Statistical Methods in Medical Research 1998; **7**: 28–48

Increasing efficiency from censored survival data by using random effects to model longitudinal covariates

Joseph W Hogan Center for Statistical Sciences, Brown University, Providence, Rhode Island, USA and **Nan M Laird** Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA

When estimating a survival time distribution, the loss of information due to right censoring results in a loss of efficiency in the estimator. In many circumstances, however, repeated measurements on a longitudinal process which is associated with survival time are made throughout the observation time, and these measurements may be used to recover information lost to censoring. For example, patients in an AIDS clinical trial may be measured at regular intervals on CD4 count and viral load.

We describe a model for the joint distribution of a survival time and a repeated measures process. The joint distribution is specified by linking the survival time to subject-specific random effects characterizing the repeated measures, and is similar in form to the pattern mixture model for multivariate data with nonignorable nonresponse. We also describe an estimator of survival derived from this model.

We apply the methods to a long-term AIDS clinical trial, and study properties of the survival estimator. Monte Carlo simulation is used to estimate gains in efficiency when the survival time is related to the location and scale of the random effects distribution. Under relatively light censoring (20%), the methods yield a modest gain in efficiency for estimating three-year survival in the AIDS clinical trial. Our simulation study, which mimics characteristics of the clinical trial, indicates that much larger gains in efficiency can be realized under heavier censoring or with studies designed for long term follow up on survival.

1 Introduction

Studies which measure an individual subject's outcome repeatedly over time are a common phenomenon in medicine and public health. These repeated measures can be used to fulfil a variety of study objectives, and often involve the simultaneous analysis of a survival time variable. In many settings, changes in the repeated measures over time are used to define the primary endpoint of the study. In this setting, failure to account for dropouts from the study may lead to erroneous inferences.[1,2] Alternatively, changes in the repeated measure are sometimes used as a surrogate measure for the event of major interest such as time to death or disease progression.[3,4] If there is censoring of the time to event, the repeated measures may also be used to increase efficiency of the estimate of the survival distribution.[5,6]

In this paper we will discuss a particular random effects model for repeated measures that can be used in either setting. The novel feature of this model which distinguishes it from the usual random effects model[7] is that the random effects are used as a device for constructing the joint distribution of the repeated measures and the associated survival time variable. In particular, the distribution of the random effects is modelled as a mixture over the survival distribution. When the outcome of primary

Address for correspondence: JW Hogan, Center for Statistical Sciences, Brown University, Box G–H, Providence, RI 02912, USA. E-mail: jhogan@stat.brown.edu

 0962-2802(98)SM150RA

interest is the repeated measures, the survival time variable is time to study dropout; when the survival time variable is the primary endpoint, for example time to death or disease progression, the repeated measures may be used as auxiliary variables to provide additional information on censored cases. In the context of surrogate markers, the repeated measures may be treated as time-varying covariates measured with error, with the unobserved random effects determining the true covariate value.

The traditional variance component model evolved as a convenient way to model multiple sources of error in the context of designed experiments with many factors. A simple version with only two components has enjoyed much popularity in repeated measures analysis, even though the model's appropriateness is often questioned.[8] With longitudinal data, where the repeated measures are taken serially on the same subject, at least two sources of variability are readily identified: subject-to-subject and within-subject. Often, subject-to-subject variability is modelled by a vector of correlated random subject effects, and the term *co*variance components is sometimes used to describe these models. Diggle[9] proposed a further partitioning of the within-subject variance by distinguishing between an autocorrelated random error and pure measurement noise. Multiple levels of nesting of subjects can be handled using additional random error terms[10].

The use of random effects – or covariance components – to analyse repeated measures is attractive for several reasons. It easily accommodates unbalanced designs, especially regarding the timing and frequency of observations; it enables explicit partitioning of the variability, which can be useful for planning purposes; and it allows one to get estimates of individual effects. From the point of view of developing models for the joint distribution of repeated measures and event times, it offers an attractive mechanism for modelling the error in the measurement of the repeated measures and for allowing the event time to depend on the underlying 'error-free' trajectory.[3,11–14]

This paper is organized as follows. In Section 2, we introduce notation and provide background on using random effects for modelling the joint distribution of repeated measures and a survival time variable. The remainder of the paper focuses on a particular aspect of the model, namely using the repeated measures to increase efficiency of the estimate of the survival distribution. In Section 3 we give a brief review of other approaches to the problem of using auxiliary information to increase efficiency of an estimated survival distribution in the presence of independent right censoring. In Section 4 we present a particular random effects model for handling repeated measures and event times jointly, and derive expressions for the likelihood; we indicate methods for estimating parameters and standard errors, including a full maximum likelihood approach using the EM algorithm and a less efficient approximation to the full likelihood approach which is not as computationally intensive. In Section 5 we apply the model to data from an AIDS clinical trial, and in Section 6 we describe the results of a simulation study in which we estimate efficiency gains under various forms of dependence between the repeated measures and the survival time and for different rates of censoring. We close with some discussion and indicate directions for further development and investigation.

2 Notation and background

We assume there are n independent study subjects, and the study protocol calls for following each subject for a fixed length of time T; subjects may of course enter at different calendar times. Observations y_{ij} are made repeatedly on the ith subject at times t_{ij}. In many cases, the repeated measures are not taken throughout the study, or are taken at less frequent intervals as the study progresses (the above-mentioned AIDS trial is one example). Let $\tilde{T} \leq T$ represent the maximum amount of time during which repeated measures will be taken.

