Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS

A. A. TSIATIS, Victor DEGRUTTOLA, and M. S. WULFSOHN*

A question that has received a great deal of attention in evaluating new treatments in acquired immune deficiency syndrome (AIDS) clinical trials is that of finding a good surrogate marker for clinical progression. The identification of such a marker may be useful in assessing the efficacy of new therapies in a shorter period. The number of CD4-lymphocyte counts has been proposed as such a potential marker for human immune virus (HIV) trials because of its observed correlation with clinical outcome. But to evaluate the role of CD4 counts as a potential surrogate marker, we must better understand the relationship of clinical outcome to an individual's CD4 count history over time. The Cox proportional hazards regression model is used to study the relationship between CD4 counts as a time-dependent covariate and survival. Because the CD4 counts are measured only periodically and with substantial measurement error and biological variation, standard methods for estimating the parameters in the Cox model by maximizing the partial likelihood are no longer appropriate. Instead, we propose a two-stage approach. In the first stage the longitudinal CD4 count data are modeled using a repeated measures random components model. In the second stage methods for estimating the parameters in a Cox model when the data are assumed to be of this form are derived. We also considered methods to account for missing data patterns. These methods are applied to CD4 data from a randomized clinical trial of AIDS patients where half of the patients were randomized to receive Zidovudine (ZDV) and the other half were randomized to receive a placebo. Although a strong correlation between CD4 counts and survival is demonstrated, we also show that CD4 counts may not serve as a useful surrogate marker for assessing treatments for this population of patients.

KEY WORDS: Longitudinal data; Proportional hazards model; Random components model; Survival analysis.

1. INTRODUCTION

A randomized clinical trial is often used to evaluate new treatments in patients with the human immune virus (HIV). In the past, most HIV trials use clinical progression as the primary outcome. A surrogate marker of clinical outcome would permit clinical trials to be completed in a shorter period and with a much smaller sample size. An acceptable surrogate marker would be one that responds rapidly to treatment and whose response implies a benefit regarding clinical outcome. Because of its observed rapid response to some antiretroviral therapies and observed correlation with clinical outcome measures, the number of CD4 T-lymphocytes (hereinafter called the CD4 count) has been proposed as such a marker for HIV trials.

To be more precise, if we use the definition of Prentice (1989), a good surrogate marker should have the following three properties:

1. The marker should be related to prognosis.
2. The distribution of values for the marker should be different for individuals receiving an effective treatment versus those receiving a placebo.
3. The beneficial effects of a good treatment should be mediated through its effect on the marker. That is, patients with the same value of a marker should have the same prognosis whether they are receiving a treatment or a placebo. In such a case the better prognosis associated with a good treatment could be explained by the change in the value of the marker for that treatment.

In order to assess Conditions 1 and 3, we need to establish the relationship of prognosis to the marker for both patients receiving treatment and those receiving placebo. In this article we specifically examine whether CD4 counts may serve as a suitable surrogate marker for survival for patients with acquired immune deficiency syndrome (AIDS). Our results are based on a completed double-blind placebo-controlled trial conducted by Burroughs-Wellcome, which treated 281 patients with advanced HIV disease. Of these, 137 patients were randomized to receive a placebo and 144 patients were randomized to receive a 250-mg dose of Zidovudine (ZDV) every four hours. In this study CD4 counts were determined prior to treatment and approximately every four weeks during therapy. The median duration of follow-up was 120–127 days, at which point the study was stopped due to the superior results of the ZDV arm in decreasing mortality. At the termination of the study, all patients actively participating in the study were offered ZDV and subsequently followed for clinical outcomes.

In this trial, as in most HIV clinical trials, CD4 counts are measured only periodically and with a substantial amount of variability. The coefficient of variation was about 50%, based on replicate measurements at baseline. This variability arises from measurement error as well from true biologic differences such as diurnal fluctuations.

The proportional hazards regression model (Cox 1972) was used to study the relationship of CD4 counts as a time-dependent covariate to survival. To apply Cox's methods for the estimation of the model parameters, it is necessary to have complete knowledge of the covariate history for all individuals while on study. In most cases, however, and certainly for the clinical trial we are considering, the time-dependent covariate is measured only periodically and with

* Anastasios A. Tsiatis is Professor of Biostatistics, Victor DeGruttola is Associate Professor, and Michael Wulfsohn is Research Assistant, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115. This research was supported by National Institutes of Health Grants RO1-AI31789, RO1-CA51962, R29-AI28905, and RO1-CA50597. The authors thank Urania Dafni for helpful discussions and Margaret Fischl for providing the data on which this work was based.

© 1995 American Statistical Association
Journal of the American Statistical Association
March 1995, Vol. 90, No. 429, Applications and Case Studies

measurement error. Even for situations where measurement error is unimportant, the individual's value of the covariate must be known at all points in time to maximize the partial likelihood (Cox 1975) to obtain parameter estimates. Because all required covariate values are rarely available, data analysts generally impute these values. For example, one might impute the most recent value by using the prior measured value or by making a linear interpolation between two measured values. If sufficient measurements are made and the time-dependent covariate does not change quickly over time, then these imputation methods will probably work well. But if the measurements are infrequent, then there is no guarantee that an ad hoc approach will lead to good answers.

Along with the problem of missing data, the error in measuring covariates leads to biased estimation of regression parameters that describe the relationship between hazard and true value of the covariate (Prentice 1982). Whether one is interested in the relationship between hazard and the true marker or the observed marker depends on how that marker is to be used. If one is interested in making regulatory decisions based on marker response to therapy, then one would require estimates of the relationship with the true marker value. Estimates of the relationship with the observed marker value, unadjusted for measurement error, might mislead one into believing that a good surrogate marker was a poor one. The reason for this is that surrogacy, or the degree to which treatment benefit for a clinical outcome is reflected in treatment response on the marker, is a characteristic of a population of patients. If the true marker value is surrogate for clinical outcome, then knowledge of the treatment response of the observed marker averaged over a population (which is much more precise than for an individual) will provide an accurate indication of the treatment's benefit on the clinical outcome of interest. Thus for regulatory decisions, estimation of the hazard relationship to the true marker is important, even if this value can never be observed.

Similarly, knowledge of the degree to which a clinical benefit of treatment is mediated through a marker, after adjustment for measurement error, gives important insight into the effect of treatment on the disease mechanism. Note that the degree to which the observed marker response of an individual patient provides prognostic information for that patient depends on measurement error. Thus the estimated hazard expressed as a function of the marker, after adjustment of measurement error, shows the maximum explanatory power of the marker that one could attain by reducing measurement error.

In this article we develop new methods appropriate for this problem and apply them to data from the placebo-controlled trial of ZDV described in Sections 5 and 7. This analysis has two specific aims: to delineate the relationship between the hazard rate of dying and an individual's CD4 count history, and to determine whether CD4 may serve as a useful surrogate marker in assessing treatment; that is, to establish whether the beneficial effect of ZDV on survival is explained through its effect on CD4 counts.

1.1 The Model and Notation

Let $Z^*(t)$ denote the hypothetical true value of CD4 count at time t, if we were able to evaluate it without measurement error, and let $\bar{Z}^*(t)$ denote the history up to time t, $\{Z^*(u), u \leq t\}$. This is in contrast to $Z(t)$, which is the measured CD4 count at time t. The primary interest is to estimate the relationship between survival and the true CD4 count history. This relationship will be described through the hazard function. If we denote by T the survival time of an individual, then the hazard rate at time t as a function of the covariate history up to time t is defined as

$$\lambda(t \mid \bar{Z}^*(t)) = \lim_{h \to 0} \frac{1}{h} \{\mathrm{pr}(t \leq T < t + h \mid T \geq t, \bar{Z}^*(t))\}.$$

As in most clinical trials, the survival data are subject to right censoring. Therefore, we observe $X = \min(T, C)$, where C corresponds to a potential censoring time, and the failure indicator Δ, which is equal to 1 if the individual is observed to fail ($T \leq C$), and zero otherwise. In such a case we can only observe the cause-specific hazard, namely,

$$\lim_{h \to 0} \frac{1}{h} \{\mathrm{pr}(t \leq X < t + h, \Delta = 1 \mid X \geq t, \bar{Z}^*(t))\}. \quad (1)$$

It is assumed, however, that censoring is noninformative, in which case the cause-specific hazard given by (1) is equal to the hazard of interest $\lambda(t \mid \bar{Z}^*(t))$.

