Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Modeling Disease Marker Processes in AIDS

Yudi Pawitan & Steve Self

To cite this article: Yudi Pawitan & Steve Self (1993) Modeling Disease Marker Processes in AIDS, Journal of the American Statistical Association, 88:423, 719-726, DOI: 10.1080/01621459.1993.10476332

To link to this article: https://doi.org/10.1080/01621459.1993.10476332

Published online: 27 Feb 2012.

Modeling Disease Marker Processes in AIDS

YUDI PAWITAN and STEVE SELF*

The importance of disease markers in understanding the progression of acquired immune deficiency syndrome (AIDS) and devising treatment strategies is well recognized. This issue is usually addressed using cross-sectional data analysis, which tends to ignore the longitudinal data collected on the individuals. Available longitudinal data for nontransfusion-related AIDS raise some technical challenges to standard longitudinal analyses due to left and right censoring as well as left truncation. We describe a likelihood method to model the disease markers as a function of time by modeling the joint distribution of the markers, the time of infection, and the time to AIDS. We address the problems of censoring and truncation using standard survival analysis techniques. We also consider the prediction of time to AIDS given a series of disease marker measurements. An illustrative example, using data from the Toronto AIDS cohort study, is given. In particular, the analysis shows that the slope of the decline in T4 cell count measurements or T4/T8 ratio is associated with the time to AIDS. We compare the prediction of the time to AIDS for an individual with or without a series of T4/T8 measurements and with a known or unknown infection time.

KEY WORDS: HIV; Likelihood estimation; Prediction; Random effects; Regression; Survival analysis.

A number of serologic and cellular markers have been shown to be associated with the development of acquired immune deficiency syndrome (AIDS) among individuals infected with human immunodeficiency virus (HIV) (see, for example, Fahey et al. 1990 and Schecter et al. 1989). The method used to show the association is typically cross-sectional, where the measurements at entry to the study or at the time of infection are correlated with the time to AIDS. Strictly speaking, the method does not really tell if, for example, a decline of T4 cell count observed *within* an individual is associated with the time to AIDS. To illustrate a limitation of this method, for a seropositive individual who has a series of T4 cell count measurements (with the time of infection known or not known), what is the predicted time to AIDS? The last value of the T4 measurements will probably be used for prediction. To use all values or some combination of values in the prediction, we need a model that includes those values as predictors of time to AIDS. Developing such a model is hard in general, because the times as well as the number of T4 count measurements vary considerably between individuals. A longitudinal study of the markers can help answer the question and in general is important for understanding the pathogenesis and natural history of the disease. On the practical level, it also contributes in the development of treatment strategies.

As was reviewed by Jewell (1990), most longitudinal data on nontransfusion-related AIDS come from follow-up studies of prevalent cohorts, groups of patients who were infected prior to entry to the study; for example, the Multicenter AIDS Cohort Study. Technical problems in analysis arise due to left censoring of the infection time and right censoring of the disease occurrence. This problem is only partially overcome, for example, by including only cases for which the infection times are observed, so the problem reduces to right censoring. The latter was considered, for example, by Wu and Bailey (1988) and Wu and Carroll (1988), who discussed also the bias and efficiency of least squares procedures when the censoring is informative. Our proposed method can accommodate both censoring due to events, which is likely to be informative, and censoring due to end of follow-up, which is not likely to be informative. It is also applicable when the relative time zero of the process—in this case, the infection time—is left censored. DeGruttola, Lange, and Dafni (1991) analyzed the longitudinal T4 cell count data from the San Francisco cohort study using the growth curve model with random effects. They treated the problem of left censoring by introducing a measurement error model for the unknown infection times, but they did not consider the problem of informative right censoring.

We describe here an approach to model disease marker processes as a function of time. This approach is based on a growth curve model with random intercepts and slopes, as in DeGruttola et al. (1991), but we use standard survival analysis techniques to deal with the censoring and truncation problem. The likelihood functions from different censoring patterns are presented in Section 1. The model accommodates the fact that, for example, the path of a disease marker is associated with the time to AIDS. In the linear model example that we give in Section 2, both the intercept and slope terms are a function of time to AIDS. In this connection we should mention Eyster, Gail, Ballard, Al-Mondhiry, and Goedert (1987), who reported that among seropositive hemophiliacs the slope of the T4 cell counts over time was strongly associated with the time to AIDS. Our approach also allows a straightforward prediction of the time to AIDS given a time series of a disease marker. This is done by estimating the conditional distribution of the time to AIDS, given the series of measurements and the (possibly censored) information about infection time. In one illustrative example in Section 2, this prediction is computed with and without assuming the series of T4/T8 ratio measurements.

1. DATA AND MODELS

A semi-ideal data set for this problem is shown in Figure 1(a); it is semi-ideal in the sense that the times of infection were known (to within a few months). This plot represents

* Yudi Pawitan is Lecturer, Department of Statistics, University College, Dublin 4, Ireland. Steve Self is a member of the Biostatistics Department, Fred Hutchinson Cancer Research Center, Seattle, WA 98104. This research was partially supported by National Institutes of Health Grant 5RO1AI29168. The authors thank V. Farewell and R. Gentleman and the Toronto AIDS Cohort Study for allowing the use of the data in this article and the referees for excellent suggestions, particularly those that led to consideration of the random effects model and Model 3.

longitudinal T4 counts data of 16 cohort members of the Toronto AIDS cohort who seroconverted during the follow-up period. As an illustration of T4 paths over time, the data from three cohort members are represented by lines, with each line representing the repeated measurements of T4 counts from an individual. The line starts at the time of infection and ends at the time of AIDS (for one case marked with "+") or last follow-up (for the other). The three cases shown are those with minimum, median, and maximum path averages, where the path average was computed as the average of all follow-up measurements from an individual.

