An Analysis of Paediatric CD4 Counts for Acquired Immune Deficiency Syndrome Using Flexible Random Curves
Author(s): Minggao Shi, Robert E. Weiss and Jeremy M. G. Taylor
Source: *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 45, No. 2 (1996), pp. 151-163
Published by: Wiley for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/2986151
Accessed: 22-04-2020 17:47 UTC

Appl. Statist. (1996)
45, *No.* 2, *pp.* 151–163

An Analysis of Paediatric CD4 Counts for Acquired Immune Deficiency Syndrome using Flexible Random Curves

By MINGGAO SHI, ROBERT E. WEISS and JEREMY M. G. TAYLOR†

University of California, Los Angeles, USA

[Received November 1994. Revised July 1995]

SUMMARY

In this paper we analyse CD4 counts from infants born to mothers who are infected with the human immunodeficiency virus. A random effects model with linear or low order polynomials in time is unsatisfactory for these longitudinal data. We develop an alternative approach based on a flexible family of models for which both the fixed and the random effects are linear combinations of B-splines. The fixed and random parts are smooth functions of time and the covariance structure is parsimonious. The procedure allows estimates of each individual's smooth trajectory over time to be exhibited. Model selection, estimation and computation are discussed. Centile curves are presented that take into account the longitudinal nature of the data. We emphasize a graphical approach to the presentation of results.

Keywords: B-spline; Centile curves; Empirical Bayes method; Graphics; Random effects; Residuals

1. Introduction

We analyse data consisting of serial CD4 counts of children born to mothers who are infected with the human immunodeficiency virus (HIV) in Los Angeles. This data set is highly unbalanced, with measurements for any one child occurring randomly in time. Children may or may not be HIV infected. Our goals in this analysis are

(a) to describe the differences in population trends of the infected and uninfected children and to compare the two groups,
(b) to model properly the group and individual changes in CD4 counts over time in both groups,
(c) to exhibit the set of changes over time and
(d) to present normal centile curves of the CD4 counts.

The paediatric data for acquired immune deficiency syndrome (AIDS) were collected by the Clinical Immunology Research Laboratory of the University of California at Los Angeles during late 1988 to mid-1994. The infection status is defined by the presence or absence of HIV antibodies 15 months after birth and by viral culture assays. Children whose first visit is at age 1 year or older are excluded

†*Address for correspondence*: Department of Biostatistics, School of Public Health, University of California at Los Angeles, Los Angeles, CA 90095-1772, USA.
E-mail: JEREMY@SUNLAB.PH.UCLA.EDU

0035–9254/96/45151

from this analysis. 92 children in the study have defined HIV status and their CD4 counts are analysed in this paper. Among them 35 (38%) are HIV infected (the infected group) and 57 (62%) are uninfected (the uninfected group). The data set consists of 508 observations of CD4 counts, 245 from the infected group and 263 from the uninfected group.

Half of the 57 uninfected children have four visits or fewer because they drop out of the study as their HIV status becomes known. In contrast over half of the 35 infected children have eight or more visits, but with few observations available for newborns. There is a large variation in the CD4 counts of the infected group with counts ranging from 14 to 9106. All observations were transformed by a fourth-root power to achieve homogeneity of within-subject variance (Taylor *et al.*, 1994). The transformed CD4 counts are presented in Fig. 1. Line segments connect consecutive observations for each child. Fig. 1 strongly suggests that the two groups should be modelled separately.

The CD4 count is a critical measure of the immune system and is used as an important marker in describing the progress to AIDS in adults. From the onset of HIV infection in adults, there is a continual loss of CD4 T-cells, with a lower CD4 count indicating more severe immune deficiency and higher risk of developing AIDS.

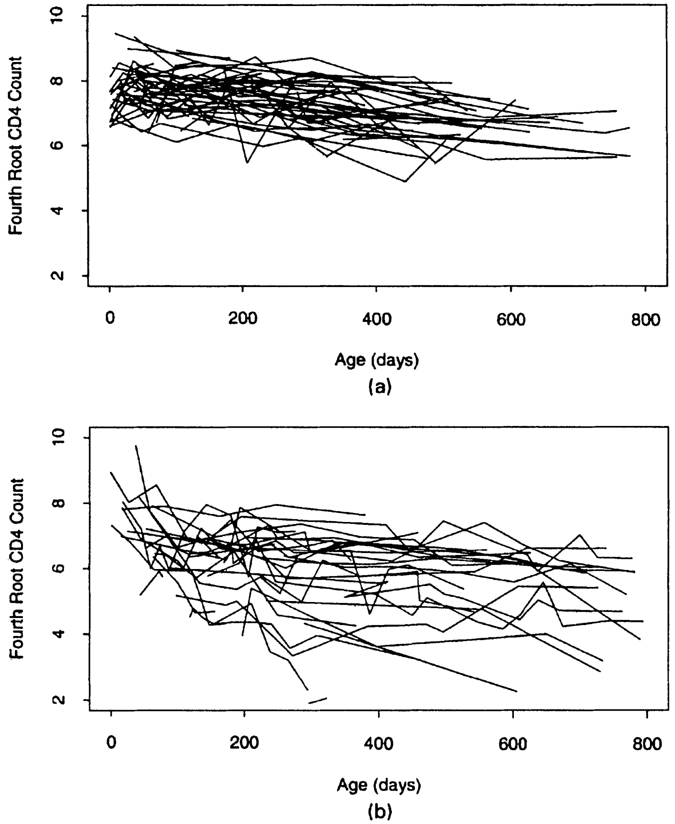

Fig. 1. Observed fourth-root CD4 counts plotted against age for (a) the uninfected group and (b) the infected group: consecutive observations from an individual are connected by line segments