The proportional hazards model introduced by Cox (1972) relates the hazard to time-dependent covariates,

$$\lambda(t \mid \bar{Z}^*(t)) = \lambda_0(t) f(\bar{Z}^*(t), \beta),$$

where $f(\bar{Z}^*(t), \beta)$ is a function of the covariate history specified up to an unknown parameter β (possibly vector valued). If the underlying hazard $\lambda_0(t)$ is left unspecified, then the parameter β is estimated by maximizing the partial likelihood given by Cox (1975), namely

$$\prod_{i=1}^{n} \left[f(\bar{Z}_i^*(X_i), \beta) \Big/ \sum_{j=1}^{n} f(\bar{Z}_j^*(X_i), \beta) Y_j(X_i) \right]^{\Delta_i}, \quad (2)$$

where $Y_j(v)$ is the indicator of being at risk time v, $I(X_j \geq v)$.

To apply this methodology, one needs the knowledge of $Z^*(t)$ for all values $t \leq X$. This is not generally available. For example, the CD4 counts are measured only at certain occasions with substantial measurement error. That is, the observable data are given by the vector $\{Z(t_1), \ldots, Z(t_m)\}$, $t_m \leq X$, where $Z(t)$ is the observed laboratory-measured CD4 count at time t. Therefore, modeling the hazard directly on the observed CD4 counts $Z(t)$ could lead to biased estimates of the true hazard relationship. Biases in the Cox model due to measurement error have been discussed by Prentice (1982).

2. THE EFFECT OF MEASUREMENT ERROR

To describe the effect of measurement error, let the observed CD4 count $Z(t)$ be equal to the true CD4 count $Z^*(t)$ plus measurement error $e(t)$. That is,

$$Z(t) = Z^*(t) + e(t). \quad (3)$$

We shall assume that $e(t)$ is random noise, so that $E(e(t)) = 0$,

$$\text{var}(e(t)) = \sigma^2 \quad \text{and} \quad \text{cov}(e(s), e(t)) = 0, \qquad s \neq t.$$

Denote by $\bar{\mathbf{Z}}(t)$ the history of observed CD4 counts up to time t, that is, $\bar{\mathbf{Z}}(t) = \{Z(t_1), \ldots, Z(t_j); t_j \leq t\}$. In such a case the observable hazard is $\lambda(t|\bar{\mathbf{Z}}(t))$ rather than the desired $\lambda(t|\bar{Z}^*(t))$. But a simple application of the law of conditional probability yields

$$\lambda(t|\bar{\mathbf{Z}}(t)) = \int \lambda(t|\bar{\mathbf{Z}}(t), \bar{Z}^*(t))\, dP(\bar{Z}^*(t)|\bar{\mathbf{Z}}(t), X \geq t).$$

If it is additionally assumed that neither measurement error nor the timing of the visits prior to time t are prognostic, then

$$\lambda(t|\bar{\mathbf{Z}}(t), \bar{Z}^*(t)) = \lambda(t|\bar{Z}^*(t)) = \lambda_0(t) f(\bar{Z}^*(t), \beta),$$

in which case

$$\lambda(t|\bar{\mathbf{Z}}(t)) \\ = \lambda_0(t) E[f(\bar{Z}^*(t), \beta)|Z(t_1), \ldots, Z(t_j), X \geq t]. \quad (4)$$

The assumption that the timing of visits prior to t is not prognostic is examined more carefully in Section 7. In that section we also consider how deviations from this assumption affect the overall results. For the time being, the results will be obtained under the assumption leading to (4).

The conditional expectation in (4) will be denoted by $E(t, \beta)$. Because (4) is also a proportional hazards relationship, if $E(t, \beta)$ were known, then β can be estimated by maximizing the partial likelihood,

$$\prod_{i=1}^{n} \left[E_i(X_i, \beta) \Big/ \sum_{j=1}^{n} E_j(X_i, \beta) Y_j(X_i) \right]^{\Delta_i}. \quad (5)$$

Analytic expressions for the foregoing conditional expectation are hard to obtain. Therefore, in the next section we shall consider methods for approximating the conditional expectations $E_i(u, \beta)$ under various assumptions.

3. APPROXIMATING THE RELATIONSHIP FOR CONDITIONAL EXPECTATION

For our application, we shall consider the case only when the hazard is a function of the current value, $Z^*(t)$, rather than a functional of the entire history, $\bar{Z}^*(t)$. The computations needed for the latter case are very similar. In Section 5 we investigate models that relate the hazard to various functions of the surrogate.

We first considered the relative risk formulation of the original Cox model (Cox 1972), where

$$f(Z^*(t), \beta) = \exp(\beta Z^*(t)). \quad (6)$$

For such a model, the value $E(t, \beta)$ is given by $E[e^{\beta Z^*(t)}|\bar{\mathbf{Z}}(t), X \geq t]$. This is the moment-generating function for the conditional distribution of $Z^*(t)$, given $\bar{\mathbf{Z}}(t)$, among the individuals at risk at time t.

To compute $E(t, \beta)$ we must be able to characterize the joint distribution of $\{\bar{\mathbf{Z}}(t), Z^*(t)\}$, given that $X \geq t$. A useful way of looking at this problem is to think of the history of true CD4 counts up to time t for an individual i, $\bar{Z}_i^*(t)$, given that the individual is at risk at time t [i.e., $(X_i \geq t)$], as a realization of a stochastic process. If we additionally make the assumption that this is a normal process, then it can be characterized by the structure of its first two moments; that is, for $u \leq t$,

$$E(Z^*(u)|X \geq t) = \mu_t(u)$$

and

$$\text{cov}(Z^*(u), Z^*(v)|X \geq t) = C_t(u, v). \quad (7)$$

If the measurement error, $e(u)$, given in (3), is also normally distributed, independent of the stochastic process $Z^*(u)$, and $X \geq t$, then the random vector $\{\bar{\mathbf{Z}}(t), Z^*(t)\}$ is jointly normal. The mean is given as the $j + 1$ vector $\{\bar{\boldsymbol{\mu}}_t, \mu_t(t)\}$, where $\bar{\boldsymbol{\mu}}(t) = \{\mu_t(t_1), \ldots, \mu_t(t_j)\}$ and the variance is the $(j + 1) \times (j + 1)$ matrix,

$$\mathbf{M}_t = \begin{bmatrix} C_t(t_1, t_1), \ldots, & C_t(t_1, t_j), & C_t(t_1, t) \\ \vdots & & \\ C_t(t_j, t_1), \ldots, & C_t(t_j, t_j), & C_t(t_j, t) \\ C_t(t, t_1), \ldots, & C_t(t, t_j), & C_t(t, t) \end{bmatrix} \\ + \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ & 0 & \sigma^2 \\ & & 0 \end{bmatrix}.$$

For jointly normal random variables, the conditional expectation $E[Z^*(t)|\bar{\mathbf{Z}}(t), X \geq t]$ is equal to

$$\mu_t(t) - (M_t^{22})^{-1}(\mathbf{M}_t^{21})[\bar{\mathbf{Z}}(t) - \bar{\boldsymbol{\mu}}(t)]', \quad (8)$$

where $\mathbf{M}_t^{-1}$ can be written as the partitioned matrix,

$$\begin{bmatrix} (\mathbf{M}_t^{11})^{j \times j} & (\mathbf{M}_t^{12})^{j \times 1} \\ (\mathbf{M}_t^{21})^{1 \times j} & (M_t^{22})^{1 \times 1} \end{bmatrix}. \quad (9)$$

In the next section we shall discuss methods for estimating the mean and covariance structure of $Z^*(u)$.

Under the assumption of joint normality, it is well known (see Rao 1973) that the moment-generating function $E(t, \beta)$ is given by

$$\exp\{\beta\mu(t|\bar{\mathbf{Z}}(t)) + \beta^2\sigma^2(t|\bar{\mathbf{Z}}(t))/2\}, \quad (10)$$

where $\mu(t|\bar{\mathbf{Z}}(t)) = E\{Z^*(t)|\bar{\mathbf{Z}}(t), X \geq t\}$ given by (8) and

$$\sigma^2(t|\bar{\mathbf{Z}}(t)) = \text{var}\{Z^*(t)|\bar{\mathbf{Z}}(t), X \geq t\},$$

which is equal to (M_t^{22}) defined in (9). Hence the estimate of β is obtained by maximizing the partial likelihood (5), substituting the expression in (10) for $E(t, \beta)$.