In the whole data set, the time of infection is largely unknown. Figure 1(b) shows the follow-up measurements of 143 individuals who were seropositive at the time of entry to the study. Three cases with minimum, median, and maximum path averages are shown. Figures 1(c) and (d) show a similar information for the T4/T8 ratio measurements. In all plots we note a certain amount of "tracking"; that is, an individual series that starts with a high value tends to remain high over time and vice versa.

To fix ideas, let t_{Ii} and t_{Di} denote infection time (relative to the beginning of the epidemic) and disease occurrence (relative to the infection time). Throughout, the subscript i corresponds to the ith individual in the cohort; B is the beginning of the ascertainment period, which is used to account for the effect of left truncation; b_i and e_i are the entry and last follow-up of the ith individual in the study; and $Z(t)$ is the marker of interest, where $t = 0$ corresponds to the beginning of the epidemic. Suppose that $Z_i(t)$ is longitudinally observed at times $s_1, \ldots, s_{m_i}$, which may vary between individuals; let $\mathbf{Z}_i$ be the vector $(Z_i(s_1), \ldots, Z_i(s_{m_i}))'$. $\mathbf{X}_{1i}$ and $\mathbf{X}_{2i}$ are vectors of other covariates; assume that these covariates are not time dependent.

Let $f_I(\cdot\,|\mathbf{X}_1, \theta_1)$ and $f_D(\cdot\,|\mathbf{X}_2, \theta_2)$ be the probability density functions of the infection times and the disease occurrence. The covariate $\mathbf{X}_1$ modifies the risk of infection, whereas $\mathbf{X}_2$ affects the occurrence of the disease after infection, there may be some overlap between $\mathbf{X}_1$ and $\mathbf{X}_2$. Joint estimation of θ_1 and θ_2 was considered by Brookmeyer and Goedert (1989). Given t_I and t_D, $\mathbf{Z}$ is modeled by a probability density

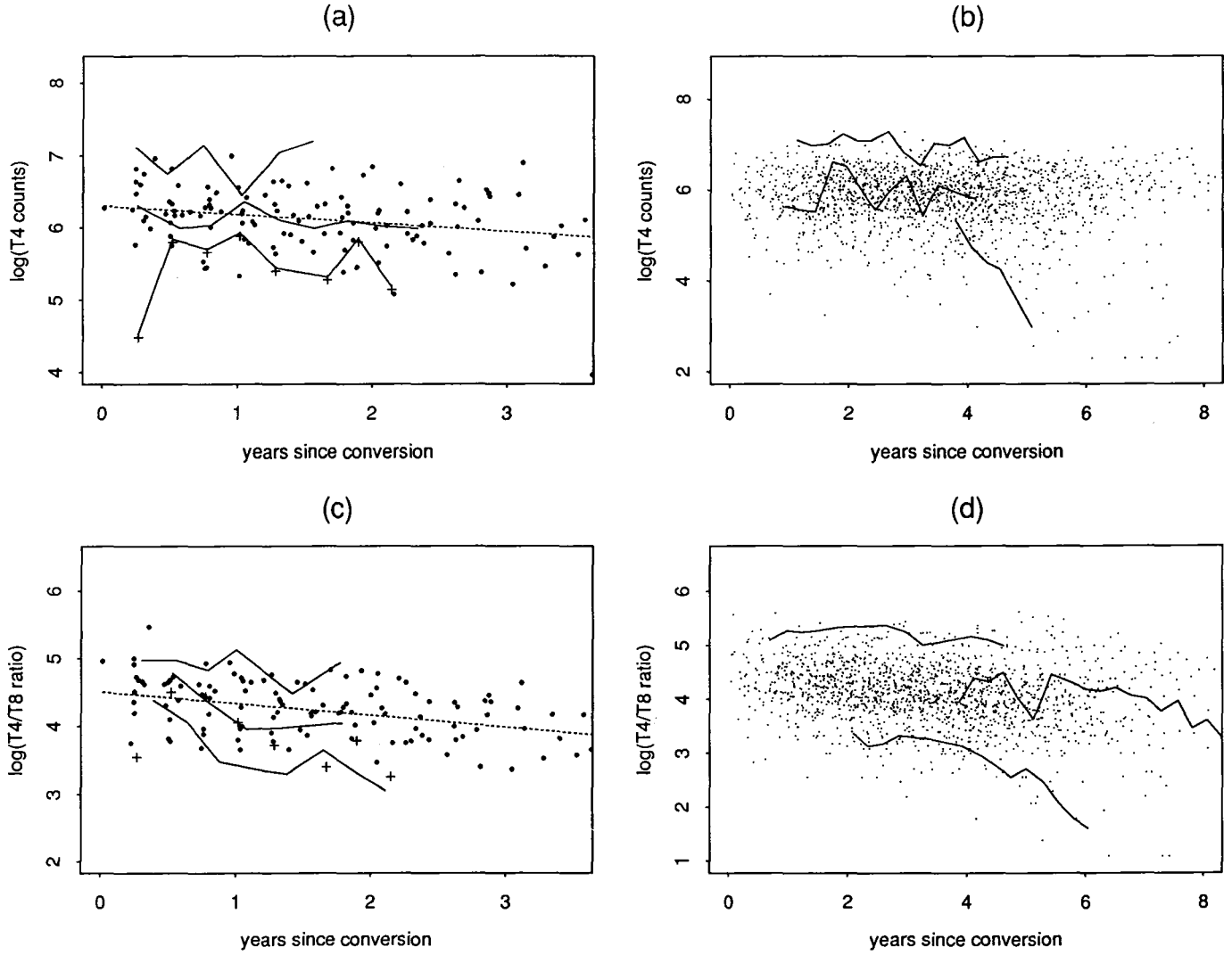

Figure 1. Time Series of T4 Count and T4/T8 Ratio Measurements. The three lines in each plot are the paths of those individuals with minimum, median, and maximum path averages. Plot (a) shows the T4 count measurements from 16 individuals with known infection times. The dotted line is the ordinary least squares fit of data. The person with "+" developed AIDS during follow-up. Plot (b) shows the T4 count measurements from 143 individuals in prevalent subcohort. The infection time is estimated as the midpoint between the first and last contact with the primary cases. Plots (c) and (d) are the same as (a) and (b), for T4/T8 ratio.