Usually, the study protocol prespecifies a vector of common times, so that $t_{ij} = t_j$ for $j = 1, \ldots, m$, where $t_1 = 0$ and $t_m = \tilde{T}$; however, due to mistiming or missing observations, the actual times and number of measurements may vary from subject to subject. We denote by n_i the number of measurements on the ith subject, and let $\mathbf{y}_i$ denote the $n_i \times 1$ vector of observations for the ith subject. In addition, we assume that each person has an event time $u_i > 0$ which may be right censored by c_i. Measurements are not made on $\mathbf{y}_i$ after censoring (i.e. $t_{i,n_i} \leq c_i$ when $c_i < u_i$); however, repeated measures sometimes continue to be taken after the event. For example, CD4 cell counts and other virologic measures may be recorded after onset of AIDS-defining illness. In the case where $\mathbf{y}_i$ denotes an auxiliary measure, u_i denotes the event which is the main outcome of interest (e.g. death or disease progression); u_i may be censored at the end of follow-up, at T, or prior to T (as in the case of loss-to-follow-up or study dropout for reasons unrelated to u_i). In clinical trials with long follow up times, it is common to have an interim analysis; in this case, staggered entry will induce an independent censoring process.

In the case where $\mathbf{y}_i$ is the primary outcome of interest, u_i might be the time of outcome-related dropout from the study. Dropouts due to factors unrelated to outcome should then be treated as 'censored' at the time of dropout. For example, in a clinical trial, removal from protocol because of disease progression would be considered an event (u_i), whereas removal from protocol due to side-effects of treatment unrelated to treatment response might be considered censoring at the time of removal. Thus, for subject i, we observe the triplet $(\mathbf{y}_i, \tilde{u}_i, \xi_i)$, where $\tilde{u}_i = \min\{u_i, c_i\}$ and $\xi_i = 1$ if $\tilde{u}_i = u_i$.

Our objective here is to describe a model for the joint distribution of $(\mathbf{y}_i, u_i)$ which can estimated from the observed data. Numerous investigators have considered models which can be used in this setting, and in our brief review we discuss models which treat the repeated measures as multivariate normal and the u_i as a continuous random variable. Little[15] discusses a similar class of models in the context of nonignorable missingness, where u_i refers to a pattern of missingness observed in a multivariate repeated measures outcome. Hogan and Laird[16] give a more general overview of these models, including cases where the $\mathbf{y}_i$ are not restricted to be continuous random variables.

The models we consider here build on the random effects model for $\mathbf{y}_i$ described by Laird and Ware,[7] which can be motivated as follows. First assume that each individual has a curve characterizing change in y_{ij} over time, which can be modelled linearly as

$$\mathbf{y}_i = \mathbf{Z}_i \boldsymbol{\beta}_i + \mathbf{e}_i \tag{2.1}$$

for some proper choice of $\mathbf{Z}_i$ and $\boldsymbol{\beta}_i$. Here $\mathbf{Z}_i$ will be a function of the t_{ij} values and $\boldsymbol{\beta}_i$ is

an individual's coefficient vector. The $\mathbf{e}_i$ are taken to have zero mean and variance $\sigma^2\mathbf{I}_{n_i}$. We view the coefficients $\boldsymbol{\beta}_i$ as random variables with a mean vector which depends on subject characteristics and variance–covariance matrix $\boldsymbol{\Phi}$. Thus we write

$$\boldsymbol{\beta}_i = \mathbf{A}_i\boldsymbol{\alpha} + \mathbf{b}_i \tag{2.2}$$

where $\mathbf{A}_i$ is a matrix which includes subject specific covariates, $\boldsymbol{\alpha}$ is the population mean vector and $\mathbf{b}_i$ is a $N(\mathbf{0}, \boldsymbol{\Phi})$ random variable which represents the ith subject's deviation from the mean population curve $\mathbf{A}_i\boldsymbol{\alpha}$. These two assumptions imply that $\mathbf{y}_i$ can be written as the sum of fixed and random effects

$$\mathbf{y}_i = \mathbf{Z}_i\mathbf{A}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i$$

where $\boldsymbol{\alpha}$ are the fixed effects and $\mathbf{b}_i$ and $\mathbf{e}_i$ are zero-mean error terms which model between subject ($\mathbf{b}_i$) and within subject ($\mathbf{e}_i$) errors. If we assume that the $\mathbf{b}_i$ values and the $\mathbf{e}_i$ values are independent normals, then the marginal distribution of $\mathbf{y}_i$ is completely specified as $N(\mathbf{X}_i\boldsymbol{\alpha}, \mathbf{Z}_i\boldsymbol{\Phi}\mathbf{Z}_i^T + \sigma^2\mathbf{I}_{n_i})$, where $\mathbf{X}_i = \mathbf{Z}_i\mathbf{A}_i$.

An attractive feature of the random effects model is that we can use the random subject effect $\boldsymbol{\beta}_i$ as a mechanism for modelling the joint distribution of $\mathbf{y}_i$ and u_i. The general idea is to specify a model for the joint distribution of $\boldsymbol{\beta}_i$ and u_i, so that integrating over $\boldsymbol{\beta}_i$ will yield a joint distribution for $(\mathbf{y}_i, u_i)$. Historically, two basic approaches are used to model the joint distribution of $(\boldsymbol{\beta}_i, u_i)$: one specifies $f(u_i \mid \boldsymbol{\beta}_i)$ and $f(\boldsymbol{\beta}_i)$ and the other uses $f(\boldsymbol{\beta}_i \mid u_i)$ and $f(u_i)$. The first approach has been termed the selection model,[17] since implicit in the model is the notion that individuals are selected out of the study (at time u_i) in a way which depends on their subject specific characteristics ($\boldsymbol{\beta}_i$). In this case, a key assumption is that $f(u_i \mid \boldsymbol{\beta}_i, \mathbf{y}_i) = f(u_i \mid \boldsymbol{\beta}_i)$. The second approach is termed the mixture model,[1] since implicit in this model is the notion that the distribution of the repeated measures depends on a subject characteristic (dropout) which may not be of direct interest; this model can be viewed as a mixture of distributions induced by dropout time. Here the key assumption usually made is that $f(\mathbf{y}_i \mid \boldsymbol{\beta}_i, u_i) = f(\mathbf{y}_i \mid \boldsymbol{\beta}_i)$.