We note that Self and Pawitan (1992) considered a similar problem but used an additive relative risk function $f(Z^*(t), \beta) = 1 + \beta Z^*(t)$ considered by Prentice and Self (1983). But if we used relative risk functions $f(Z^*(t), \beta)$ that are more complex than the Cox model (6) or the additive relative risk model, then simple expressions for $E(t, \beta)$ could not be obtained, even with the assumption of joint normality. For example, we may wish to use the Cox model with higher-order polynomial terms. This is used later for comparison to model (6) to check adequacy of fit. Or we may wish to check the adequacy of the proportional hazards assumption by introducing an interaction term between time and CD4 counts. For such models, we considered a simple first-order approximation to $E(t, \beta)$, namely

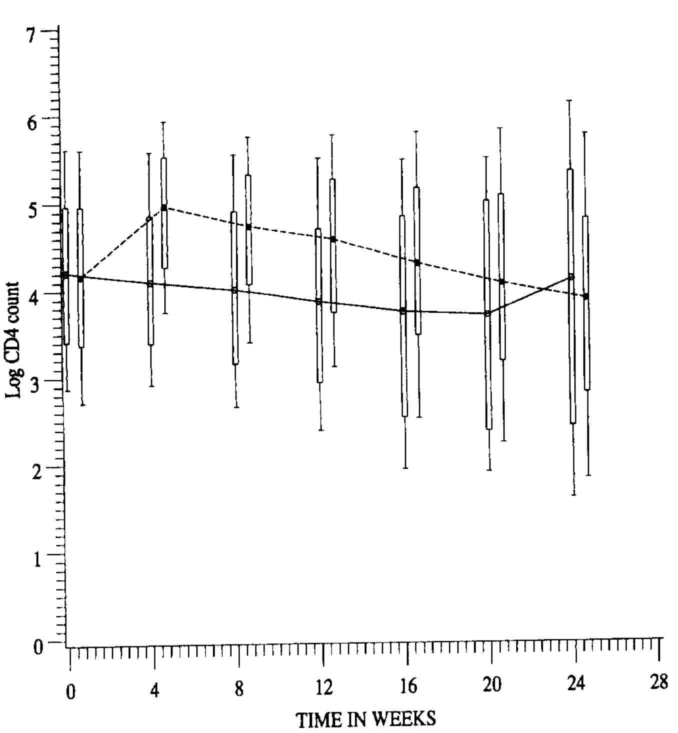

Figure 1. Boxplots of Log CD4 Counts Over Time Measured From Randomization. The solid line represents patients randomized to placebo; the dashed line represents the patients randomized to receive ZDV.

$$E(t,\beta) = E[f(Z^*(t),\beta)\,|\,\bar{\mathbf{Z}}(t), X \geq t)] \approx f(E(Z^*(t)\,|\,\bar{\mathbf{Z}}(t), X \geq t), \beta). \qquad (11)$$

The adequacy of this approximation for general models is under study. But when this approximation was used for the Cox model (6), the estimate for β is obtained by maximizing the partial likelihood

$$\prod_{i=1}^{n}\left[\exp\{\beta\mu_i(X_i\,|\,\bar{\mathbf{Z}}_i(X_i))\} \div \sum_{j=1}^{n}\exp\{\beta\mu_j(X_i\,|\,\bar{\mathbf{Z}}_j(X_i))\}Y_j(X_i)\right]^{\Delta_i}.$$

The resulting estimates from this approximation were virtually the same as those derived using (10). The approximation given by (11) is useful, because standard software for the Cox model can be used.

To get tractable answers, we have made the assumption that the covariate process is Gaussian among individuals at each risk set. This assumption is not technically reasonable, as it would necessitate the existence of covariate and failure time processes that induce the family of jointly Gaussian distributions conditional on being at risk at each time t. Such processes may not even exist. Nonetheless, we show in later sections that these assumptions are practically reasonable in that they are good approximations to the data at each risk set.

4. DESCRIBING THE STOCHASTIC PROCESS

To apply the methods described in Section 8, we must transform the covariate data to approximate normality. From here on we shall refer to the transformed data as $Z(u)$. Also, within each risk set, $X \geq t$, we must be able to model the mean and covariance structure of $Z(u)$, $0 \leq u \leq t$ conditional on $X \geq t$. To do this, we consider a class of flexible, parsimonious models to aid us in the choice of the mean and covariance structure.

Due to the nature of the data and available software, we modeled the covariate data using linear random components models as described by Laird and Ware (1982). We found that transforming to a log scale normalized the data. Also, the log CD4 counts appeared to decline over time in a linear fashion for patients randomized to placebo, whereas patients randomized to ZDV had an increase for the first four weeks, followed by a linear decline.

Figure 1 depicts the log CD4 over time for both ZDV patients and placebo patients using box plots at times where CD4 counts were measured. Figure 2 plots the longitudinal data for the first five patients randomized to ZDV and five patients on placebo. These two figures illustrate the log-linear decline of CD4 after four weeks for ZDV patients, as well as the large degree of within-patient variability.

Therefore, for placebo patients we used linear growth curve models in each risk set to model the log CD4 counts. That is, given $X \geq t$, it is assumed that the ith patient has log CD4 counts given by

$$Z_i(u) = Z_i^*(u) + e_i(u), \qquad u \leq t,$$

where

$$Z_i^*(u) = \theta_{0i} + \theta_{1i}u$$

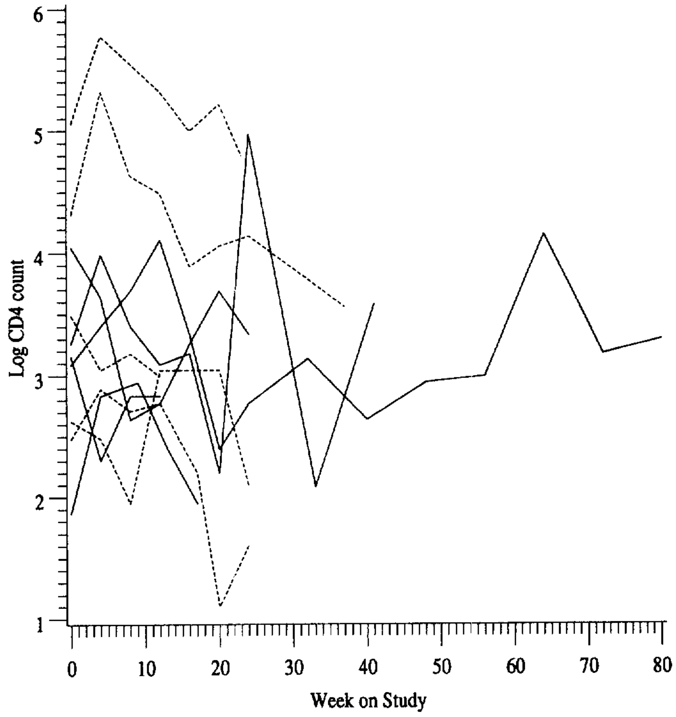

Figure 2. Individual Log CD4 Count Trajectories Over Time. The solid lines are the trajectory of log CD4 counts for the first five patients randomized to ZDV; the dashed lines are for the first five patients randomized to placebo.

and $e_i(u)$ corresponds to random measurement error. In this model it is assumed that each individual randomized to placebo follows a linear trajectory with his or her own slope and intercept, θ_{1i} and θ_{0i}, which are themselves considered to be independent realizations of a bivariate normal random variable. That is, $(\theta_{0i}, \theta_{1i})'$ are iid normal bivariate vectors with mean $(\theta_0^{(t)}, \theta_1^{(t)})'$ and covariance matrix

$$\begin{bmatrix} \sigma_{\theta_0\theta_0}^{(t)} & \sigma_{\theta_0\theta_1}^{(t)} \\ \sigma_{\theta_0\theta_1}^{(t)} & \sigma_{\theta_1\theta_1}^{(t)} \end{bmatrix}.$$

We note that the mean and covariance structure among individuals at risk at time t are allowed to change by letting the mean and variance of the random components be a function of t.