The CD4 count is also an important immunological marker for HIV-infected children; however, the CD4 counts vary strongly with age in childhood. Newborn infants have substantially higher CD4 counts than adults. With increasing age, the CD4 counts in infants decline and begin to approximate those of adults; a decline in children's CD4 counts is therefore normal. The rate of CD4 count decline is often much more rapid in infected infants than in infected adults (Koup and Wilson, 1994). Others have shown that HIV-infected infants have lower CD4 counts than uninfected infants (de Martino *et al.*, 1991).

A standard approach to modelling unbalanced longitudinal data is to fit a random effects model, based on random linear or low order polynomials. We fit the random quadratic model

$$y_i(t) = f(t) + s_i(t) + \epsilon_{it}, \tag{1}$$

for $i = 1, \ldots, n$, with $f(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2$ and $s_i(t) = \gamma_{i0} + \gamma_{i1} t + \gamma_{i2} t^2$ to the fourth-root CD4 counts for each group, where $y_i(t)$ is the observation at time t from person i. The $\alpha^{\mathrm{T}} = (\alpha_0, \alpha_1, \alpha_2)$ are parameters and the function $f(t)$ is the population mean response or the fixed effect. The random vector $\gamma_i^{\mathrm{T}} = (\gamma_{i0}, \gamma_{i1}, \gamma_{i2})^{\mathrm{T}} \sim N(0, D)$ and $s_i(t)$ is the ith individual's mean and is called the random effect. The $\epsilon_{it} \sim N(0, \sigma^2)$ is random error due to measurement error or short-term fluctuations. All ϵ_{it} and all γ_i are assumed independent. The fitted individual trajectories and residuals for both infected and uninfected groups are presented in Fig. 2. Scientifically, it is implausible that CD4 counts on any scale follow quadratic trends for any extended period of

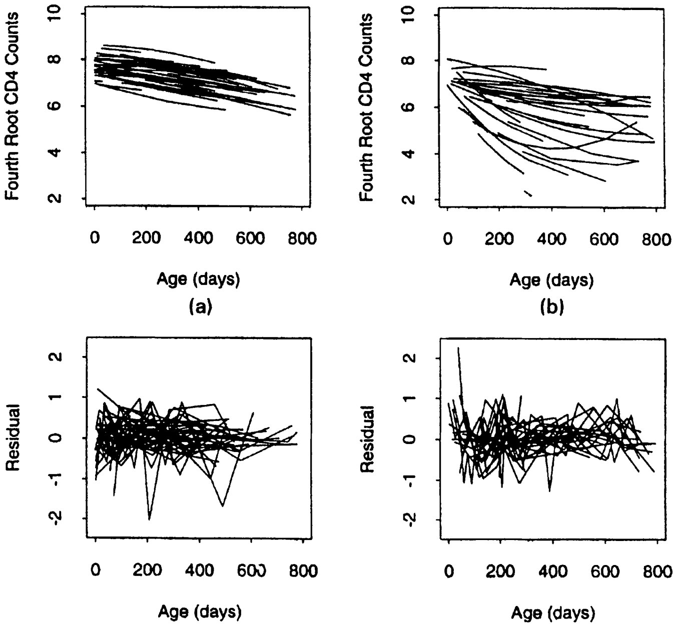

Fig. 2. Random quadratic model: (a), (b) fitted curves for individuals; (c), (d) residuals; (a), (c) uninfected group; (b), (d) infected group

time. Additionally, the residuals for the infected infants are strictly positive from birth to age 1 month, suggesting that this model is not adequate.

In recent work, model (1) has been extended by allowing the population mean $f(t)$ to take general forms (Zeger and Diggle, 1994; Wang and Taylor, 1995). But the random effects and covariance structures are still assumed to have special parametric forms which may not properly model the data. For example, Zeger and Diggle (1994) considered $s_i(t)$ as realizations of a stationary stochastic process. However, for these CD4 data and others (Taylor *et al.*, 1994) the assumption of stationarity does not seem reasonable. Rice and Silverman (1991) developed a nonparametric curve estimation method by modelling each individual curve by a smooth mean function plus the sum of a small number of eigenfunctions with random amplitudes. But their methodology cannot be applied to irregularly spaced longitudinal data. In this paper, we develop and evaluate a flexible class of semiparametric statistical models for longitudinal data on the basis of *B*-splines (de Boor, 1978). The point of view that we take is to let the data show us the appropriate functional form of both the fixed and the random effects. The model will be defined in the next section. Estimation and computation are discussed in Section 3. The CD4 data are analysed in Section 4. We conclude with a discussion.