We now describe some specific examples of mixture and selection models. Wu and Carroll[11] used the selection model approach in the setting where

$$\mathbf{Z}_i = \begin{bmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{bmatrix}$$

and the distribution of u_i is specified on a set of $\mathcal{J}$ intervals defined by $t_0^* < t_1^* < \cdots < t_{\mathcal{J}}^*$ as $(0 \leq u_i \leq t_j^* \mid \boldsymbol{\beta}_i) = M(\boldsymbol{\psi}^T\boldsymbol{\beta}_i, t_j^*)$. In other words, the distribution of u_i depends on $\boldsymbol{\beta}_i$ only through the linear combination $\boldsymbol{\psi}^T\boldsymbol{\beta}_i$, where $\boldsymbol{\psi}$ is a vector of unknown regression parameters. Models suggested for $M(\cdot, \cdot)$ include proportional hazards, logistic and probit regression. Notice that the distribution of u_i given $\boldsymbol{\beta}_i$ is independent of $\mathbf{y}_i$. Using the normality assumption for $(\mathbf{b}_i, \mathbf{e}_i)$, the distribution of the data is completely specified when $M(\cdot, \cdot)$ is modelled as logistic or probit; an additional assumption for the baseline hazard is needed if the proportional hazards model is used. Wu and Carroll[11] suggest using marginal maximum likelihood with the probit model in order to make the computations tractable. Even then, full maximum likelihood for $(\boldsymbol{\alpha}, \boldsymbol{\psi})$

and the variance–covariance components $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\Phi})$ is numerically intensive. They suggest an approximation whereby $\boldsymbol{\theta}$ is first estimated via a method-of-moments from each individual's OLS estimates $\hat{\boldsymbol{\beta}}_i = (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{y}_i$, and then assumed fixed for the purpose of maximizing the likelihood to estimate $\boldsymbol{\psi}$ and $\boldsymbol{\alpha}$.

This work has been extended by Tsiatis *et al.*,[3] Self and Pawitan[13] and Wulfsohn and Tsiatis,[14] who were primarily interested in using the model to assess the dependence of failure time on a covariate process, say $v_i(u)$, which is measured with error at discrete times $t_{i_1}, \ldots, t_{i,n_i}$. Each uses a proportional hazards model

$$\lambda_i(u) = \lambda_0(u)\exp[\psi_0 + \psi_1 v_i(u)]$$

to specify the dependence, where $\lambda_0(\cdot)$ is left unspecified, $\boldsymbol{\psi}^T = (\psi_0, \psi_1)$ are the regression parameters as before, and $v_i(u)$ is a linear function of $\boldsymbol{\beta}_i$. The observable data $\mathbf{y}_i$ are again related to $\boldsymbol{\beta}_i$ by the model specified in equations (2.1) and (2.2). In this case, the estimated parameters include $\boldsymbol{\alpha}, \boldsymbol{\psi}, \boldsymbol{\theta}$ and the underlying hazard function λ_0, which is estimated at each distinct failure time. Wulfsohn and Tsiatis[14] apply the EM algorithm[18] to maximize the likelihood based on the joint distribution of $\mathbf{y}_i$ and u_i, again permitting u_i to be subject to censoring. One difficulty in applying full ML estimation to this problem using EM is slow convergence, especially in the variance components. Faucett and Thomas[19] take a Bayes approach and use Gibbs sampling to compute posterior means and variances.

De Gruttola and Tu[12] and Schuchter[20] independently suggested a different version of this model which assumes that u_i (or some transformation of u_i) is normal with mean $\boldsymbol{\psi}^T\boldsymbol{\beta}_i$ and variance τ^2. This implies that $(\mathbf{y}_i, u_i)$ are jointly multivariate normal, but with a mean and variance–covariance matrix which are nonlinear functions of $\boldsymbol{\alpha}, \boldsymbol{\psi}, \tau^2$ and $\boldsymbol{\theta}$. The EM can be used to obtain MLEs of all the parameters even when the u_i are subject to independent right censoring. Notice that because $(u_i, \boldsymbol{\beta}_i)$ is bivariate normal, the conditional distribution of $\boldsymbol{\beta}_i$ given u_i is again normal, with a mean which depends linearly on the observed u_i. Further, since $\mathbf{y}_i = \mathbf{Z}_i\boldsymbol{\beta}_i + \mathbf{e}_i$, the conditional distribution of $\mathbf{y}_i$ given u_i is normal, with a mean depending linearly on u_i, so that this model can be viewed as either a selection model or a mixture model.

Wu and Bailey[21,22] used this last representation to motivate what they called the 'conditional linear mean model' (CLMM), whereby the OLS estimate $\hat{\boldsymbol{\beta}}_i$ is modelled as a linear function of u_i and the covariates in $\mathbf{A}_i$, with error variance equal to $\sigma^2(\mathbf{Z}_i^T\mathbf{Z}_i)^{-1} + \boldsymbol{\Phi}$. Denote the conditional mean of $\hat{\boldsymbol{\beta}}_i$, given $(u_i, \mathbf{A}_i)$, as $\mathbf{W}_i\boldsymbol{\delta}$. They propose a weighted regression to estimate δ using method-of-moment estimators for $\boldsymbol{\theta}$. An advantage of their approach is that it does not require specifying a distribution for u_i (because the analysis is conditional on the observed u_i), but a drawback is that it cannot handle cases with censored u_i in the regression analysis.

Because the regression in the CLMM is conditional on u_i, none of the regression coefficients $\boldsymbol{\delta}$ corresponds directly to $\boldsymbol{\alpha}$ from (2.2); to estimate the unconditional mean of the repeated measures, one must integrate over the distribution of u_i (hence the connection to mixture models). Hogan and Laird[23] proposed a generalization which uses a nonparametric distribution for u_i and permits censoring. Assuming that the estimated distribution of u_i puts mass only at observed event times, it is easy to write down a likelihood whose unknown parameters consist of the regression coefficients in the mean of $\mathbf{y}_i$ given both u_i and the covariates in $\mathbf{A}_i$, $\boldsymbol{\delta}, \boldsymbol{\theta}$ and the mass points in the

distribution of u_i at the observed death times. The likelihood readily accommodates censored u_i (if no u_i are censored, the estimated survival distribution is simply the empirical cdf). The estimation methods are directly related to more general ones described by Redner and Walker;[24] that is, one can view the repeated measures data as arising from a mixture density where the mixing components are only partially labelled and the mixing proportions are unknown. Redner and Walker also discuss conditions under which parameter estimates are identifiable from the data, an important issue in drawing inference from incomplete multivariate data. Treating the repeated measures as outcomes from a mixture density in which the event times are the mixture components is the basis of the approach described in Section 4.