function $\phi(\cdot \mid t_I, t_D, \theta_3)$. Let $\theta \equiv (\theta_1, \theta_2, \theta_3)$ be the vector of unknown parameters and assume that f_I, f_D, and ϕ are some known parametric models. The basic likelihood component of a person with observed t_I and t_D is

$$L(\theta \mid t_I, t_D) = f_I(t_I \mid \mathbf{X}_1, \theta_1) f_D(t_D \mid \mathbf{X}_2, \theta_2) \phi(\mathbf{Z} \mid t_I, t_D, \theta_3), \quad (1)$$

where we have assumed that, conditional on $\mathbf{X}_1$ and $\mathbf{X}_2$, t_I and t_D are independent. Joint estimation of θ_1, θ_2, and θ_3 is necessary due to left and right censoring.

1.1 Modeling f_I and f_D

For the infection time distribution, we consider the Weibull regression model with the log survival function given by

$$\log(F_I(t \mid \mathbf{X}_1)) = -(\lambda_{0I} t e^{\delta_I' \mathbf{X}_1})^{\alpha_I}, \quad (2)$$

where α_I is the shape parameter, which is equal to 1 for the exponential case, and δ_I is a vector of regression parameters. Similarly, for the time of disease occurrence, we model

$$\log(F_D(t \mid \mathbf{X}_2)) = -(\lambda_{0D} t e^{\delta_D' \mathbf{X}_2})^{\alpha_D}. \quad (3)$$

1.2 Modeling $Z(t)$

In general, we can formulate a generalized linear model for the longitudinal data $\mathbf{Z}$ (see McCullagh and Nelder 1989). But a full likelihood specification is necessary whenever the infection time or the time to AIDS are censored. This limits the types of models that can be used for dependent non-Gaussian outcomes. Under the common assumption that, conditional on the individual, the measurements are independent, it is straightforward to use, for example, the binomial or Poisson likelihoods. The mean function for $Z_i(t)$, conditional on t_{Ii} and t_{Di}, is modeled as

$$\mu_i(t) \equiv \mu_i(t \mid t_{Ii}, t_{Di}, \theta_3) = h(\eta_i(t \mid t_{Ii}, t_{Di}, \theta_3)),$$

where h is the inverse link function and η_i is a linear model involving the covariates t_{Ii} and t_{Di}.

For our application, we describe in detail the case where there is a normalizing transformation g so that, denoting gZ to be the transform of Z,

$$gZ_i(t) = \gamma_{0i} + \gamma_{1i}(t - t_{Ii})^+ + e_i(t), \quad (4)$$

where γ_{0i} and γ_{1i} are *random* intercept and slope parameters, $x^+ = \max(0, x)$, and the error terms $e_i(t)$ are iid $N(0, \sigma_e^2)$ for all i and t, independent of γ_{0i} and γ_{1i} (cf. Laird and Ware 1982). The linear model is derived based on the assumption that the mean level of $gZ_i(t)$ is relatively constant prior to infection and then decreases linearly from that point on. Both the original level γ_{0i} and the slope γ_{1i} are potentially associated with the time to AIDS and are modeled as

$$\gamma_{0i} = \beta_0 + \beta_1 / t_{Di} + \pi_i \quad (5)$$

$$\gamma_{1i} = \beta_2 + \beta_3 / t_{Di}, \quad (6)$$

where β_0 to β_3 are *fixed* parameters and π_i is a random person effect, assumed to be iid $N(0, \sigma_\pi^2)$, independent of t_I, t_D, and $e_i(t)$. The transformation $1/t_{Di}$ scales the time axis in units of the time to AIDS. A transformation is necessary for the following reason. We expect that lower original level or steeper decline of $gZ(t)$ is associated with shorter time to AIDS; that is, a positive association. But a simple model—$\gamma_{1i} = \beta_2 + \beta_3 t_{Di}$, with $\beta_3 > 0$—will clash with the biologically plausible constraint of negative γ_{1i}. In the transformed model the slope is always negative whenever β_2 and β_3 are negative. Different transformations may achieve the same goal, but the proposed one has a meaningful interpretation.

These lead to an overall model

$$gZ_i(t) = \mu_i(t) + w_i(t),$$

$$\mu_i(t) = \beta_0 + \beta_1 / t_{Di} + \beta_2 (t - t_{Ii})^+ + \beta_3 (t - t_{Ii})^+ / t_{Di}, \quad (7)$$

where $(w_i(s_1), \ldots, w_i(s_{m_i}))$ is jointly normal with mean 0 and covariance matrix $\mathbf{\Sigma}_i$. The matrix $\mathbf{\Sigma}_i$ is an m_i by m_i exchangeable covariance matrix with $\sigma^2 = \sigma_\pi^2 + \sigma_e^2$ on the diagonal and σ_π^2 off the diagonal. We will express the latter parameter in terms of within-person correlation $\rho \equiv \sigma_\pi^2 / \sigma^2$, which measures the proportion of variability of $w_i(t)$ due to person effects. The model can be easily enlarged by including covariates that may affect the intercept or the slope or by adding terms to capture possible nonlinear patterns over time. We tried to fit, for example, the quadratic and piecewise linear models. In the model (7), β_1 reflects the association between the baseline level of $gZ(t)$ (at the time of, or prior to, infection) and the disease occurrence, β_2 is the rate of change of $gZ(t)$ after infection, and β_3 measures the interaction between the rate of change and disease occurrence. The term $(t - t_{Ii})^+ / t_{Di}$ can be interpreted as the scaled incubation time, because it takes value 0 at the time of infection and 1 at the time of AIDS. So β_3 is the average change of $gZ(t)$ by the time of AIDS occurrence.