2. Flexible Semiparametric Model

In this section, we propose a flexible class of semiparametric models for modelling unbalanced repeated measures data by expressing the functions $f(t)$ and $s_i(t)$ as cubic splines. Cubic splines are smooth curves with continuous first and second derivatives and piecewise continuous third derivatives that are discontinuous at certain specified locations called knots. We use *B*-splines to parameterize our splines since the *B*-spline basis functions are numerically superior to other alternative bases such as the truncated power series. The *B*-spline basis functions consist of a set of overlapping, smooth, non-negative, unimodal functions, which sum to 1 at every value of t. The *B*-splines are nearly orthogonal since any given basis function is non-zero only over a span of at most five distinct knots. The algebraic definition of *B*-splines is opaque and the interested reader should see de Boor (1978).

Let the $J \times 1$ vector $B(t)$ be a cubic *B*-spline basis evaluated at time t. We assume that the functions in model (1) have the form $f(t) = \alpha^{\mathrm{T}} B(t)$ and $s_i(t) = \gamma_i^{\mathrm{T}} B(t)$ and consider the model

$$y_i(t) = \alpha^{\mathrm{T}} B(t) + \gamma_i^{\mathrm{T}} B(t) + \epsilon_{it}, \tag{2}$$

where $\alpha = (\alpha_k)$ is a vector of fixed effect coefficients which determines the shape of the population mean function. It is a property of *B*-splines that smooth coefficients α_k will produce smooth functions $f(t)$ and if the α_k are monotone, $\alpha_k < \alpha_{k+1}$, then the mean function $f(t)$ will be monotone (Kelly and Rice, 1990). Also, if all α_ks are equal, then $f(t)$ is a constant. The random quadratic effects model is a special case of our model since a polynomial of degree 2 is in the span of the *B*-spline basis. The vector γ_i is a vector of random effect coefficients associated with child i. Let $\{\gamma_i\} = \{\gamma_i;\ i = 1, \ldots, n\}$; the $\{\gamma_i\}$ are assumed mutually uncorrelated random vectors with a common distribution $\gamma_i \sim N_J(0, D)$.

In model (2), both the function $f(t)$ and the random effects $s_i(t)$ are linear combinations of the *B*-spline basis. This allows the average trajectory $f(t)$ to be a general smooth function and allows each child to have their own smooth trajectory $f(t) + s_i(t)$ and does not force all trajectories to be similar, i.e. from the same small parametric family. This is very desirable in many applictions and, like all random effects models, it permits the investigator to examine both the population trajectory as well as the child trajectories. In contrast, the linear or low order polynomial model forces the trajectories to be linear or polynomial which can be very unrealistic.

This flexible modelling structure is likely to be overparameterized. Model selection can be used to reduce the number of parameters involved. The covariance matrix D may be degenerate or nearly degenerate or it may have some simple structure. Although we can also do model selection on the fixed effects, the focus here is to simplify model (2) through model selection on the covariance structure D. We use a principal components decomposition of D to reduce the number of covariance parameters while accounting for most of the variance of $y_i(t)$ explained by the random effects. Let the singular value decomposition of D be

$$D = \Delta \Lambda \Delta^{\mathrm{T}},$$

where Δ is an orthogonal matrix consisting of the eigenvectors Δ_j of D. The matrix $\Lambda = \mathrm{diag}(\lambda_j)$ is a diagonal matrix and λ_j is the jth largest eigenvalue of D. The linear combinations $C_j(t) = \lambda_j^{1/2}\, \Delta_j^{\mathrm{T}}\, B(t)$ are a transformation of the basis functions $B(t)$ and are called the transformed B-splines. With this notation, we write

$$y_i(t) = \alpha^{\mathrm{T}}\, B(t) + \sum_{j=1}^{J} \delta_{ij}\, C_j(t) + \epsilon_{it} \tag{3}$$

where the $\delta_{ij} = \gamma_i^{\mathrm{T}} \Delta_j \lambda_j^{-1/2}$ are independent and identically distributed $N(0, 1)$ random variables. Most of the variability of the $y_i(t)$ will be accounted for by the first few terms of $\{\delta_{ij}\, C_j(t)\}$. Hence model (3) can be rewritten as

$$\begin{aligned} y_i(t) &= \alpha^{\mathrm{T}}\, B(t) + \sum_{j=1}^{K} \delta_{ij}\, C_j(t) + \sum_{j=K+1}^{J} \delta_{ij}\, C_j(t) + \epsilon_{it} \\ &\approx \alpha^{\mathrm{T}}\, B(t) + \sum_{j=1}^{K} \delta_{ij}\, C_j(t) + \epsilon_{it} \end{aligned} \tag{4}$$

for some $K \ll J$. This model has the simplified random effect portion

$$s_{Ki}(t) = \sum_{j=1}^{K} \delta_{ij}\, C_j(t);$$

the usefulness of the terms $s_{Ki}(t)$ may be summarized in terms of the variance explained.