3 Using auxiliary variables to improve survival estimation

Several researchers have studied methods for using auxiliary variables to recover information on censored individuals in survival studies. The methods can be useful for interim analyses in randomized trials, where censoring due to late entry may be heavy but can usually be regarded as noninformative. Cox[25] outlines a likelihood-based parametric method for using auxiliary information in an exponential regression survival model, and gives expressions for the proportion of information recovered. Lagakos,[26] Finkelstein and Schoenfeld[27] and Gray[28] consider the case where the auxiliary variable is also a failure time variable assumed to be related to the one of primary interest.

Malani[5] and Murray and Tsiatis[6] investigate the use of discrete longitudinal covariate processes to increase efficiency in the Kaplan–Meier (KM) estimator[29] of a survival distribution (and hence increase power in test statistics based on the estimator). They describe the properties of weighted KM estimators of the form $\hat{S}(t) = \sum_k w_k \hat{S}_k(t)$, where $\hat{S}_k(t)$ is the KM estimator for those with covariate profile k, and w_k is the proportion of subjects taking that path. Murray and Tsiatis[6] demonstrate that modest efficiency gains can be made in some practical situations. More recently, Murray and Tsiatis[30] have studied properties of test statistics based on their weighted KM estimator. Pepe and Fleming[31] describe a class of test statistics based on the weighted KM estimator; Robins and Rotnitzky[32] develop a class of estimators for regression models which can be applied to censored survival data when surrogate markers are available. Within this context, they derive a modified logrank test which is more powerful than the ordinary logrank test in some cases. Slud and Korn[33] give semiparametric methods for both testing and estimation in which subjects are stratified on a 'response indicator' which is observed shortly after randomization and which is related to survival. Jewell and Kalbfleisch[34] review several parametric and semi-parametric methods for using longitudinal prognostic information in survival analysis, and provide analytic expressions for the efficiency gains in some special cases.

Fleming *et al.*[4] develop a nonparametric approach which is similar to that of Cox[25] in that it uses an 'augmented likelihood' for the parameter in the survival time regression model. Here the auxiliary variable $\mathbf{y}_i$ can be vector valued and time varying, and inferences are desired about the distribution of u_i given covariates $\mathbf{Z}_i$. Letting ψ represent the regression parameter for survival, their likelihood function for ψ is

$$L(\psi) = \prod_{\xi_i=1} P_\psi(u_i|\mathbf{Z}_i) \prod_{\xi_i=0} P_\psi(U > u_i^*) P_\psi(\mathbf{y}_i \mid U > u_i^*, \mathbf{Z}_i)$$

Fleming *et al*. assume a proportional hazards model for the covariates and use the Breslow estimator of the hazard, which puts mass only at the observed event times. Since

$$P_\psi(\mathbf{y}_i \mid U > u_i^*, \mathbf{Z}_i) = \int_{u_i^*}^{\infty} P_\psi(u \mid U > u_i^*, \mathbf{Z}_i) P(\mathbf{y}_i \mid u, \mathbf{Z}_i)\, \mathrm{d}u$$

in order to estimate $P_\psi(\mathbf{y}_i \mid U > u_i^*, \mathbf{Z}_i)$ they need only to estimate $P(\mathbf{y}_i \mid u, \mathbf{Z}_i)$ at the observed death times. This is easily done nonparametrically when the sample space for $\mathbf{y}_i$ is small. Our approach is similar to Fleming *et al*. except we assume a parametric model for $P(\mathbf{y}_i \mid u_i)$ and a full likelihood-based analysis which incorporates information on $\mathbf{y}_i$ for uncensored subjects.

4 Random effects model for *f*(y, *u*)

Here we describe a modification to the mixture model described in Section 2 for the joint distribution of the repeated measures $\mathbf{y}_i$ and the possibly right censored survival time u_i. Using the unobserved random effects characterizing $\mathbf{y}_i$, the model links $\mathbf{y}_i$ to u_i via

$$f(\mathbf{y}_i, u_i) = \int f(\mathbf{y}_i|\boldsymbol{\beta}_i, u_i)\, f(\boldsymbol{\beta}_i|u_i)\, f(u_i)\, \mathrm{d}\boldsymbol{\beta}_i$$

where $\boldsymbol{\beta}_i$ are the random effects as in (2.1). Also as in (2.1), the observed repeated measurements consist of the vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{i,n_i})^{\mathrm{T}}$, where y_{ij} is observed at time t_{ij}. Subjects can have different numbers of measurements, and the set of measurement times need not be common across subjects.

We consider a particular parameterization of association between $\mathbf{y}_i$ and u_i in which the repeated measures distribution is specified as a two-component mixture of normals. The mixture components are defined by the binary indicator of survival beyond a prespecified time τ. The choice of τ can be motivated by a desire to obtain a point estimate of $S(\tau)$. In an interim analysis, for example, τ can be elapsed time from the start of the study to the time of analysis. In a completed trial, τ might correspond to a follow up time near the end of the trial.