The measurement error $e_i(u)$ is also assumed to be distributed as a normal random variable with mean zero and variance σ_e^2 independent of the random vector $(\theta_{01}, \theta_{1i})$. This conceptualization yields a mean and nonstationary covariance structure in (8) given by

$$\mu_t(u) = \theta_0^{(t)} + \theta_1^{(t)}u$$

and

$$C_t(u, v) = \sigma_{\theta_0\theta_0}^{(t)} + \sigma_{\theta_0\theta_1}^{(t)}(u + v) + \sigma_{\theta_1\theta_1}^{(t)}uv. \qquad (12)$$

To distinguish between the models for patients randomized to placebo and patients randomized to ZDV, we shall denote all the parameters for models describing ZDV patients by using ψ's instead of θ's. For each patient randomized to ZDV, we assume that $Z_i^*(u)$ is a piecewise linear spline with a knot at four weeks to reflect the possible increase that occurs at that time. That is, we assume that

$$Z_i^*(u) = \psi_{0i} + \psi_{1i}u + \psi_{2i}(u - 4)_+,$$

where x_+ is equal to x for $x \geq 0$ and to 0 for $x < 0$. Here we assume that $(\psi_{0i}, \psi_{1i}, \psi_{2i})$ are independent realizations from a normal distribution with mean $(\psi_0^{(t)}, \psi_1^{(t)}, \psi_2^{(t)})$ and covariance matrix

$$\begin{bmatrix} \sigma_{\psi_0\psi_0}^{(t)} & \sigma_{\psi_0\psi_1}^{(t)} & \sigma_{\psi_0\psi_2}^{(t)} \\ \sigma_{\psi_0\psi_1}^{(t)} & \sigma_{\psi_1\psi_1}^{(t)} & \sigma_{\psi_1\psi_2}^{(t)} \\ \sigma_{\psi_0\psi_2}^{(t)} & \sigma_{\psi_1\psi_2}^{(t)} & \sigma_{\psi_2\psi_2}^{(t)} \end{bmatrix}.$$

The methods of Section 8 require knowledge of the mean and covariance structure at each risk set t, $\mu_t(u)$ and $C_t(u, v)$, given by (12). For these models, this is equivalent to knowing the population parameters $\theta^{(t)}$, $\psi^{(t)}$, $\sigma_\theta^{(t)}$, $\sigma_\psi^{(t)}$, and σ_e^2. Of course, these quantities are not known and must be estimated by the data, $\{\bar{Z}_i(t)\}$, for all i such that $\{X_i \geq t\}$. These parameters were estimated using restricted maximum likelihood, as described by Laird and Ware (1982) and using software provided by Lindstrom and Bates (1988).

The values $E(t, \beta)$ in the partial likelihood of (5) are functions of the unknown population parameters. We thus used $\hat{E}(t, \beta)$, which are the values of $E(t, \beta)$ with the restricted maximum likelihood estimates of these parameters substituted for their true value. An important question that still needs to be addressed is the impact of using $\hat{E}(t, \beta)$ instead of $E(t, \beta)$ in the partial likelihood (4). Simulation studies have implied that empirical estimates of the variance of β were close to the variance obtained from the Cox model.

Linear random effects models were used primarily for convenience. Any model that reasonably approximates the mean and covariance structure of $Z(u)$ among individuals at risk at time t might be used.

Some diagnostics were performed to assess the adequacy of these models. For example, to determine whether the linear growth curve assumptions were reasonable, we also fitted quadratic growth curve models to the log CD4 data and considered the increase in twice the log-likelihood as a diagnostic for fit. Unfortunately, we cannot use the usual likelihood theory here, because the null hypothesis falls on the boundary of the parameter space for such random-components models. Although there is some asymptotic theory on the behavior of likelihood ratio tests in such situations (see Self and Liang 1987), we considered deriving the distribution of the likelihood ratio test using a parametric bootstrap approach. That is, using the parameters that were estimated from the linear growth curve model on the actual CD4 data, we randomly generated data from distributions corresponding to this normal linear growth curve model. For each such set of data, we computed the maximized likelihood from fitting a linear growth curve model and a quadratic growth curve model. Therefore, we were able to empirically compute the approximate null distribution of the likelihood ratio test. We then could compare the likelihood ratio test for the actual set of data to the simulated distribution of the likelihood ratio test to determine significance. We applied this procedure to each of the risk sets both for patients randomized to placebo and for patients randomized to ZDV. In no case did we find a significant increase in the maximized log-likelihood.

As an illustration, when we applied the foregoing methods to patients randomized to placebo who were at risk at 16 weeks, we obtained the following results:

Maximized Log-Likelihood

	Actual	Simulation	
Linear	−141.63	−135.2	(mean)
		−135.8	(median)
Quadratic	−140.99	−133.3	(mean)
		−134.4	(median)
Twice the Difference	1.27	3.72	(mean)
		3.36	(median)
		9.50	(95% percentile)

Clearly, the observed likelihood ratio test of 1.27 is not a significantly large value. Therefore, the assumption that the log CD4 counts follow a linear growth curve model seems quite reasonable.

Because many of the foregoing results also depend on the normality of the random components, as a rough check of this assumption we computed Q-Q plots of the empirical Bayes estimates of $\hat{\theta}_{0i}^t$ and $\hat{\theta}_{1i}^t$ for patients randomized to placebo and $\hat{\psi}_{0i}^t$, $\hat{\psi}_{1i}^t$, $\hat{\psi}_{2i}^t$ for patients randomized to ZDV at different risk times t. These plots generally indicate a close approximation to normality. In Figure 3 we present the Q-Q plot for placebo patients at week 17 for the purpose of illustration.

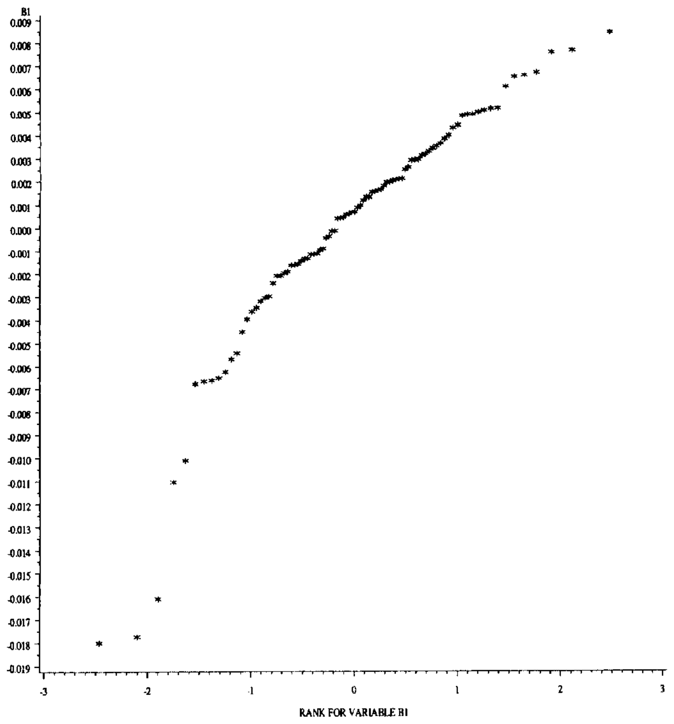

Figure 3. Q-Q Plot of the Empirical Bayes Estimates of Slopes for Placebo Patients at Risk at 17 Weeks.