3. Estimation and Model Selection

Suppose that the ith child is observed at times $t_{i1}, \ldots, t_{in_i}$ and let $Y_i = (y_i(t_{i1}), \ldots, y_i(t_{in_i}))^{\mathrm{T}}$ denote the observed fourth-root CD4 counts. For the observed data, conditionally on K and after rotation, model (4) has the form

$$Y_i = X_i\alpha + Z_i b_i + e_i, \tag{5}$$

where X_i is an $n_i \times J$ matrix with jth column the jth B-spline evaluated at times $t_i = (t_{i1}, t_{i2}, \ldots, t_{in_i})^{\mathrm{T}}$. The matrix Z_i is an $n_i \times K$ matrix with jth column the jth transformed B-spline $C_j(t)$ evaluated at the times t_i. Both matrices X_i and Z_i depend on i through t_i. The $\{b_i\}$ are individual effects, assumed independent and identically normally distributed with mean 0 and unknown covariance matrix Ω_K. The $\{e_i\}$ are independent of $\{b_i\}$ and are distributed $N(0, \sigma^2 I_{n_i})$, where I_{n_i} is an $n_i \times n_i$ identity matrix. Model (5) has the form of a typical random effects model. Both X_i and Z_i depend on the number of knots and their locations and the number of random effects, all of which need to be selected by the analyst.

The choice of the number of knots is a bias–variance trade-off. Many knots will result in estimates with small bias but large variance, and few knots will give larger bias and low variance. In practice we try several sets of knots and choose the one which gives the 'best' results. On the basis of our experience, a model with 5–10 knots will often give satisfactory results. The knots should be placed at time points where many data are collected; more knots should be put in places where there is more curvature in the response.

Several specific methods are available for choosing J and K. A maximum likelihood approach is to select J and K on the basis of likelihood ratio tests. Even if a formal test is not done, the relative values of the log-likelihood will aid in determining J and K as shown in the analysis of the CD4 counts. A second approach is cross-validation. The cross-validation score is defined as

$$\mathrm{CV}(J, K) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\{y_i(t_{ij}) - \hat{y}_{JKi}^{(-ij)}(t_{ij})\}^2,$$

where $N = \Sigma_{i=1}^{n} n_i$ and $\hat{y}_{JKi}^{(-ij)}(t_{ij})$ is the predicted value for the ith person at time t_{ij} by fitting model (4) with J knots and K transformed B-splines to the data with observation $y_i(t_{ij})$ deleted. Fitted values include both the fixed and the random components defined as $\hat{y}_i(t) = \hat{f}(t) + \hat{s}_i(t)$. The J and K with the smallest cross-validation score can be selected. A third approach is to choose J and K by checking the contribution of $C_j(t)$ to the reduction of the estimates of the variance σ^2. The selection process is stopped if the additional knot or term does not reduce the variance by a reasonable amount. Finally, for given J, the relative percentage of the variation explained by the random effects $(100R^2)$ can also be used to select K, where

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\{y_i(t_{ij}) - \hat{y}_{Ki}(t_{ij})\}^2}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\{y_i(t_{ij}) - \hat{f}(t_{ij})\}^2}.$$

In this expression $\hat{f}(t) = \hat{\alpha}^{\mathrm{T}} B(t)$ is estimated by using model (4) with $K = 0$ and $\hat{y}_{Ki}(t)$ is the fitted curve of the ith subject by model (4) with K transformed B-splines. The value of K may be chosen on the basis of the contributions of $C_K(t)$ to R^2.

The parameters and random effects can be estimated by maximum likelihood methods, restricted maximum likelihood methods or fully Bayesian methods. For fixed J there are two steps in our estimation procedures for the model proposed. The first step involves fitting model (5) with

$$Z_i = X_i = (B(t_{i1}), \ldots, B(t_{in_i}))^{\mathrm{T}}.$$

In this case, D is the $J \times J$ covariance matrix of the random coefficients $\{b_i\}$ and can be estimated by maximum likelihood. The EM algorithm and computing formulae (3.1)–(3.7) of Laird *et al.* (1987) are used to find estimates of D in the analysis in Section 4. The eigenvalues and eigenvectors of the estimated matrix D are then combined with the B-spline basis to form the $C_j(t)$. These transformed B-splines are fixed in our further estimation.

In the second step, we do model selection by using the transformed B-splines. For fixed $K = 1, 2, \ldots$, we refit model (5) and calculate cross-validation scores, the log-likelihood, $\hat{\sigma}^2$ and R^2. The log-likelihood is equal to

$$-\frac{1}{2} \sum_{i=1}^{n} \{n_i \log(2\pi) + \log|\hat{\Sigma}_i| + (y_i - X_i\hat{\alpha})^{\mathrm{T}} \hat{\Sigma}_i^{-1} (y_i - X_i\hat{\alpha})\},$$

where $\hat{\Sigma}_i = \hat{\sigma}^2 I + Z_i \hat{\Omega}_K Z_i^{\mathrm{T}}$. Again the EM program of Laird *et al.* (1987) is used to compute all estimates. The number of random effect terms (K) is chosen by the model selection methods outlined earlier. We then analyse the selected model conditionally on K. The parameters α, Ω_k, σ^2 and the random coefficients $\{b_i\}$ are all estimated. Because of the definition of $\{C_j(t)\}$, the estimated matrix Ω_K in this second stage will be very similar to the identity matrix. The estimated population mean function is $\hat{f}(t) = \hat{\alpha}^{\mathrm{T}} B(t)$, the child trajectories are

$$\hat{f}(t) + \hat{s}_i(t) = \hat{\alpha}^{\mathrm{T}} B(t) + \sum_{j=1}^{K} \hat{b}_{ij} C_j(t),$$