Specifically, we assume that

$$[\mathbf{y}_i \mid \boldsymbol{\beta}_i, u_i] \sim N\big(\mathbf{Z}_i\boldsymbol{\beta}_i(u_i), \sigma^2(u_i)\mathbf{I}_{n_i}\big)$$

where $\sigma^2(u) = \sigma_a^2\, I\{u \le \tau\} + \sigma_b^2\, I\{u > \tau\}$. The distribution of the random effects also is a two-part mixture depending on u_i. To write the model as a regression in which $E(y_{ij}|u_i)$ is a linear function of time, define

$$\mathbf{W}_i = \mathbf{W}(u_i) = \begin{bmatrix} 1 & I\{u_i > \tau\} & 0 & 0 \\ 0 & 0 & 1 & I\{u_i > \tau\} \end{bmatrix}$$

and let $\boldsymbol{\alpha} = (\alpha_{0a}, \Delta_0, \alpha_{1a}, \Delta_1)^{\mathrm{T}}$. Then

$$[\boldsymbol{\beta}_i \mid u_i] = \mathbf{W}_i\boldsymbol{\alpha} + \mathbf{b}_i$$

where $\mathbf{b}_i$ have a bivariate normal distribution with zero mean and variance $\boldsymbol{\Phi}_a I\{u \leq \tau\} + \boldsymbol{\Phi}_b I\{u > \tau\}$. In this notation, α_{0a} and α_{1a} are, respectively, the intercept and slope for those with $u_i < \tau$; the vector $\boldsymbol{\Delta} = (\Delta_0, \Delta_1)^{\mathrm{T}}$ is the difference in the mean of $\boldsymbol{\beta}_i$ between those with $u \leq \tau$ and $u > \tau$. This model formulation can be used to obtain directly the conditional distribution of $\mathbf{y}_i$ given u_i, namely

$$[\mathbf{y}_i \mid u_i] = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i$$

where $\mathbf{X}_i = \mathbf{Z}_i\mathbf{W}_i$ and $\mathbf{e}_i \sim N(\mathbf{0}, \sigma^2(u_i)\mathbf{I}_{n_i})$.

This model can be viewed as a version of the pattern mixture model described by Little[1] for modelling incomplete multivariate data. In applying this model to survival estimation, our objective is to estimate the mixing distribution from partially observed (censored) data; in missing data applications, the mixing distribution is estimated from completely observed missing data indicators and is usually regarded as a nuisance parameter. Other model formulations are possible; following Wu and Bailey[22] for example, one might assume that the random effects are polynomial functions of u; such structures are easily incorporated through proper specification of $\mathbf{W}_i$.

We turn now to parameterization of $f(u)$. Cox and Oakes[35] (pp 175–77) describe the KM estimator in terms of a multinomial distribution. Their description will prove useful as we investigate the relative efficiencies of the unweighted KM estimator and the weighted estimator derived from the mixture model. Let the unique ordered observed failure times be $u_1^*, u_2^*, \ldots, u_m^*$, so that the set of labels (subscripts) is $\{1, 2, \ldots, m\}$. For an observed failure at u_j^*, let $\mathcal{U}_i = \{j\}$. For a censored observation at u_j^* or in the interval $[u_j^*, u_{j+1}^*)$, let $\mathcal{U}_i = \{j+1, j+2, \ldots, m\}$. Thus, the observed outcome for subject i can be represented as a proper subset of the labels $\{1, 2, \ldots, m\}$. This multinomial model is parameterized by the vector $\boldsymbol{\pi}^{\mathrm{T}} = (\pi_1, \pi_2, \ldots, \pi_m)$, where $\pi_j = \mathrm{P}\{\text{failure at } u_j\}$. The observed data log-likelihood for $\boldsymbol{\pi}$ is

$$\ell(\boldsymbol{\pi}; \mathcal{U}) = \sum_{i=1}^{n} \log\left(\sum_{j \in \mathcal{U}_i} \pi_j\right)$$

where $\mathcal{U} = \{\mathcal{U}_1, \ldots, \mathcal{U}_n\}$.

In the model for joint distribution, missingness in u_i (e.g. censoring) is *missing at random* (MAR) in the Little and Rubin[36] hierarchy. This implies that censoring is independent of failure time conditional on observed data $\mathbf{y}_i$, but that censoring may be dependent on u_i unconditional on $\mathbf{y}_i$. This suggests that our model also has potential utility for correcting bias in the KM estimator induced by dependent censoring. The ML estimate of $\boldsymbol{\pi}$ derived from the joint distribution model can be viewed as a weighted average over the sample space of $\mathbf{y}$, similar to the weighted KM estimators described in Section 3.[5,6]

One can express the conditional distribution of u_i given $\boldsymbol{\beta}_i$ through a logistic regression model implied by our specification of $f(\mathbf{y}_i, u_i)$. Consider the case where $\sigma_a^2 = \sigma_b^2$. We have

$$\log\left[\frac{\text{pr}(u > \tau \mid \boldsymbol{\beta}_i)}{\text{pr}(u \le \tau \mid \boldsymbol{\beta}_i)}\right] = \log\left[\frac{S(\tau)}{1-S(\tau)}\right] + \frac{1}{2}\log(|\boldsymbol{\Phi}_a|/|\boldsymbol{\Phi}_b|) + \frac{1}{2}(\boldsymbol{\beta}_i - \boldsymbol{\alpha}_a)^{\mathrm{T}}\boldsymbol{\Phi}_a^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\alpha}_a) - \frac{1}{2}(\boldsymbol{\beta}_i - \boldsymbol{\alpha}_b)^{\mathrm{T}}\boldsymbol{\Phi}_b^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\alpha}_b)$$

This expression shows that when estimating $\boldsymbol{\pi}$ from censored data using the mixture model, relative weight is assigned to $\text{pr}(u > T)$ using the underlying survival function and by comparing the normalized distances from $\boldsymbol{\beta}_i$ to $\boldsymbol{\alpha}_a$ and $\boldsymbol{\alpha}_b$, respectively. Also evident is that the $\boldsymbol{\beta}_i$ can be viewed as random effects which induce heterogeneity in both the survival and repeated measures distributions (see Lancaster and Intrator[37] for an example of using a common random effect to induce dependence between a Poisson process and a survival time). A similar expression can be derived for the conditional distribution of u_i given $\mathbf{y}_i$. For simplicity, assume that $\boldsymbol{\Phi}_a = \boldsymbol{\Phi}_b = \boldsymbol{\Phi}$ and that $\sigma_a^2 = \sigma_b^2 = \sigma^2$. Let $\boldsymbol{\Sigma}_i = \mathbf{Z}_i\boldsymbol{\Phi}\mathbf{Z}_i + \sigma^2\mathbf{I}_{n_i}$. Then