5. RELATIONSHIP OF CD4 COUNTS TO SURVIVAL

The growth curve random components model described in Section 4 is useful in describing the history or trajectory of CD4 counts over time. At any point in time t, the trajectory of CD4 counts up to that time could be summarized by the vectors $\boldsymbol{\theta}_i$ or $\boldsymbol{\psi}_i$ for patients randomized to placebo or to ZDV. For example, patients randomized to placebo have linear trajectories in the log CD4 that are summarized by their slope and intercept $(\theta_{1i}, \theta_{0i})$. Therefore, the hazard function can be modeled to the past history of CD4 counts as

$$\lambda_i(t) = \lambda_0(t) f(\boldsymbol{\theta}_i, t, \beta), \qquad (13)$$

where $f(\boldsymbol{\theta}_i, t, \beta)$ is some function of the CD4 history up to time t (summarized by $\boldsymbol{\theta}_i$), specified as a function of unknown parameters β, which must be estimated. For example, a model that relates the hazard function to both the current value of CD4 as well as the slope of CD4 can be written as

$$\lambda_i(t) = \lambda_0(t)\exp\{\beta_1(\theta_{0i} + \theta_{1i}t) + \beta_2\theta_{1i}\}. \qquad (14)$$

As discussed in Section 8, the hazard function induced by (13) as a function of the observed history of CD4 counts is approximately equal to

$$\lambda(t|\bar{\mathbf{Z}}_i(t)) = \lambda_0(t) f(\hat{\boldsymbol{\theta}}_i^{(t)}, t, \beta),$$

where

$$\hat{\boldsymbol{\theta}}_i^{(t)} = \hat{E}(\boldsymbol{\theta}_i | \bar{\mathbf{Z}}_i(t), X \geq t).$$

The values $\hat{\boldsymbol{\theta}}_i^t$ are the so-called empirical Bayes estimates of the individual random effects as described by Laird and Ware (1982), evaluated at risk time t. Therefore, to estimate the parameters β in model (13), we first fit a separate growth curve random components model at each risk set time t. We then substituted the empirical Bayes estimates $\hat{\boldsymbol{\theta}}_i^{(t)}$ for $\boldsymbol{\theta}_i$ in the partial likelihood

$$\prod_{i=1}^{n} \left\{ f(\hat{\boldsymbol{\theta}}_i^{(X_i)}, X_i, \beta) \Big/ \sum_{j=1}^{n} f(\hat{\boldsymbol{\theta}}_j^{(X_i)}, X_i, \beta) Y_j(X_i) \right\}^{\Delta_i}.$$

We analyzed the data from the patients randomized to ZDV separately from those randomized to placebo. For each treatment group, we considered different aspects of CD4 trajectory, individually and in combination, as those to be included as time-dependent covariates in the Cox model, similar to the example given in (14). For patients randomized to placebo, we consider baseline CD4 (θ_{0i}), slope of decline (θ_{1i}), and current value, $(\theta_{0i} + \theta_{1i}t)$. For patients randomized to ZDV, we considered baseline CD4 (ψ_{0i}), slope of decline after four weeks $(\psi_{1i} + \psi_{2i})$, initial increase (ψ_{1i}), and current value $(\psi_{0i} + \psi_{1i}t + \psi_{2i}(t-4)_+)$.

The separate analysis of each treatment group showed similar results—namely, that CD4 counts are significantly predictive of survival. Also, the current value of CD4 was the most predictive aspect of the trajectory; other features of the path did not add significantly to the log-likelihood.

Therefore, the model we considered for the hazard relationship was the standard Cox model,

$$\lambda(t|\bar{Z}^*(t)) = \lambda_0(t)\exp(\beta Z^*(t)),$$

which we fit separately for patients randomized to receive placebo and for patients randomized to receive ZDV. We checked the adequacy of this model in two ways. First, we checked whether the log-linear relationship of the relative risk was appropriate. We did this by considering the higher-order quadratic model; that is,

$$\lambda(t|\bar{Z}^*(t)) = \lambda_0(t)\exp\{\beta_1 Z^*(t) + \beta_2(Z^*(t))^2\}.$$

Using a Wald test, we found that β_2 was not significantly different from zero (p value = .67 for patients randomized to ZDV and p value = .18 for patients randomized to placebo). Cubic terms were also fit with no significant increase in model fit.

Next, to check the adequacy of the proportional hazards assumption, we introduced an interaction term of time and CD4 count into our model. Namely, we considered the model

$$\lambda(t|Z^*(t)) = \lambda_0(t)\exp\{\beta_1 Z^*(t) + \beta_2 t Z^*(t)\}.$$

Using a Wald test, we found that β_2 was not significantly different from zero (p value = .63 for patients randomized to ZDV and p value = .12 for patients randomized to placebo).

Because all of the model diagnostics seem to indicate that the simple Cox model given by (6) was adequate, we obtained our final estimate of β by maximizing the partial likelihood (5) using (10) to compute $E_i(X_i, \beta)$. We computed the estimate of the underlying hazard rate $\lambda_0(t)$ by smoothing the Breslow estimate of cumulative hazard (Breslow 1974). In particular, the estimate of the cumulative underlying hazard $\hat{\Lambda}_0(t) = \int_0^t \lambda_0(u)\,du$ is given by

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \left[\Delta_i I(X_i \le t) \Big/ \left\{ \sum_{j=1}^{n} \hat{E}_j(X_i, \hat{\beta}) Y_j(X_i) \right\} \right],$$

and the estimate of hazard is given by

$$\hat{\lambda}_0(t) = \{ \hat{\Lambda}_0(t + h_1) - \hat{\Lambda}_0(t - h_2) \} / (h_1 + h_2), \quad (15)$$

where h_1 and h_2 were chosen experimentally to include sufficient failures in the window $(t - h_2)$ to $(t + h_1)$ to give a sufficiently smooth estimator. Hence the estimate of the hazard rate at time t as a function of the true CD4 count, $Z^*(t)$, is given by $\hat{\lambda}_0(t)\exp(\hat{\beta}Z^*(t))$.

To contrast this hazard relationship for patients randomized to placebo versus patients randomized to ZDV, we shall denote the hazard relationship for placebo patients as $\hat{\lambda}_0^P(t)\exp(\hat{\beta}^P Z^*(t))$ and that for ZDV patients as $\hat{\lambda}_0^Z(t)\exp(\hat{\beta}^Z Z^*(t))$. These data-analytic techniques were applied to the data from the Burroughs–Wellcome 02 clinical trial. Figure 4 shows the estimated relationship of the hazard function to CD4 counts at 6, 12, and 18 months after start of therapy for patients randomized to ZDV. This figure clearly shows that the hazard rate increases as CD4 declines, with the greatest effect occurring for patients with CD4 counts less than 50. It also demonstrates a substantial effect of time trend, that is, the hazard rate increases as a function of time from treatment initiation even after adjusting for CD4 values.

The estimated relationship of the hazard rate as a function of CD4 counts for patients randomized to placebo was computed at three months; results are shown in Figure 5. This study was closed after seven months because of a significantly decreased mortality for ZDV patients at that time. Because the patients randomized to placebo were then offered ZDV, we could only analyze the hazard relationship for placebo patients in the first seven months. For comparison, in Figure 5 we also plotted the hazard rate at three months for the patients randomized to receive ZDV.

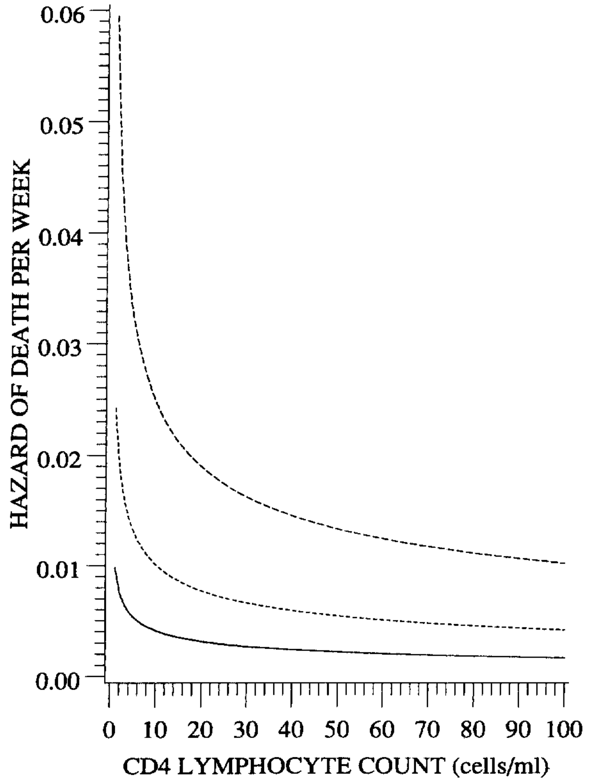

Figure 4. Estimated Hazard Rate as a Function of CD4 Counts for Patients Randomized to Receive ZDV. The solid line is at 6 months after randomization, the middle dashed line is at 12 months after randomization, and the outer dashed line is at 18 months after randomization.