$i = 1, 2, \ldots, n$, and the residuals are $y_i(t) - \hat{f}(t) - \hat{s}_i(t)$ (Weiss and Lazaro, 1992). The maximum likelihood or restricted maximum likelihood estimators can also be computed by Newton–Raphson iteration (Jennrich and Schluchter, 1986) or by the EM algorithm (Laird and Ware, 1982). Bayesian posteriors can be calculated by using the Gibbs sampler (Gelfand *et al.*, 1990).

4. Analysis of Acquired Immune Deficiency Syndrome Data

In this section, we continue our analysis of the CD4 counts data set with our proposed method. We first choose the knots of the B-spline basis. For each group, we consider four sets of knots with the number of knots (J) equal to 3, 5, 7 and 10, with the locations at the percentiles of the distribution of $\{t_{ij}\}$. For these data we found that, for the same number of knots, the model with knots at the percentiles of $\{t_{ij}\}$ fit the data better than the model with knots equally spaced in time, both in terms of likelihood and in terms of σ^2. Models with different values of K as well as different sets of knots were fitted to the data and compared. Preliminary analyses suggested that $K = 2$ random effects should provide a satisfactory fit to the data for both

TABLE 1
Knot selection criteria for the infected group

J (K = 2)	*Cross-validation scores*	*Log-likelihood*	$\hat{\sigma}^2$
3	0.413	−275.76	0.388
5	0.416	−268.81	0.320
7	0.325	−241.83	0.215
10	0.371	−246.68	0.218

groups. For each set of knots and $K = 2$, we look at the cross-validation score, the log-likelihood and $\hat{\sigma}^2$. On the basis of the results in Table 1, seven knots were selected for the infected group; they are $t = 0, 113, 189, 264, 398, 563$ and 794 days. Using a similar method, five knots at 1, 71, 205, 350 and 778 days were selected for the uninfected group.

Model (4) was fitted to the transformed CD4 counts for each group. The leading two transformed *B*-splines ($C_j(t), j = 1, 2$) for each group are shown in Fig. 3. The full curves indicate the leading random effect term. For the uninfected group this

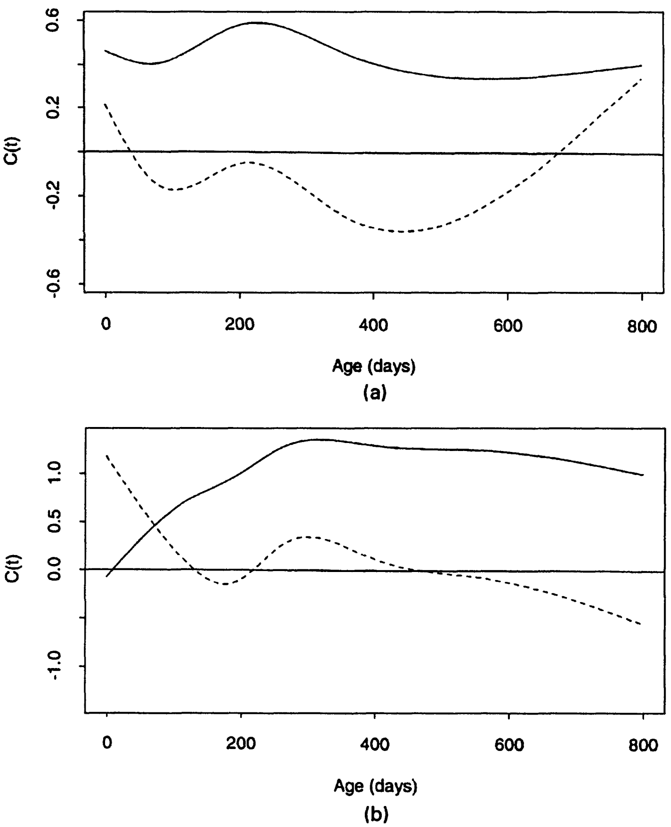

Fig. 3. Leading transformed *B*-spline (———) and second transformed *B*-spline (- - - -) plotted against age for (a) the uninfected group and (b) the infected group

TABLE 2
Model selection criteria

K	*Cross-validation scores*	*Log-likelihood*	$\hat{\sigma}^2$	$100R^2$
Uninfected group, $J=5$				
0	0.520	−272.42	0.465	0
1	0.324	−242.19	0.268	51.6
2	0.319	−239.71	0.240	59.6
3	0.318	−238.57	0.225	63.8
Infected group, $J=7$				
0	1.32	−375.74	1.251	0
1	0.393	−262.06	0.318	79.2
2	0.325	−241.83	0.215	87.5
3	0.323	−239.40	0.198	88.9

curve is approximately constant, which could be interpreted as a random intercept model. To determine the number of terms in the random effects, we compared the cross-validation scores, the log-likelihoods, $\hat{\sigma}^2$ and R^2 as shown in Table 2. The model with $K=2$ is selected for both groups.