$$\begin{aligned}\log\left[\frac{\text{pr}(u > \tau \mid \mathbf{y}_i)}{\text{pr}(u \le \tau \mid \mathbf{y}_i)}\right] &= \log\left[\frac{S(\tau)}{1-S(\tau)}\right] \\ &\quad + \frac{1}{2}(\mathbf{Z}_i\boldsymbol{\alpha}_a)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{Z}_i\boldsymbol{\alpha}_a) - \frac{1}{2}(\mathbf{Z}_i\boldsymbol{\alpha}_b)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{Z}_i\boldsymbol{\alpha}_b) + (\mathbf{Z}_i\boldsymbol{\Delta})^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{y}_i \\ &= \log\left[\frac{S(\tau)}{1-S(\tau)}\right] + \gamma_0 + \sum_{j=1}^{n_i}\gamma_j y_{ij}\end{aligned}$$

which is linear in $\mathbf{y}_i$ on the log-odds scale. Note that $\mathbf{Z}_i\boldsymbol{\Delta}$ is the expected difference in $\mathbf{y}_i$ between those who fail before and after τ.

The EM algorithm can be employed to find the ML estimate of $\boldsymbol{\pi}$ because the complete-data log-likelihood for $\boldsymbol{\pi}$ is a linear function of sufficient statistics. Although the EM is never used in practice to obtain the KM estimator, it is useful to review the derivation here because it extends easily to accommodate information on surrogate markers when iterative methods are necessary.

Let K_{ij} be an indicator of failure of subject i at time u_j, let $K_j = \sum_i K_{ij}$ represent the number of failures at u_j, and let $\mathbf{K}^{\mathrm{T}} = \{K_1, K_2, \ldots, K_m\}$. The complete data log-likelihood for $\boldsymbol{\pi}$ is

$$\ell(\pi; \mathbf{K}) = \sum_{l=1}^{m-1} K_j \log \pi_j + K_m \log\left(1 - \pi_1 - \pi_2 - \cdots - \pi_{m-1}\right)$$

with vector of sufficient statistics $\mathbf{K}$ and MLE $\hat{\boldsymbol{\pi}} = \mathbf{K}/n$. With censored data, the EM algorithm is defined as follows. Let $\hat{\boldsymbol{\pi}}$ be the current parameter update, and let $\delta_{ij} = I\{j \in \mathcal{U}_i\}$. At the E-step, the expected sufficient statistics are

$$\hat{K}_j = E[K_j \mid \mathcal{U}, \boldsymbol{\pi}] = \sum_{i=1}^{n}\left[\delta_{ij}\pi_j \Big/ \sum_{k=1}^{m}\delta_{ik}\pi_k\right], \quad j = 1, \ldots, m$$

The M-step updates π_j with $\pi_j' = \hat{K}_j/n$ $(j = 1, \ldots, m)$. At convergence, $\hat{S}(t) = \sum_{j:u_j \ge t}\hat{\pi}_j$ is the KM estimator.

In making use of longitudinal prognostic information with the parametric mixture model, the EM algorithm for estimating $\boldsymbol{\pi}$ remains essentially the same. Let $f_{\boldsymbol{\theta}}(\mathbf{y}_i \mid u_i)$ represent the model for the conditional distribution of $\mathbf{y}_i$ given u_i (equivalently, $\mathbf{y}_i$ given $\mathbf{K}_i$), with parameter vector $\boldsymbol{\theta}$. Suppose now that $\boldsymbol{\theta}$ is known. The E-step uses information on both $\mathbf{Y}^{\mathrm{T}} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ and $\mathcal{U}$ to calculate expectations of the K_j; that is

$$\begin{aligned}\hat{K}_j &= E[K_j \mid \mathbf{Y}, \mathcal{U}, \boldsymbol{\pi}, \boldsymbol{\theta}] \\ &= \sum_{i=1}^{n} E[K_{ij} \mid \mathbf{y}_i, \mathcal{U}_i, \boldsymbol{\pi}, \boldsymbol{\theta}] \\ &= \sum_{i=1}^{n} \left[\delta_{ij} \pi_j f_{\boldsymbol{\theta}}(\mathbf{y}_i \mid K_{ij} = 1) \Big/ \sum_{k=1}^{m} \delta_{ik} \pi_k f_{\boldsymbol{\theta}}(\mathbf{y}_i \mid K_{ik} = 1) \right] \end{aligned} \tag{4.1}$$

The M-step remains unchanged. Note that in the E-step, $E[K_{ij}|\text{observed data}] = 1$ when failure occurs at u_j, meaning that if no subjects are censored, $\hat{\boldsymbol{\pi}}$ is the empirical cdf. For censored cases, the expectation is the probability of failing at u_j^*, conditional on $\mathbf{y}_i$ and $u_i > u_j^*$.

Because the log likelihood for the model $f_{\boldsymbol{\theta}}(\mathbf{Y}|\mathbf{K})$ is a linear combination of sufficient statistics for $\boldsymbol{\theta}$, the EM algorithm can be used to estimate simultaneously all model parameters. Let $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{\pi}\}$, let the complete data be $\mathcal{C} = \{(\mathbf{y}_1, u_1), (\mathbf{y}_2, u_2), \ldots, (\mathbf{y}_n, u_n)\}$, and let the observed data be denoted with $\mathcal{O} = \{(\mathbf{y}_1, \mathcal{U}_1), (\mathbf{y}_2, \mathcal{U}_2), \ldots, (\mathbf{y}_n, \mathcal{U}_n)\}$. The complete data log-likelihood is

$$\ell(\boldsymbol{\Theta}; \mathcal{C}) = \sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(\mathbf{y}_i | u_i) + \log f_{\boldsymbol{\pi}}(u_i)$$

