing: the equivalent variable kernel method. *Annals of Statistics* 1984; **12**: 898–916.
- 29 Tsiatis AA, DeGruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error. applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* 1995; **90**(429): 27–37.
- 30 Sy JP, Taylor JMG, Cumberland WG. A stochastic model for the analysis of bivariate longitudinal AIDS data. *Biometrics* 1997; **53**: 542–555.

- 31 Jones RH. *Longitudinal Data with Serial Correlation: a State-Space Approach*. London: Chapman & Hall, 1993.
- 32 Cox DR, Miller HD. *The Theory of Stochastic Processes*. New York: Wiley, 1965.
- 33 DeGruttola V, Tu XM. Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics* 1994; **50**: 1003–1014.
- 34 Faucett CL, Thomas DC. Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statistics in Medicine* 1996; **15**: 1663–1685.
- 35 Pawitan Y, Self S. Modeling disease marker processes in AIDS. *Journal of the American Statistical Association* 1993; **88**(423): 719–726.
- 36 Doksum KA, Normand SLT. Gaussian models for degradation processes – Part I: Methods for the analysis of biomarker data. *Lifetime Data Analysis* 1995; **1**: 131–144.
- 37 Berman SM. Perturbation of normal random vectors by nonnormal translations, and an application to HIV latency time distributions. *The Annals of Applied Probability* 1994; **4**(4): 968–980.
- 38 Doksum KA. Degradation models for failure time and survival data. *CWI Quarterly* 1991; **4**(4): 199–205.
- 39 Hamilton JD, Hartigan PM, Simberkoff MS, *et al.*. A controlled trial of early versus late treatment with zidovudine in symptomatic Human

- Immunodeficiency Virus infection. *The New England Journal of Medicine* 1992; **326**: 437–443.
- 40 O'Brien WA, Hartigan PM, Martin D, *et al.* Changes in plasma HIV-1 RNA and CD4+ lymphocyte counts and the risk of progression to AIDS. *The New England Journal of Medicine* 1996; **334**: 426–431.
- 41 Zimmerman DL, Nuñez-Anton V. Structured antedependence models for longitudinal data. Technical Report 253, University of Iowa, September 1996.