The fitted fourth-root CD4 counts and residuals for both groups are presented in Fig. 4. Compared with the random quadratic model (Fig. 2), our model fits the data well, as shown in both the fitted curves and in the residual plots, especially for the infected group near age 0. The two sets of individual curves are quite similar for the

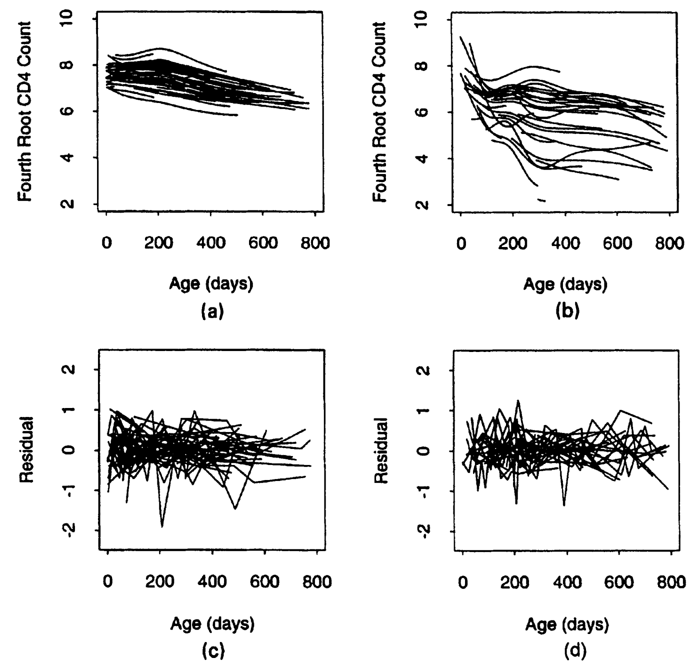

Fig. 4. Results from the *B*-spline model: (a), (b) fitted curves for individuals; (c), (d) residuals; (a), (c) uninfected group; (b), (d) infected group

uninfected group but are noticeably different for the infected group. Our model has larger likelihood and smaller σ^2 than the random quadratic model. For the random quadratic model, the log-likelihood and σ^2 are -242.7 and 0.243 for the uninfected group and are -263.2 and 0.278 for the infected group. The random quadratic model has three random coefficients per person instead of our more parsimonious choice of $K=2$.

The estimated population trajectories for both groups are displayed in Fig. 5 in the untransformed scale, along with confidence bands and the observed CD4 counts. The 95% pointwise confidence bands are defined by

$$\hat{f}(t) - 1.96\,\sigma_f(t) \leqslant f(t) \leqslant \hat{f}(t) + 1.96\,\sigma_f(t),$$

where

$$\sigma_f^2(t) = B^{\mathrm{T}}(t)\left\{\sum_{i=1}^{n} X_i^{\mathrm{T}}(\hat{\sigma}^2 I + Z_i\hat{\Omega}_K Z_i^{\mathrm{T}})^{-1}X_i\right\}^{-1} B(t).$$

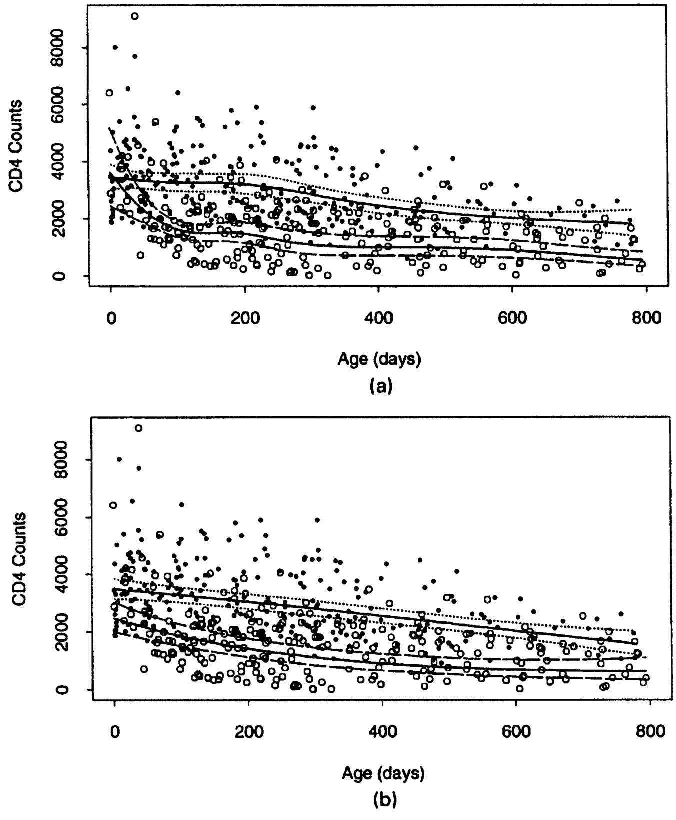

Fig. 5. CD4 measurements based on (a) the B-spline approach and (b) the random quadratic model (●, uninfected group; ○, infected group; ——, estimated mean trajectories (back transformed) of counts (upper curves for the uninfected group; lower curves for the infected group); ·········, – – – –, 95% confidence bands for the population means)