Under (7), the likelihood contribution of the ith individual associated with the parameter θ_3 in (1) is proportional to

$$\det(\mathbf{\Sigma}_i)^{-1/2} \exp\{-(g\mathbf{Z}_i - \mu_i)' \mathbf{\Sigma}_i^{-1} (g\mathbf{Z}_i - \mu_i)/2\}. \quad (8)$$

1.3 Likelihood Formula for Censored Cases

We potentially observe the following patterns of censoring for the time of infection and the time to AIDS, each with a different likelihood formula:

a. $t_{Ii} < b_i$ and t_{Di} is observed

$$L_i(\theta) = \int_0^{b_i} L(\theta \mid t, t_{Di})\, dt;$$

b. $t_{Ii} < b_i$ and $t_{Ii} + t_{Di} > e_i$

$$L_i(\theta) = \int_0^{b_i} \int_{e_i - t}^{\infty} L(\theta \mid t, u)\, du\, dt;$$

c. t_{Ii} is observed and $t_{Ii} + t_{Di} > e_i$

$$L_i(\theta) = \int_{e_i - t_{Ii}}^{\infty} L(\theta \mid t_{Ii}, u)\, du;$$

d. $t_{Ii} > e_i$

$$L_i(\theta) = \int_{e_i}^{\infty} \int_0^{\infty} L(\theta \mid t, u)\, du\, dt.$$

The left censoring $t_{Ii} < b_i$ is basically an interval censoring $0 < t_{Ii} < b_i$. Modification for more general interval censoring

on t_{Ii} and t_{Di} is straightforward by taking an appropriate integral of $L(\theta \mid t_I, t_D)$. The total likelihood is then computed as

$$L(\theta) = \prod_i L_i(\theta), \tag{9}$$

where $L_i(\theta)$'s are the individual likelihood terms as defined previously.

1.4 Left Truncation

To account for individuals who develop the disease prior to the beginning of the study, we divide the individual likelihood term by the probability of observing the disease occurring after B, which is

$$1 - \int \int_0^B \int_0^{B-t} f_I(t; \mathbf{X}_1, \theta_1) f_D(u; \mathbf{X}_2, \theta_2)\, du\, dt\, dP(\mathbf{X}_1, \mathbf{X}_2).$$

The integral over $\mathbf{X}_1$, $\mathbf{X}_2$ is estimated by

$$\frac{1}{N} \sum_i \int_0^B \int_0^{B-t} f_I(t; \mathbf{X}_{1i}, \theta_1) f_D(u; \mathbf{X}_{2i}, \theta_2)\, du\, dt,$$

where the summation is over the individuals in the cohort.

1.5 Computation

Maximum likelihood estimates are obtained by maximizing (9) using the simplex algorithm, a derivative-free optimization technique that needs only a function evaluation. All integrations are evaluated numerically using a 10-point Gauss–Legendre quadrature formula (which will give an exact result for up to 19-degree polynomials). The integration on the infinite range is truncated at 25; this is unlikely to affect the results because, using the estimated parameter values, the probability of $t_D > 25$ is less than 5×10^{-8}. The standard errors of the estimates are computed from the inverse of the estimated information matrix, which is computed numerically. Further detail is given in the next section.

1.6 Prediction of the Time to AIDS

We now consider the prediction of the time to AIDS t_D given a time series vector $\mathbf{Z}$, observed at times $s_1, \ldots, s_m$, with known or unknown infection time t_I. It is worth noting that there are no straightforward regression techniques in this case, because the length of $Z(t)$ varies among individuals. The likelihood approach in this section suggests the following solution: Find the conditional density of t_D given $\mathbf{Z}$ and t_I under the *estimated* parameter values. When t_I is left censored, then it is integrated over the appropriate range. For example, given $t_I < b$ and a vector $\mathbf{Z}$, the conditional density of t_D is given by

$$\frac{\int_a^b L(\theta \mid u, t)\, du}{\int_{e-b}^{\infty} \int_a^b L(\theta \mid u, s)\, du\, ds}$$

for $t > e - b$, where b and e are the beginning and the end of follow-up for the individual, $a = \max(0, e - t)$, and L is given by (1).

The conditional density is then interpreted as a predictive density. A prediction point of t_D can be chosen to be the mode or the expected value of this conditional density, and the prediction interval can be computed from the percentile points. The expected value of the conditional density can be interpreted as an empirical Bayes estimate of t_D.

2. APPLICATION: TORONTO AIDS COHORT STUDY

2.1 Data Description

The Toronto AIDS cohort study has been described previously (Coates et al. 1990; Struthers and Farewell 1989). Briefly, it is as follows. A total of 249 healthy homosexual males were recruited between July 1984 and July 1985 and followed for an average of 5.5 years. These were partners of males who had been diagnosed with AIDS at most 1 year prior to enrollment. Some baseline data collected at enrollment included the date of first and last contact with the primary case. A battery of immunologic tests was performed at baseline and repeated about every 3 months.

At the time of recruitment 143 cases were seropositive, and a total of 16 cases seroconverted during follow-up. The analysis will be restricted to these 159 cases. The 90 patients who did not seroconvert by the end of follow-up are considered noninfectible for various reasons and excluded from analysis. We censor 31 cases who were treated by zidovudine (AZT) at the time of treatment initiation. A total of 39 AIDS cases were observed during follow-up. Following Coates et al. (1990), we assume that no infection could occur prior to January 1, 1978, so we set B to be 6.5.