At the E-step, using the current update $\hat{\boldsymbol{\Theta}}$, we calculate the expected log-likelihood with respect to the observed data. Recall that u_j^* is the jth ordered failure time. Let $\mathcal{C}_{ij} = \{\mathbf{y}_i, (u_i, \xi_i) = (u_j^*, 1)\}$; this represents observed data configured according to failure at $u = u_j^*$. The expected log-likelihood is found by taking a weighted average of complete-data likelihoods across j

$$\begin{aligned} Q(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}) &= E[\ell(\boldsymbol{\Theta}; \mathcal{C}) \mid \hat{\boldsymbol{\Theta}}, \mathcal{O}] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} E\left[K_{ij} \mid \hat{\boldsymbol{\Theta}}, \mathcal{O}\right] \ell_i(\boldsymbol{\Theta}; \mathcal{C}_{ij}) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{K}_{ij} \ell_i(\boldsymbol{\Theta}; \mathcal{C}_{ij}) \end{aligned}$$

where $\hat{K}_{ij}$ is calculated as in (4.1), substituting $\hat{\boldsymbol{\pi}}$ for $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$. The M-step is a matter of maximizing a weighted complete data log-likelihood for $\boldsymbol{\Theta}$; the task separates nicely into two parts because the complete data log-likelihood factors over $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$. The update for π_j is $\sum_i \hat{K}_{ij}/n$. It is possible to use either ML or restricted maximum likelihood (REML) to update $\boldsymbol{\theta}$ by applying weighted versions of formulas supplied by Laird *et al.*[38]

Although the principles of ML or REML estimation in the mixture model are somewhat transparent, in practice it can be difficult to apply the estimation methods. One limitation is slow convergence of the EM to the variance component estimates;[38] another is that the algorithm can become unstable in cases where the estimates of variance components are near the boundary of the parameter space; in our experience, these become especially problematic under heavy censoring.

Another potential difficulty arises in the proper calculation of standard errors for $\hat{\boldsymbol{\pi}}$. In the KM estimator, for example, we found via simulation that standard errors calculated from the observed information matrix[39] sometimes under-estimated empirical standard errors calculated via simulation. In samples of size 150 generated from a lognormal survival distribution, we found information-based standard errors to be fairly accurate when survival is recorded on a highly discretized scale, but less accurate in situations with few (one or two) events per observed event time.

For situations where censoring is independent of survival, a simple alternative to full implementation of the EM algorithm is to apply a two-step estimation procedure which gives consistent but not fully efficient estimates under a correctly specified model for $f_{\boldsymbol{\theta}}(\mathbf{y}|u)$ and under the assumption that censoring is independent of u_i conditional on $\mathbf{y}_i$. At the first step, obtain an estimate $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ from the combined group of subjects who have either an observed event in $[0, \tau]$ or have $\tilde{u}_i > \tau$. This can be done using a standard software package such as Proc Mixed for SAS (SAS Institute, Cary, NC, USA), BMDP 5V (BMDP Statistical Software, Los Angeles, CA, USA), or the LME program[40] designed for S-Plus (MathSoft, Seattle, WA, USA). Next, use (4.1), with $\tilde{\boldsymbol{\theta}}$ in place of $\boldsymbol{\theta}$, to obtain an estimate $\tilde{\boldsymbol{\pi}}$ of $\boldsymbol{\pi}$. Unless the survival distribution is highly discrete – in which case one can reasonably assume independence between $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\pi}}$ and apply the delta method – standard errors should be calculated via resampling techniques such as the bootstrap.[41]

5 Analysis of time to disease progression using repeated measures of CD4

Protocol 128 of the ACTG was a randomized double-blind clinical trial designed to compare two doses of zidovudine therapy for mild to moderately symptomatic HIV-infected children.[42] The study enrolled 424 mild to moderately symptomatic HIV-infected children who were randomized to receive zidovudine at either standard dose (180 mg/m^2 body surface area, every 6 h) or reduced dose (90 mg). The investigators reported comparisons on several time to event outcomes, including time to death, time to AIDS-defining illness, and time to clinical disease progression. (Occurrence times for these events are not necessarily mutually exclusive.) For this example, we analyse time to disease progression data and use longitudinally measured CD4 cell count for recovering information. Children were followed for up to 52 months, but the last clinical progression was recorded at about month 42 (3.5 years).

For the purposes of our example, we concern ourselves with estimating the probability of remaining event-free for three years ($\tau = 36$ months), and use longitudinal CD4 cell counts data to augment information lost to censoring. We analyse data for the 208 children on the standard dose arm, and use only outcomes collected up to year 3. Forty of the children (19%) were observed to remain event free

up to year 3; approximately 20% were censored prior to that point. Subjects are censored if they leave the study or if they are switched to another treatment regimen (as happened in a small number of cases). The KM estimator of event free probability at year 3 is 0.355, with bootstrap standard error 0.0373.

CD4 counts were measured approximately once per month for two years, but somewhat less regularly thereafter. Because the event of interest is disease progression and not death, CD4 data are available on some subjects even after the event time; however, CD4 was not recorded beyond a censoring time. Figures 1–3 show observed CD4 (on the square root scale) for samples of patients who progress prior to year 3 (Figure 1), who remain progression free for 3 years (Figure 2), and who are censored prior to year 3 (Figure 3). Figure 4 is an exploratory plot which shows observed CD4 data for uncensored patients stratified on survival time; quantile regression lines are overlaid to indicate location and scale. Those observed to fail before year 3 exhibit considerably more variability.

Normal quantile plots of observed CD4 data at baseline and at selected time points suggest that a square root transformation is appropriate for a normal error model. We model the distribution of square root CD4 as a two-component mixture of random