The estimates show that the two groups are clearly different with the uninfected group generally maintaining a higher population mean CD4 T-cell count than the infected group. The differences are 1000 or more for children over 4 months old. The CD4 counts in both groups have a continual decline, with a rapid decline for very young infected infants. This fact may conflict with the findings of de Martino *et al.* (1991), who found that for symptom-free children both the uninfected and the infected groups have higher CD4 counts at age 1–6 months than at age 0–1 month.

For comparison, in Fig. 5, we also present the population means and confidence bands estimated from the random quadratic model. The fitted mean CD4 counts of the infected group fails to model the rapid decline from birth to age 3 months and furthermore this quadratic model shows a slight increase from age 700 to 800 days.

Another aim of our analysis is to provide reference centile curves based on repeatedly measured data. Centile charts are commonly used in preliminary medical diagnoses to establish whether some measure of interest on an individual lies within a typical range. The general form of such a chart is a series of smoothed curves, showing how selected percentiles for a measurement change as a function of time. There are several methods for fitting centile curves (Cole and Green, 1992; Thompson and Theron, 1990; Efron, 1991; Aldhous *et al.*, 1994). Most of those references assume independent observations or ignore the repeated measures nature of the data. Using our proposed model, we can estimate smooth reference centile curves from unbalanced and irregular longitudinal data, incorporating the correlation structure into the estimation procedure.

From model (4), we have that for any fixed time t the $100\nu\%$ centile for $y(t)$ is estimated by

$$y_\nu(t) = \hat{\alpha}^{\mathrm{T}} B(t) + z_\nu\{C^{\mathrm{T}}(t)\hat{\Omega}_K C(t) + \hat{\sigma}^2\}^{1/2}, \tag{6}$$

where $C(t) = (C_1(t), C_2(t), \ldots, C_K(t))^{\mathrm{T}}$ and z_ν is the 100ν-percentile of the standard normal distribution. We fitted the centile curves to the CD4 counts for the uninfected group. The estimated $\{3, 5, 10, 25, 50, 75, 90, 95, 97\}$ centiles curves are shown in Fig. 6. Our estimated centile curves are similar to the CD4 centile curves published by Wade *et al.* (1992).

5. Discussion

We have presented and illustrated a fairly simple, flexible, semiparametric method for the analysis of unbalanced longitudinal data. The estimation and implementation of the computations are simple since the expression of model (5) enables us to use existing algorithms to compute the maximum likelihood estimators of the parameters. In the analysis of AIDS data, the EM algorithm of Laird *et al.* (1987) was programmed in S-PLUS (Becker *et al.*, 1988) to compute the estimates. We used the same *B*-spline basis in model (2) for the fixed and random effects, but we could in principle use a different basis for each of the fixed effects and random effects. For example, we might use a basis with more knots for the fixed effects than for the random effects.

In our approach we have assumed that $s_i(t)$ can be described by a smooth curve, parameterized by using *B*-splines. There are alternative approaches. One of the referees preferred to view $s_i(t)$ as a realization of a specific stochastic process. For

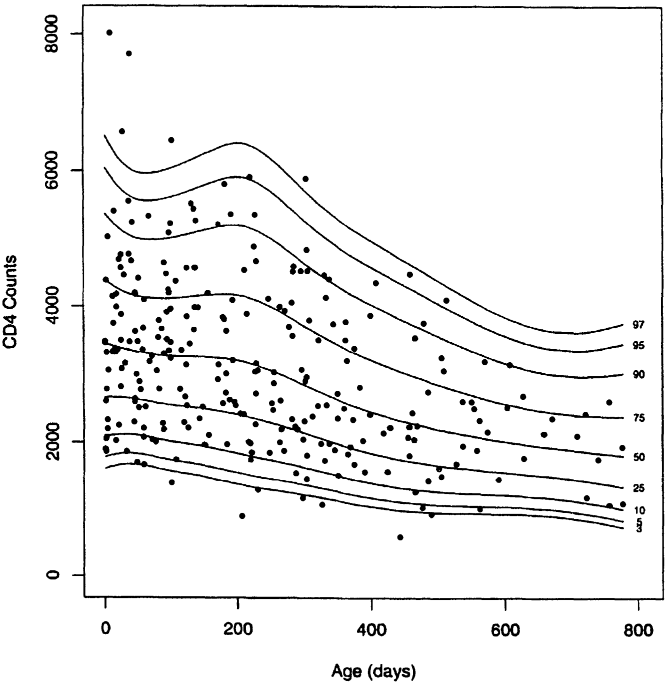

Fig. 6. Estimated centile curves for the uninfected group: ●, individual CD4 measurements; ——, back-transformed estimated reference centile curves