As the markers of interest, we will consider the T4 counts and T4/T8 ratio (multiply by 100). It is now well known that T4 cells are the main target of HIV and that depletion of T4 cells usually follows an HIV infection. But, it is also known that there is a high variability in individual T4 count measurements. In contrast, T4/T8 ratio values are considered much more reliable, because they are obtained directly from flow-cytometric measurements. Available for analysis is a total of 1,768 T4 counts and 1,771 T4/T8 ratio measurements, ranging from 1 to 20 measurements per individual.

2.2 Three Models

We compare three models according to different assumptions about the time of infection. For Model 1 we assume the midpoint between the first and last contact, with the primary case as the infection time. For Model 2 we assume the infection time to be interval-censored between the first and last contact. These two assumptions carry more information than left censoring, but they may also induce some bias because they depend on the quality of the primary contact information. For Model 3 we consider the infection times for individuals who were seropositive at entry to be left censored (or interval censored between 0, the beginning of the epidemic on January 1, 1978, and the time of entry). In this analysis we do not include any covariate in the model for infection time and time to AIDS distributions.

To compute good starting values for the regression parameters β_0 and β_2 and the total variance σ_2, we fit a linear regression model of $\log Z(t)$ as a function of time only among the cases for which the seroconversion times are known. Some guesses are used for the other regression parameters, whereas initial estimates of the Weibull parameters are found by fitting a separate Weibull model to the time of infection

Table 1. Parameter Estimates for T4 Counts Models

Variable	Parameter	Model 1 Estimate	Model 1 Standard error	Model 2 Estimate	Model 2 Standard error	Model 3 Estimate	Model 3 Standard error
t_I	λ_{0I}	.16	.004	.15	.004	.19	.02
	α_I	4.35	.37	5.01	.44	2.49	.55
t_D	λ_{0D}	.10	.01	.11	.01	.11	.01
	α_D	1.92	.25	1.94	.28	3.23	.75
log(T4)	β_0	6.07	.08	6.06	.08	6.09	.06
	β_1	−.01	.35	−.06	.34	0	—
	β_2	.14	.03	.17	.04	.26	.06
	β_3	−1.54	.21	−1.65	.21	−2.49	.37
	σ^2	.30	.03	.30	.03	.26	.03
	ρ	.46	.05	.46	.05	.42	.06

NOTE: Model 1 is computed assuming the midpoint between the first and last contact as the infection time. For Model 2, the infection time is interval censored between the first and last contact. For Model 3, the infection time is censored between time zero and entry to the study, and we set $\beta_1 = 0$.

and the time to AIDS using the midpoint between the date of first and last contact as the date of infection.

2.3 Results

The summary results for T4 count models are given in Table 1. The regression estimates computed using the midpoints are quite similar to those computed using interval-censored infection times, except for β_3 estimate in Model 3. Due to more censoring, the variance estimates in Model 3 are generally greater than those in Model 2, which in turn are greater than those in Model 1. (In the next subsection we discuss which model is the most reasonable.) The estimated value of β_1 at −.01 (s.e. = .35) or −.06(.34) from Model 1 or Model 2 may indicate that the level of T4 count prior to infection is not associated with the time to AIDS. The parameter β_1 is set to 0 in Model 3 because if it is in the model, its estimate is too strongly correlated with the estimate of β_0, causing a very large variance. The slope of decline appears to be associated with the time to AIDS, as the estimates of β_3 are highly significant in all models. The significant estimates of the last parameter ρ, which measures the proportion of T4 variability due to person effects, confirm the tracking pattern in the individual T4 paths, as shown in Figure 1.

The summary results for T4/T8 ratio models are given in Table 2. Here we can apply the same comments as for the T4 models earlier. We also note that, except for Model 3, the parameter estimates for the infection time and the time to AIDS distributions are similar to those from the T4 count models.

In summary, there is a significant amount of tracking in the T4 and T4/T8 ratio paths over time and a fair amount of variation in rate of decline of T4 or T4/T8 ratio among individuals. This rate of decline is strongly associated with the time to AIDS, which suggests that the scaled incubation time may be a more natural scale for viewing the progression of HIV infection. We have also tried more complicated models, including quadratic and piece-wise linear with a change in slope prior to AIDS, but we could not improve the fit of the simple linear models. Basically, there is not enough resolution in the data to warrant nonlinear models over time.

2.4 Model Adequacy and Comparison

A formal check of the model adequacy is quite complicated. We would need to estimate $\mu_i(t)$ in (7) for every observation using the joint conditional distribution for t_{Ii} and t_{Di} given the data and then compute the residuals. Here we check the normality and Weibull assumptions and present an external check, in which we compare the estimates derived from the models with other estimates in the literature.

Figures 2(a) and (b) show that there is no strong evidence against the normality assumption for the log(T4) and log(T4/T8 ratio) measurements. Only the seroconverters are in-

Table 2. Parameter Estimates for T4/T8 Ratio Models

Variable	Parameter	Model 1 Estimate	Model 1 Standard error	Model 2 Estimate	Model 2 Standard error	Model 3 Estimate	Model 3 Standard error
t_I	λ_{0I}	.16	.004	.15	.004	.17	.01
	α_I	4.35	.37	5.07	.45	3.13	.56
t_D	λ_{0D}	.10	.01	.11	.01	.11	.01
	α_D	1.85	.23	1.89	.26	2.78	.75
log(T4/T8)	β_0	4.74	.07	4.69	.08	4.80	.07
	β_1	−.20	.31	−.06	.34	0	—
	β_2	.01	.02	.05	.03	.13	.06
	β_3	−1.64	.16	−1.80	.17	−2.44	.33
	σ^2	.23	.03	.23	.03	.23	.03
	ρ	.63	.05	.65	.05	.67	.05

NOTE: Model 1 is computed assuming the midpoint between the first and last contact as the infection time. For Model 2, the infection time is interval censored between the first and last contact. For Model 3, the infection time is censored between time zero and entry to the study, and we set $\beta_1 = 0$.

cluded in Figures 2(a) and (b), so that we are certain that the measurements are local around a fixed point of the process, which in this case is the time of seroconversion. Figure 1 also indicates that equal variance assumption over time is reasonable.