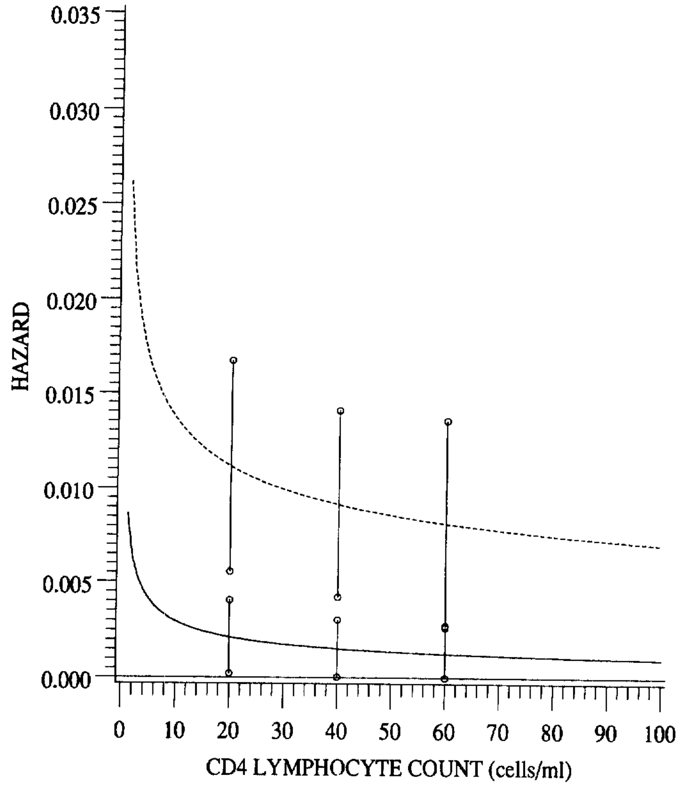

Figure 5. Estimated Hazard Rate as a Function of CD4 Counts at 3 Months After Randomization. The solid line is for patients randomized to receive ZDV; the dashed line is for patients randomized to placebo. Pointwise 95% confidence intervals for the estimated hazard rate at CD4 counts 20, 40, and 60 are also provided.

To get some sense of the precision of the hazard estimate given by (15), in Figure 5 we also computed the 95% confidence interval for the estimated hazard at CD4 counts 20, 40, and 60. We obtained the estimate for the asymptotic variance of (15) by computing

$$\text{var}(\hat{\lambda}_0(t)) = \text{var}\{ \hat{\Lambda}_0(t + h_1) - \hat{\Lambda}_0(t - h_2) \} / (h_1 + h_2)^2,$$

where $\text{var}\{ \hat{\Lambda}_0(t + h_1) - \hat{\Lambda}_0(t - h_2) \}$ was obtained using formulas derived for the estimated cumulative hazard from the proportional hazards model as given by Tsiatis (1981) and Andersen and Gill (1982). Because asymptotic normality should follow using the arguments of Ramlau-Hansen (1983), the 95% confidence interval is the estimate ± two standard errors.

Among patients randomized to ZDV, there were very few deaths prior to six months. Therefore, we cannot expect the relationship of the hazard rate as a function of CD4 counts at three months to be very precise. The shape of this relationship given in Figure 5 is mediated primarily by the relationship of the hazard ratio to CD4 counts at later times and the assumption of proportional hazards. Although we did a formal test of proportional hazards and failed to reject this assumption, we should still consider the curve in Figure 5 for ZDV patients as a rough extrapolation resulting from this assumption. More important than the actual shape of the curve is the fact that the hazard rate is substantially less for patients randomized to ZDV versus those randomized to placebo, regardless of CD4 counts.

6. ADJUSTING FOR MISSING DATA PATTERNS

The methods described so far make the assumption that the hazard rate at time t is not related to the timing of the visits prior to t. This assumption may not be appropriate. For example, patients who are sicker may be less likely to continue coming in for visits and may also have a higher risk of death. As suggested by one of the referees, we conducted a thorough analysis of this question. In particular, we considered the following model. We let $D(t)$ denote the length of time between t and the visit prior to t. Recall that the history of observed counts $\bar{\mathbf{Z}}(t)$ is given by $\{Z(t_1), \ldots, Z(t_j), t_j \le t\}$. Therefore, $D(t) = t - t_j$. We feel that $D(t)$ is the important feature of the timing of the visits that may be prognostic, in addition to the true history of CD4 counts. We now make the assumption that

$$\lambda(t \mid \bar{\mathbf{Z}}(t), \bar{Z}^*(t)) = \lambda_0(t)\exp\{(\beta_0 Z^*(t)) + \gamma D(t)\}. \qquad (16)$$

In other words, we allow the hazard at time t to be not only a function of the true CD4 counts, but also a function of the timing of the events. In this exercise we do not necessarily believe that (16) is the correct model, but rather consider it as a first-order extension of the model given in (4), which would correspond to $\gamma = 0$. This will allow us to examine the effect of nonrandom missingness of visits on the overall results.

Using the law of conditional probability, as derived in Section 2, we get that the hazard rate as a function of the observed CD4 history is given by

$$\begin{aligned}\lambda(t \mid \bar{\mathbf{Z}}(t)) &= \int \lambda(t \mid \bar{\mathbf{Z}}(t), \bar{Z}^*(t)) \\ &\quad \times f(\bar{Z}^*(t) \mid \bar{\mathbf{Z}}(t), X \ge t)\, d\bar{Z}^*(t) \\ &= \int \lambda_0(t)\exp(\gamma D(t))\exp(\beta_0 Z^*(t)) \\ &\quad \times f(\bar{Z}^*(t) \mid \bar{\mathbf{Z}}(t), X \ge t)\, d\bar{Z}^*(t) \\ &= \lambda_0(t)\exp(\gamma D(t)) E[e^{\beta_0 Z^*(t)} \mid \bar{\mathbf{Z}}(t), X \ge t].\end{aligned} \qquad (17)$$

To compute $E[e^{\beta_0 Z^*(t)} \mid \bar{\mathbf{Z}}(t), X \ge t]$, we must model the relationship of observed and true CD4 counts among individuals at risk at time t. Again we use growth curve models similar to those described in Section 4. But here we also allow the distribution of the true CD4 counts to be dependent on the amount of missingness, $D(t)$. Specifically, among individuals at risk at time t, we assume that for patients randomized to ZDV,

$$Z_i^*(u) = \psi_{0i} + \psi_{1i}u + \psi_{2i}(u-4)_+ + \delta_t^Z D_i(t); \qquad u \le t,$$

and for patients randomized to placebo,

$$Z_i^*(u) = \theta_{0i} + \theta_{1i}u + \delta_t^P D_i(t); \qquad u \le t.$$

The θ's and ψ's are random effects assumed normally distributed, as given in Section 3. In this model the degree of dependency of the true CD4 counts on the amount of missingness is given by the magnitude of the fixed-effect parameters δ_t^Z and δ_t^P. As in Section 3, we also assume that the observed CD4 count at time u is given by

$$Z_i(u) = Z_i^*(u) + e_i(u),$$

where $e_i(u)$ is random normal noise with mean zero and common variance.

These random-components models are computed separately at each death time among individuals at risk at these times. The estimates for δ_t^Z, δ_t^P at different death times t were generally negative, indicating that individuals with more missed visits were more likely to have lower CD4 counts. This corresponds with our intuition that the sicker individuals are more likely to miss visits.

Using this model, we can compute the conditional mean and variance of $Z^*(t)$ given $\bar{\mathbf{Z}}(t) = \{Z(t_i), \ldots, Z(t_j), D(t)\}$, which are denoted by $\mu(t \mid \bar{\mathbf{Z}}(t))$ and $\sigma^2(t \mid \bar{\mathbf{Z}}(t))$ using standard growth curve techniques as described by Laird and Ware (1982). Because all of the variables are jointly normal, we get that

$$\begin{aligned}E[e^{\beta Z^*(t)} \mid \bar{\mathbf{Z}}(t), X \ge t] \\ = \exp\{\beta\mu(t \mid \bar{\mathbf{Z}}(t)) + \beta^2\sigma^2(t \mid \bar{\mathbf{Z}}(t))/2\}.\end{aligned}$$

Therefore, the hazard rate as a function of the observed CD4 history, (17), is equal to

$$\begin{aligned}\lambda(t \mid \bar{\mathbf{Z}}(t)) = \lambda_0(t)\exp[\gamma D(t) + \beta\mu(t \mid \bar{\mathbf{Z}}(t)) \\ + \beta^2\sigma^2(t \mid \bar{\mathbf{Z}}(t))/2].\end{aligned} \qquad (18)$$

Because (18) is a proportional hazards model, we can derive estimates for γ as well as for β by maximizing the partial likelihood. We conducted this exercise separately for patients randomized to ZDV and patients randomized to placebo. In both cases the effect of delay was significant; that is, γ was significantly positive, with a p value $< .0001$ for ZDV patients and a p value of .02 for placebo patients.