example, Zeger and Diggle (1994) and Diggle (1988) used adaptations of AR(1) processes to model the covariance structure of the observations. Both these processes are stationary, whereas our model allows the variance structure to change with time, something which we have found desirable for the CD4 count data. Non-stationary parametric stochastic processes are possible; for example Wang and Taylor (1995) used a different adaptation of an AR(1) process, and Taylor *et al.* (1994) assumed Brownian motion or an integrated Ornstein–Uhlenbeck process. In essence, all these assumptions can just be viewed as different covariance structures for the observations.

In this paper we have emphasized the construction of smooth individual trajectories; such trajectories could then be used as input to further analysis. For example, Tsiatis *et al.* (1995) modelled individual CD4 counts by using a linear random effects model and then used the individual curves as predictors in a further survival analysis. The individual trajectories estimated by our approach could also be used as input in their survival analysis. Finally, the methodology developed in this paper can be generalized without technical difficulty to analyse models with covariates and different treatment groups.

Acknowledgements

The authors thank Susan Plaeger and Diana Liao for making the AIDS data available. This work was supported by an academic senate grant from the University of California at Los Angeles and by National Institutes of Health grants AI 07370, GM50011 and AI 29196.

References

Aldhous, M. C., Raab, G. M., Doherty, K. V., Mok, J. Y. Q., Bird, A. G. and Froebel, K. S. (1994) Age-related ranges of memory, activation, and cytotoxic markers on CD4 and CD8 cells in children. *J. Clin. Immunol.*, **14**, no. 5, 1–10.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Pacific Grove: Wadsworth.

de Boor, C. (1978) *A Practical Guide to Splines*. New York: Springer.

Cole, T. J. and Green, P. J. (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. *Statist. Med.*, **11**, 1305–1319.

Diggle, P. J. (1988) An approach to the analysis of repeated measurements. *Biometrics*, **44**, 959–971.

Efron, B. (1991) Regression percentiles using asymmetric squared error loss. *Statist. Sin.*, **1**, 93–125.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990) Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Am. Statist. Ass.*, **85**, 972–985.

Jennrich, R. I. and Schluchter, M. (1986) Unbalanced repeated measures models with structured covariance matrices. *Biometrics*, **42**, 805–820.

Kelly, C. and Rice, J. (1990) Monotone smoothing with application to dose–response curves and the assessment of synergism. *Biometrics*, **46**, 1071–1085.

Koup, R. A. and Wilson, C. B. (1994) Clinical immunology of HIV-infected children. In *Pediatric AIDS* (eds P. A. Pizzo and C. M. Wilfert), 2nd edn. Baltimore: Williams and Wilkins.

Laird, N., Lange, N. and Stram, D. (1987) Maximum likelihood computations with repeated measures: application of the EM algorithm. *J. Am. Statist. Ass.*, **82**, 97–105.

Laird, N. M. and Ware, J. H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

de Martino, M., Tovo, P. A., Galli, L., Gabiano, C., Cozzani, S., Gotta, C., Scarlatti, G., Fiocchi, A., Cocchi, P., Marchisio, P., Canino, R., Mautone, A., Chiappe, F., Campelli, A., Consolini, R., Izzi, G., Laverda, A., Alberti, S., Tozzi, A. E. and Duse, M. (1991) Prognostic significance of immunologic changes in 675 infants perinatally exposed to human immunodeficiency virus. *J. Pediatr.*, **119**, 702–709.

Rice, J. A. and Silverman, B. W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc.* B, **53**, 233–243.

Taylor, J. M. G., Cumberland, W. G. and Sy, J. P. (1994) A stochastic model for analysis of longitudinal AIDS data. *J. Am. Statist. Ass.*, **89**, 727–736.

Thompson, M. L. and Theron, G. B. (1990) Maximum likelihood estimation of reference centiles. *Statist. Med.*, **9**, 539–548.

Tsiatis, A. A., DeGruttola, V. and Wulfsohn, M. S. (1995) Modeling the relationship of survival to longitudinal data measured with error: applications to survival and CD4 counts in patients with AIDS. *J. Am. Statist. Ass.*, **90**, 27–37.

Wade, A. M., Ades, A. E., Dunn, D. T., Newell, M.-L., Peckham, C. S. and Demaria, A. (1992) Age-related standards for T lymphocyte subsets based on uninfected children born to human immunodeficiency virus 1-infected women. *Pediatr. Infect. Dis. J.*, **11**, 1018–1026.

Wang, Y. and Taylor, J. M. G. (1995) Inference for smooth curves in longitudinal data with application to an AIDS clinical trial. *Statist. Med.*, **14**, 1205–1218.

Weiss, R. E. and Lazaro, C. (1992) Residual plots for repeated measures. *Statist. Med.*, **11**, 115–124.

Zeger, S. L. and Diggle, P. J. (1994) Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, **50**, 689–699.