For the infection and time to AIDS densities, we try the three-parameter generalized gamma family that includes the Weibull model as a special case (Kalbfleisch and Prentice 1980, p. 27). We find that for all models, the estimated non-Weibull parameter to be very close to one and all regression estimates to be virtually unchanged, thus indicating that the Weibull model is adequate.

The lack of association between the level of T4 counts prior to seroconversion and the time to AIDS, as indicated by the estimates of β_1 in Table 1, is consistent with the result in Phair et al. (1992). In other cohort studies the average of T4 counts prior to seroconversion is known to be around 1,000, about 6.9 in log scale, which is much greater than the estimate of β_0 between 6.07 and 6.09 in Table 1. But we find that the average of log T4 counts among the 90 excluded *seronegative* cases on their first visit is 6.14. So for some reason, the Toronto cohort exhibit lower baseline T4 counts level.

The positive estimates for β_2 are somewhat unexpected, especially in Table 1, because it would mean that the T4 counts level is increasing for individuals with long incubation time (>9.5 years for Model 3). A more complex model is probably warranted; for example, we may set the slope $\gamma_{1i} = \min(\beta_2 + \beta_3/t_{Di}, 0)$. For this model we find that the previous estimates are practically unchanged (but the likelihood decreases slightly). This means that individuals with a long incubation time have a constant T4 counts level. For such individuals, it seems more sensible to fit a model with a late drop in the T4 counts level, but the follow-up data is not long enough.

In both tables and for all models, the shape parameter α_D of the time to AIDS distribution is significantly greater than 1, indicating an increasing hazard function. This fact is con-