We ultimately are interested in deriving estimates for the hazard rate as a function of the true CD4 count, $\lambda(t \mid Z^*(t))$. What we have derived at this point is an estimate of the relationship to both true CD4 count and the effect of delay, given by

$$\lambda(t \mid Z^*(t), D(t)) = \lambda_0(t)\exp\{\beta Z^*(t) + \gamma D(t)\}.$$

The law of conditional probability yields

$$\begin{aligned}\lambda(t \mid Z^*(t)) = \int \lambda(t \mid Z^*(t), D(t)) \\ \times f(D(t) \mid Z^*(t), X \ge t)\, dD(t),\end{aligned} \qquad (19)$$

and $f(D(t) \mid Z^*(t), X \ge t)$ equals

$$\frac{f(Z^*(t)\,|\,D(t), X \geq t) f(D(t)\,|\,X \geq t)}{\int f(Z^*(t)\,|\,D(t), X \geq t) f(D(t)\,|\,X \geq t)\, dD(t)}. \qquad (20)$$

The distribution of $Z^*(t)$ conditional on $D(t)$ for patients randomized to placebo is normal with mean $\theta_0 + \theta_1 t + \delta_t^P D(t)$ and variance $\sigma_{\theta_0}^2 + t^2\sigma_{\theta_1}^2 + 2t\sigma_{\theta_0\theta_1}$, and the distribution for patients randomized to ZDV is similar but with an extra parameter.

Using (19) and (20), we propose to estimate $\lambda(t\,|\,Z^*(t))$ by

$$\frac{\hat{\lambda}_0(t)\exp(\hat{\beta}Z^*(t)) \sum_{j=1}^{N(t)} \exp(\hat{\gamma}D_j) f(Z^*(t)\,|\,D_j) f(D_j)}{\sum_{j=1}^{N(t)} f(Z^*(t)\,|\,D_j) f(D_j)},$$

where $N(t)$ is the number of unique delay times among all individuals at risk at time t. The estimates $\hat{\beta}$, $\hat{\gamma}$ are obtained by the maximum partial likelihood estimates given by (18). The estimate $\hat{\lambda}_0(t)$ is obtained by smoothing the Breslow estimate of the underlying cumulative hazard, as described in Section 5. The densities $f(Z^*(t)\,|\,D_j)$ are the normal densities, which are functions of the parameter θ, ψ, δ, σ_θ, σ_ψ all estimated using the restricted maximum likelihood estimates of Laird and Ware. Finally, $f(D_j)$ is estimated using the sample probability functions of the delays among individual at risk at time t.

We shall refer to these estimates as the missing data adjusted hazard estimates. Figure 6 contrasts the adjusted hazard rates as a function of CD4 count at three months both for patients randomized to ZDV and for patients randomized to placebo. The missing data adjusted estimates of hazard rate were not substantially different from estimates that did not account for the effect of missing visits given in Figure 5.

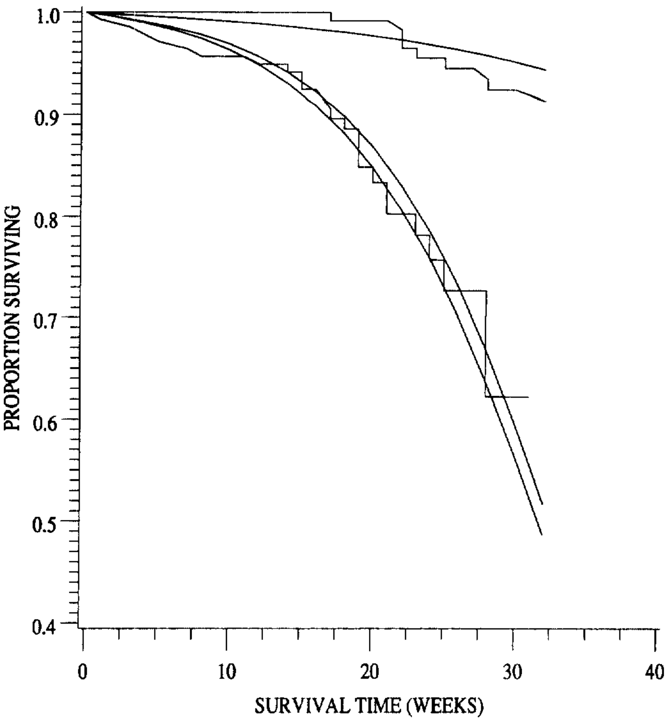

Figure 7. Observed and Predicted Survival Curves. The lower step function is the Kaplan–Meier estimate for patients randomized to placebo; the upper step function is the Kaplan–Meier estimate for patients randomized to receive ZDV. The three smooth curves are, from bottom to top, the predicted survival curves $\hat{S}^1_{AVE}(u)$, $\hat{S}^2_{AVE}(u)$, and $\hat{S}^3_{AVE}(u)$.

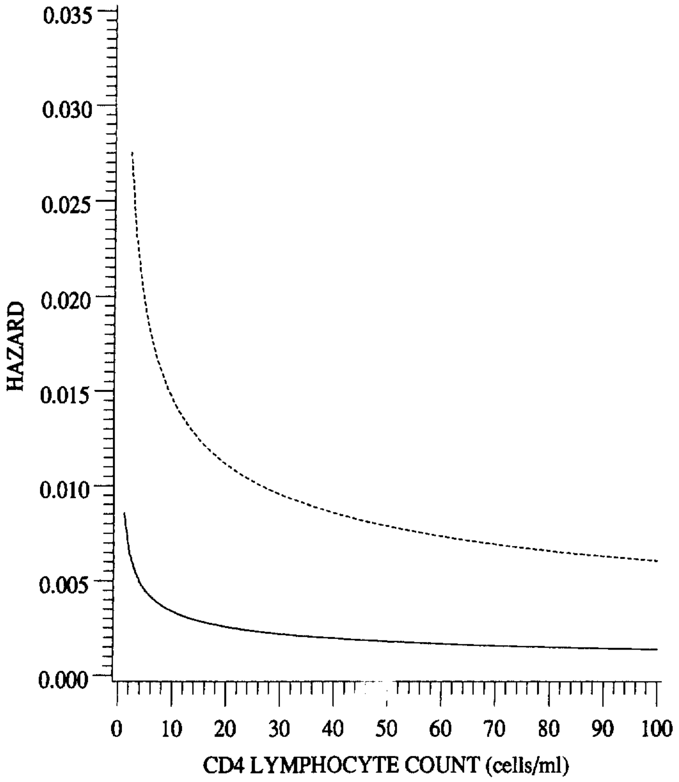

Figure 6. Estimated Hazard Rate as a Function of CD4 Counts at 3 Months After Randomization Adjusting for Nonrandom Missing Data. The solid line is for patients randomized to receive ZDV; the dashed line is for patients randomized to placebo.

7. CD4 COUNTS AS A SURROGATE FOR SURVIVAL

Figure 7 gives the Kaplan–Meier (1958) survival curves for patients randomized to receive ZDV versus patients randomized to receive placebo. There was significantly better survival for those patients randomized to ZDV, with a p value $< .0006$ (two-sided) using a log-rank test. Patients randomized to ZDV had a substantial increase in CD4 counts compared to patients randomized to placebo (see Fig. 1). Also, patients with higher CD4 counts had smaller hazard rates than those with lower CD4 counts (see Figs. 4, 5, and 6).

All of these facts together may lead clinical investigators to consider CD4 counts as a good surrogate marker for death. In Section 1 we gave the three conditions of Prentice (1982) that "a good surrogate marker" should have. We note that CD4 counts satisfy Conditions 1 and 2. However, Condition 3, which states that the hazard rate as a function of CD4 count should be the same for both treatment and placebo, is not satisfied. This is clearly illustrated in Figures 5 and 6.

To further illustrate the inadequacy of CD4 as a surrogate marker, we conducted the following exercise. Using an average CD4 trajectory for patients randomized to placebo, we computed the predicted survival curve for such a CD4 trajectory using the hazard relationship derived for placebo patients. (We use P and Z as superscripts to indicate placebo patients versus ZDV patients.) Specifically, we used $Z^{*P}_{\text{AVE}}(u) = \hat{\theta}_0 + \hat{\theta}_1 u$ as the average trajectory of log CD4 counts for placebo patients, where $\hat{\theta}_0$, $\hat{\theta}_1$ are the restricted maximum likelihood estimates of the population intercept and slope of the growth curve model. This is described in Section 4. We derived the estimated survival curve using

$$\hat{S}^1_{\text{AVE}}(u) = \exp\left\{-\int_0^u \hat{\lambda}_0^P(x)\exp(\hat{\beta}^P Z^{*P}_{\text{AVE}}(x))\,dx\right\},$$

as depicted in Figure 7. This predicted survival curve for the average trajectory of placebo patients is very close to the actual estimated survival curve for placebo patients.

To see how much survival benefit could have been predicted by just the increase in CD4 counts that were observed for patients randomized to receive ZDV, we computed the predicted survival curve for an average CD4 trajectory of a patient randomized to ZDV using the same hazard relationship as that derived for the placebo patients. That is, using

$$Z^{*Z}_{\text{AVE}}(u) = \hat{\psi}_0 + \hat{\psi}_1 u + \hat{\psi}_2(u-4)_+,$$

we computed

$$\hat{S}^2_{\text{AVE}}(u) = \exp\left\{-\int_0^u \hat{\lambda}_0^P(x)\exp(\hat{\beta}^P Z^{*Z}_{\text{AVE}}(x))\,dx\right\}.$$

This curve is also depicted in Figure 7. Even though there was a marked increase in CD4 counts among patients randomized to ZDV, this explained little of the increase in observed survival for these patients.

In contrast, if we compute the predicted survival curve for an average CD4 trajectory of ZDV patients using the hazard relationship derived for these patients, then

$$\hat{S}^3_{\text{AVE}}(u) = \exp\left\{-\int_0^u \hat{\lambda}_0^Z(x)\exp(\hat{\beta}^Z Z^{*Z}_{\text{AVE}}(x))\,dx\right\}.$$

This curve is also depicted in Figure 7. We note that this curve is very close to the estimated survival curve for the ZDV patients.

If CD4 counts serve as a useful surrogate marker, then much of the beneficial survival effect of ZDV should be explained through the effect on CD4 counts; therefore, we would expect $\hat{S}^2_{\text{AVE}}(u)$ to be close to $\hat{S}^3_{\text{AVE}}(u)$. This is clearly not the case. In fact, the increase in survival attributable to the increase in CD4 counts was very small compared to the actual differences. Most of the difference in the survival curves are therefore due to factors that are not captured by the value of CD4 count alone.

One important consideration in this research is the effect of missing data on the evaluation of CD4 as a surrogate. If patients with low CD4 counts are more likely to be missing measurements, this could lead to overestimates of the hazard of death at higher CD4 counts. To adjust for the effects of missing data, we proposed a model that allowed for the possibility that poor follow-up of CD4 counts may be associated with lower counts. The choice of a linear relationship between the true CD4 value and the time since last measurement of CD4 was not made because we know this to be an accurate description of the missingness process, but because this model constitutes a significant departure from the assumption of noninformative missingness. Our results indicate that departures of this type would not greatly alter the basic findings of our analyses.

8. CONCLUDING REMARKS

A referee expressed concern that using different shrinkage estimates in the two treatment groups may introduce bias in the treatment comparison. Specifically, the referee remarked that "if we shrink separately in the two groups, then the treatment group covariates will get shrunk to one value while the control group covariates will get shrunk to some other value. The differential shrinkage causes the covariates to be confounded with the treatment effect." Consequently, this confounding will make CD4 counts look less like a surrogate.

In Section 3 we outlined why replacing the covariates by their empirical Bayes estimates would reduce the bias of the regression parameter estimate. Because the bias introduced by measurement error may differ by treatment group, we feel that adjusting separately by treatment group provides a better estimate of the true treatment difference. Supporting evidence for this position is provided by empirical studies conducted by Dafni and Tsiatis (1994). In their simulation studies, data were generated according to the following model. The hazard rate was related to both the true CD4 counts and treatment by the proportional hazards model,

$$\lambda(t \mid \bar{Z}^*(t), \text{Trt}) = \lambda_0(t)\exp(\beta_1 Z^*(t) + \beta_2 \text{Trt}), \quad (21)$$

where Trt denotes treatment indicator. In (21), β_2 was set equal to zero to correspond to the situation where CD4 counts are a good surrogate marker, in that the treatment effect is evidenced entirely through the CD4 process. The observed CD4 counts followed a linear random effects model as described in Section 4, with varying amounts of measurement error.

We analyzed the simulated data using the observed CD4 counts as well as the two-stage method we described earlier. Specifically, in the two-stage method we replaced the covariates by their empirical Bayes estimates computed separately by treatment. Two main conclusions resulted from these simulation studies:

1. When the data were analyzed using the observed CD4 counts, the estimate of β_2 became increasingly different from zero with increasing measurement error.
2. When the data were analyzed using the empirical Bayes estimates, this bias was reduced substantially.

For example, in one set of simulations the parameters of the model were chosen to correspond closely to those derived from our analysis of the placebo controlled clinical trial conducted by Burroughs-Wellcome. With measurement error variance about 20% of the total variance, the average estimate of β_2 was $-.22$ when the analysis was based on the observed CD4 counts and $-.02$ when the analysis was based on the empirical Bayes estimates. This pattern was consistent across all simulation scenarios considered. We believe that the results of these simulation studies support the contention that the use of the two-stage method is more likely to find a good surrogate marker rather than less likely, as posited by the referee.

In conclusion, these methods permit us to estimate the relationship of hazard to CD4 counts, in the absence of measurement error and missing data. This work implies that CD4 counts may not serve as a useful endpoint for evaluating treatment effects in advanced patients. Because most of the treatment-related improvement in survival is not explained by increases in CD4 counts, even after adjustment for measurement error and missing data, clinical trials using CD4 counts as the primary endpoint may lead to erroneous conclusions regarding treatment effects on survival. We note that this is a much stronger conclusion than would be possible without such adjustment, because no marker measured with error could meet the Prentice conditions, although it would be possible for a marker measured with considerable error to provide reliable information about clinical effects of therapy. This work also has implications for investigating the biological action of antiviral drugs, through examination of the degree to which treatment benefits are mediated through markers. It is an important cautionary note that even markers that respond to therapy and are highly predictive of survival may be of little value in predicting treatment benefits on survival.

[Received August 1992. Revised June 1994.]

REFERENCES

Andersen, P. K., and Gill, R. D. (1982), "Cox Regression Model for Counting Processes: A Large-Sample Study," *The Annals of Statistics*, 10, 1100–1120.

Breslow, N. (1974), "Covariance Analysis of Censored Survival Data," *Biometrics*, 30, 89–99.

Cox, D. R. (1972), "Regression Models and Life Tables," *Journal of the Royal Statistical Society*, Ser. B, 34, 187–220.

——— (1975), "Partial Likelihood," *Biometrika*, 62, 269–276.

Dafni, U. G., and Tsiatis, A. A. (1994), "Evaluating Surrogate Markers of Clinical Outcome when Measured with Error," unpublished manuscript.

Kaplan, E. L., and Meier P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481.

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

Lindstrom, M. J., and Bates, D. M. (1988), "Newton–Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data," *Journal of the American Statistical Association*, 83, 1014–1022.

Prentice, R. L. (1982), "Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model," *Biometrika*, 69, 331–342.

——— (1989), "Surrogate Endpoints in Clinical Trials: Definition and Operational Criteria," *Statistics in Medicine*, 8, 431–440.

Prentice, R. L., and Self, S. G. (1983), "Asymptotic Distribution Theory for Cox-Type Regression Models with General Relative Risk Form," *The Annals of Statistics*, 11, 804–813.

Ramlau-Hansen, H. (1983), "Smoothing Counting Process Intensities by Means of Kernel Functions," *The Annals of Statistics*, 11, 453–466.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, New York: John Wiley.

Self, S. G., and Liang, K. Y. (1987), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," *Journal of the American Statistical Association*, 82, 605–610.

Self, S. G., and Pawitan, Y. (1992), "Modeling a Marker of Disease Progression and Onset of Disease," in *Statistical Methodology for Study of the AIDS Epidemic*, eds. K. Dietz, V. Farewell, and N. P. Jewell, Boston: Birkhäuser.

Tsiatis, A. A. (1981), "A Large-Sample Study of Cox's Regression Model," *The Annals of Statistics*, 9, 93–108.