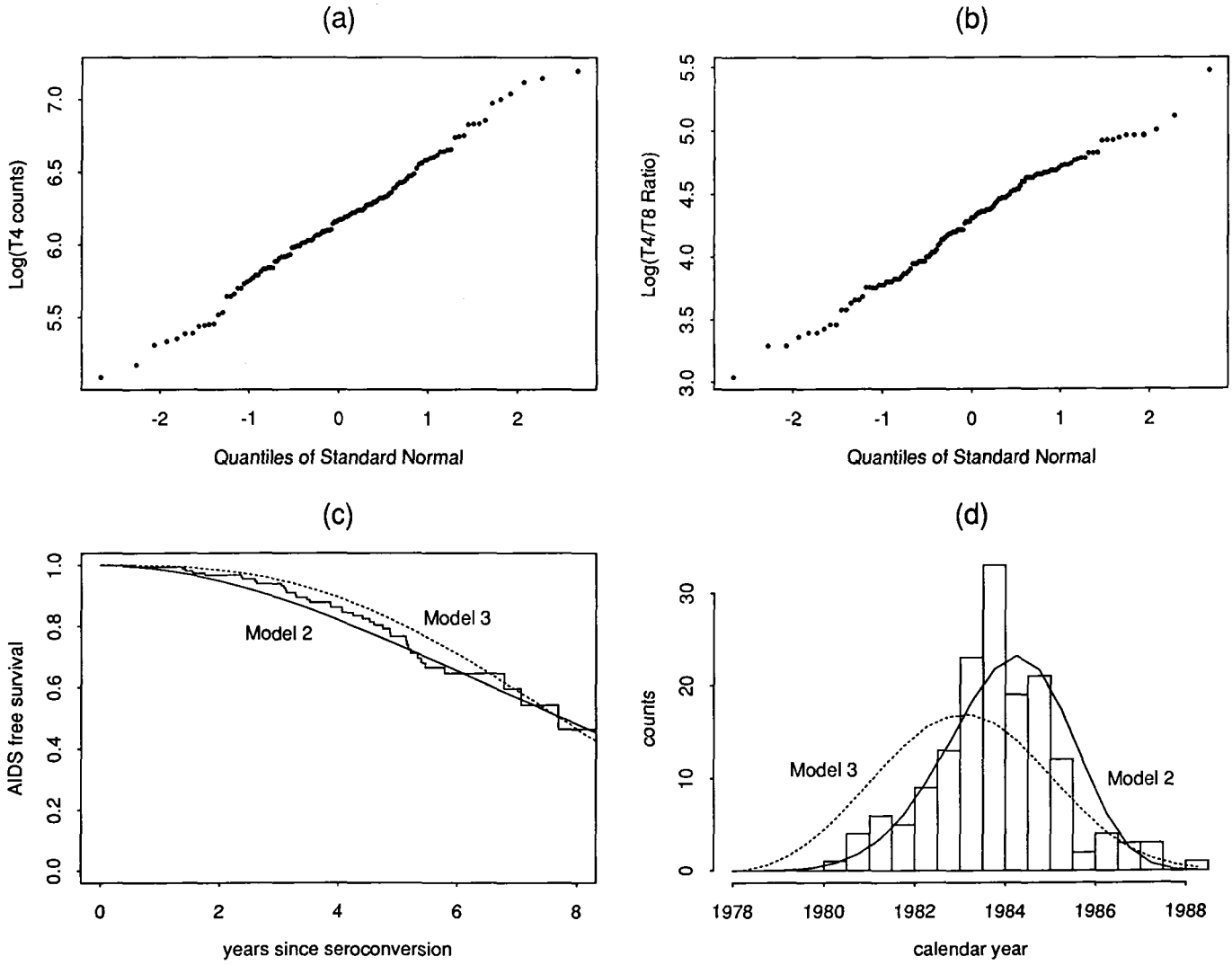

Figure 2. Normality Assumption Seems Reasonable. The primary contact information may be doubtful; (a) Normal QQ-plot for 128 T4 count measurements which are within 3 years after seroconversion, for seroconverters only. (One T4 count at 90 cells/mm³ just after seroconversion, which appears to be an outlier, is removed.) (b) Normal QQ-plot for 132 T4/T8 ratio measurements which are within 3 years after seroconversion, for seroconverters only. (c) AIDS incubation distribution: The Kaplan–Meier curve is estimated using the midpoints as infection times. The solid and dotted curves are the Weibull survival curves computed from Model 2 and Model 3 estimates of T4/T8 ratio data. (d) Infection distribution: The histogram is that of infection times estimated at midpoints of first and last contact for individuals who were seropositive at entry. The solid and dotted curves are the Weibull densities computed from Model 2 and Model 3 estimates for T4/T8 ratio model.

sistent with figure 1 of Coates et al. (1990). Figure 2(c) shows the Kaplan–Meier curve of the AIDS-free survival estimated using the midpoints as infection times (Model 1) compared to the curves computed from Model 2 and Model 3 estimates for T4/T8 ratio data. The Model 3 estimate shows a higher survival early on but drops to a similar level with the other models after 7.5 years. The Model 2 and Model 3 estimates each yield an estimated median incubation time around 7.7 years. This is comparable to a (parametric) estimate of 7.3 years for transfusion-related AIDS given by Kalbfleisch and Lawless (1988) and 7.8 years for the San Francisco City Clinic Cohort given by Lui, Darrow and Rutherford (1988), but less than the nonparametric estimate of 9.8 years for the San Francisco City Clinic Cohort given by Bacchetti and Moss (1989). Model 3 estimates based on T4 count data yield a median incubation time of 8.4 years, which is nearer to the nonparametric estimate.

Figure 2(d) shows the histogram of the infection times estimated at midpoints (Model 1) compared to the Weibull densities computed from Model 2 and Model 3 estimates for T4/T8 ratio data. Models 1 and 2 are relatively close but are quite dramatically different from Model 3. This difference may put some doubt on the primary contact assumption used in Models 1 and 2. With a peak infection around 1983, the estimated density from Model 3 is more consistent with that of the San Francisco Clinic Cohort (see fig. 1 in Bacchetti and Moss 1989 or fig. 4 in DeGruttola et al. 1991). So, considering that Model 3 uses no assumption about the infection from the primary contact, it appears that it is better than either Model 1 of Model 2.

2.5 Prediction of the Time to AIDS

To provide some examples of prediction of time to AIDS, we use the estimated Model 3 for T4/T8 ratio data. Figure 3(a) shows typical traces of T4/T8 ratio time series for two seropositive cases. The infection time for individual 1 is interval-censored within 1.5 years of entry but is known for individual 2 (whose series starts just after seroconversion.) Figure 3(b) shows the predictive densities for these two cases in solid and dotted curves. The modes are at 1.5 and 4.3 years from the last follow-up. The difference in the two curves is due to the fact that individual 1 has been infected longer and has a more consistent drop in T4/T8 ratio.

For comparison, the dashed curve in Figure 3(b) shows the predictive density for individual 1 assuming no information on the disease marker. In this case the individual is predicted to live longer, with the mode around 3.6 years. This is of course related to the fact that the T4/T8 series are associated with the time to AIDS. (Individual 1 was actually known to develop AIDS about 1 year after the last follow-up shown in Fig. 3(a).)

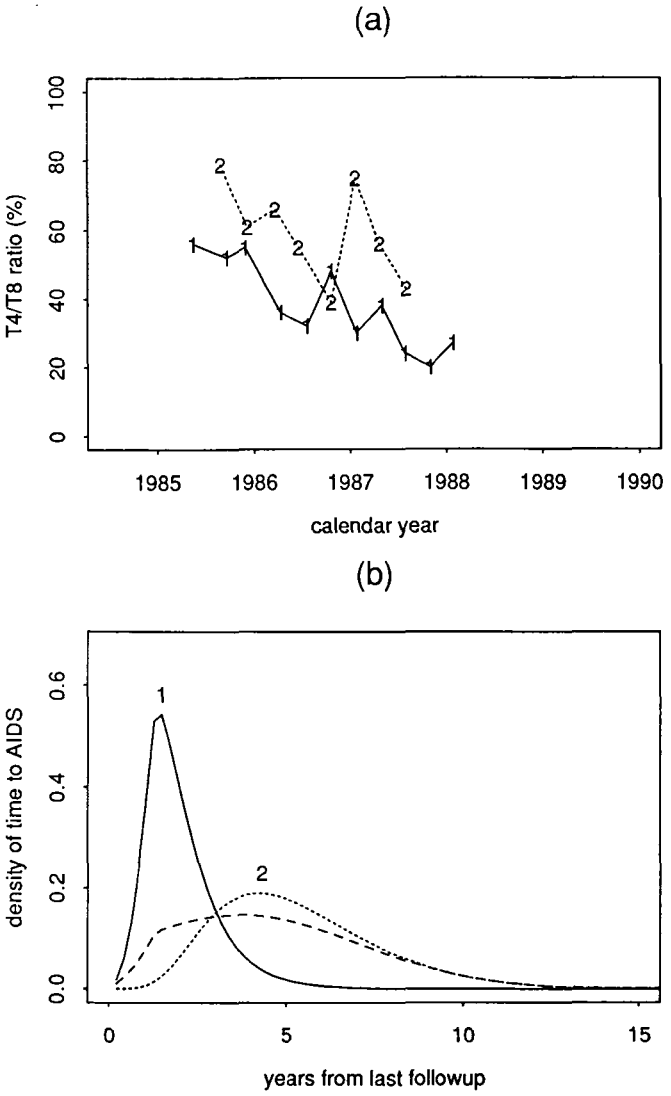

Figure 3. Prediction of Time to AIDS Given a Series of T4/T8 Ratio Measurements. (a) Two sample paths of T4/T8 ratio. (b) The corresponding predictive density of time to AIDS. The solid and dotted curves are for individuals 1 and 2. The dashed curve is for individual 1, assuming no information on T4/T8 ratio.

3. CONCLUSION

We have presented a method for analyzing longitudinal data arising in AIDS cohort studies. The method addresses the technical difficulties commonly found in analyzing such studies, namely left censoring due to unknown infection times, informative right censoring due to AIDS occurrence, and left truncation. Our approach is fully parametric, but less-rigid models can typically be incorporated by considering larger numbers of parameters, with the price of increased computation. It will be of interest to develop an inference which is robust with respect to the likelihood specification.

One referee pointed out that modeling the time to AIDS t_D given a disease marker $Z(t)$ seems to be more meaningful. We mention here three reasons to consider modeling $Z(t)$ given the time to AIDS. First, we are interested in $Z(t)$. The form t_D given $Z(t)$ is useful for predicting t_D, but does not tell us about the progression of $Z(t)$ over time. AIDS researchers tend to draw the T4 trajectories for different individuals with the time of infection as the time zero and the time of AIDS as the same endpoint. The model that we consider, including the scaled incubation time, seems to capture this idea. (Unconditional progression of $Z(t)$ over time can be found by taking the expectation of (7) over t_D.) Second, we essentially model the joint distribution of $(t_{Ii}, t_{Di}, Z(t))$. Our choice of conditioning deals nicely with the data structure (e.g., *varying amount* of $Z(t)$ information per individual), which in turn provides a flexible way to predict the time to AIDS using all of the marker data. Third, the con-

ditioning allows us to deal with the informative right-censoring problem (cf. Wu and Bailey 1988).

[*Received November 1991. Revised February 1993.*]

REFERENCES

Bacchetti, P., and Moss, A. R. (1989), "Incubation Period of AIDS in San Francisco," *Nature,* 338, 251–253.

Brookmeyer, R., and Goedert, J. J. (1989), "Censoring in an Epidemic With Application to Hemophilia-Associated AIDS," *Biometrics,* 45, 325–335.

Coates, R. A., Farewell, V. T., Raboud, J., Read, S. E., MacFadden, D. K., Calzavara, L. M., Johnson, J. K., Shepherd, F. A., and Fanning, M. M. (1990), "Cofactors of Progression to AIDS in a Cohort of Male Sexual Contacts With Men With HIV Disease," *American Journal of Epidemiology,* 132, 717–722.

DeGruttola, V., Lange, N., and Dafni, U. (1991), "Modeling the Progression of HIV Infection," *Journal of the American Statistical Association,* 86, 569–577.

Eyster, M. E., Gail, M. H., Ballard, J. O., Al-Mondhiry, H., and Goedert, J. J. (1987), "Natural History of HIV Infections in Hemophiliacs: Effects of T-Cell Subsets, Platelet Counts, and Age," *Annals of Internal Medicine,* 107, 1–6.

Fahey, J. L., Taylor, J. M. G., Detels, R., Hoffman, B., Melmed, R., Nishanian, P., and Giorgi, J. (1990), "The Prognostic Value of Cellular and Serologic Markers in Infection With HIV Type 1," *New England Journal of Medicine,* 322, 166–172.

Jewell, N. P. (1990), "Some Statistical Issues in Studies of the Epidemiology of AIDS," *Statistics in Medicine,* 9, 1387–1416.

Kalbfleisch, J. D., and Lawless, J. F. (1988), "Estimating the Incubation Period for AIDS Patients" *Nature,* 333, 504–505.

Kalbfleisch, J. D., and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data,* New York: John Wiley.

Laird, N. M., and Ware, J. H. (1982), "Random Effect Models for Longitudinal Data," *Biometrics,* 38, 963–974.

Lui, K. J., Darrow, W. W., and Rutherford, G. W. (1988), "Model Based Estimate of the Mean Incubation Period for AIDS in Homosexual Men," *Science,* 240, 1333–1335.

McCullagh, P., and Nelder, J. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Phair, J., Jacobson, L., Detels, R., Rinaldo, C., Saah, A., Schrager, L., and Munoz, A. (1992), "Acquired Immune Deficiency Syndrome Occurring Within 5 Years of Infection With Human Immunodeficiency Virus Type-1: The Multicenter AIDS Cohort Study," *Journal of AIDS,* 5, 490–496.

Schecter, M. T., Craib, K. J. P., Le, T. N., Willoughby, B., Douglas, B., Sestak, P., Montaner, J. S. G., Weaver, M., Elmslie, K. D., and O'Shaughnessy, M. V. (1989), "Progression to AIDS and Predictors of AIDS in Seroprevalent and Seroincident Cohorts of Homosexual Men," *AIDS,* 3, 347–353.

Struthers, C. A., and Farewell, V. T. (1989), "A Mixture Model for Time to AIDS Data With Left Truncation and an Uncertain Origin," *Biometrika,* 76, 814–817.

Wu, M. C., and Bailey, K. (1988), "Analyzing Changes in the Presence of Informative Right Censoring Caused by Death and Withdrawal," *Statistics in Medicine,* 7, 337–346.

Wu, M. C., and Carroll, R. J. (1988), "Estimation and Comparison of Changes in the Presence of Informative Right Censoring: Modeling the Censoring Process," *Biometrics,* 44, 175–188.