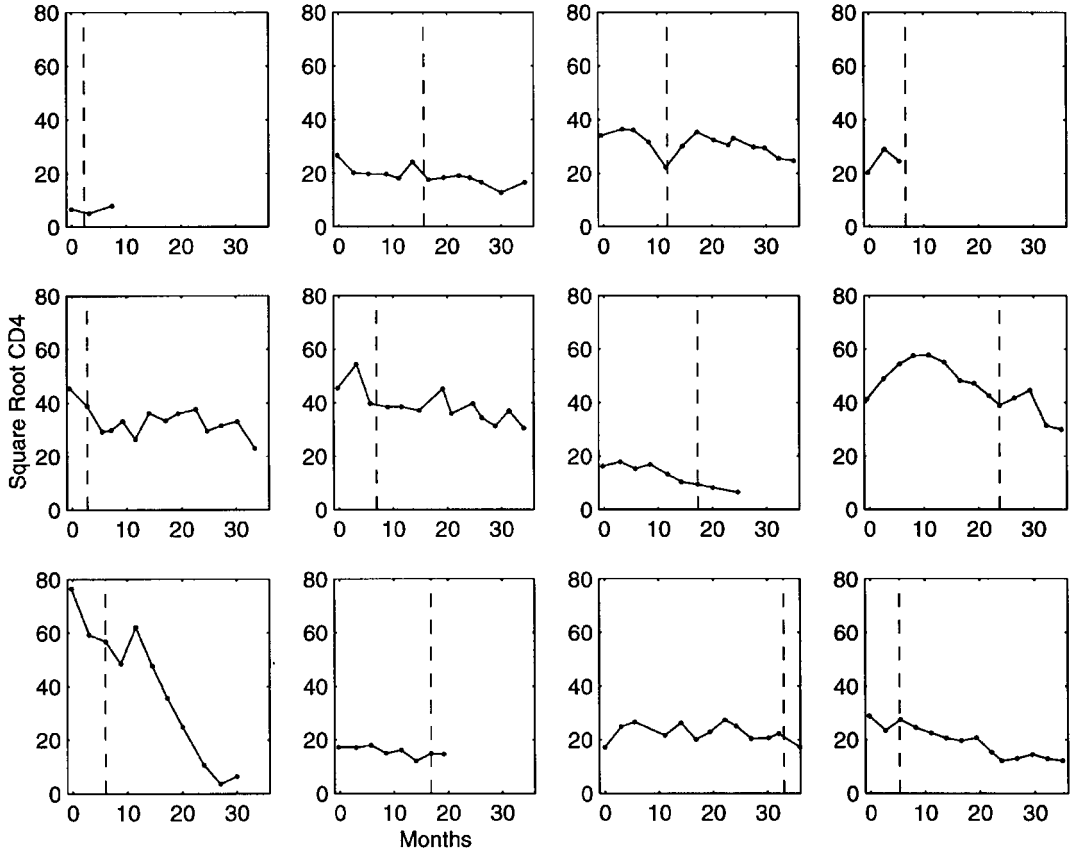

Figure 1 Longitudinal square root CD4 data for 12 patients observed to experience disease progression prior to year 3. Vertical hashed line represents time of disease progression.

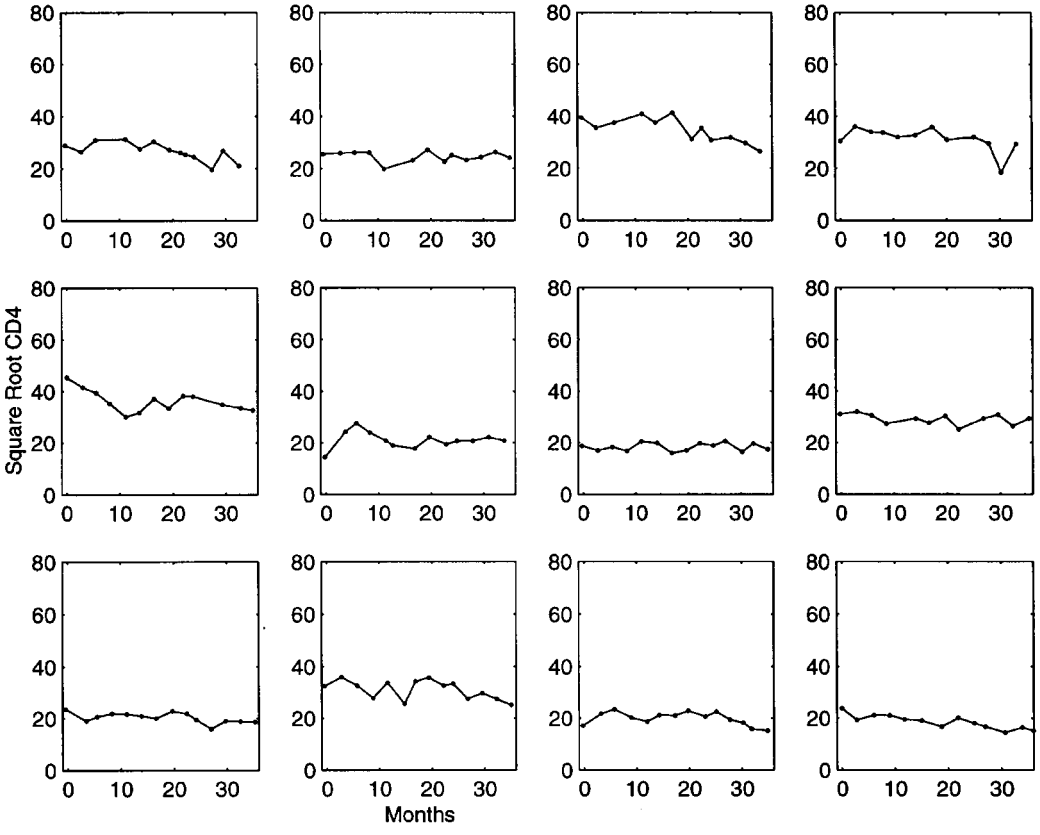

Figure 2 Longitudinal square root CD4 data for 12 patients observed to be disease free for three years

effects models which specify a linear trend over time; the two components of the mixture are defined by whether or not progression occurs prior to year 3. Timing of CD4 measurements is recorded in weeks, but we rescale the time axis using $t_{ij}^{*} = t_{ij}/24$ to facilitate stable estimation of the slope variances.

We fit the model using both the two-step procedure and using the EM algorithm. For estimating variance components, we used REML, which is typically preferred to maximum likelihood to avoid downward bias. Standard errors for the estimates of $\boldsymbol{\theta}$ and $S(\tau)$ are estimated via bootstrap resampling (250 resamples). We checked the linearity assumption by adding a random quadratic term to the model and found no substantial improvement in fit (assessed using Akaike's information criterion[43]).

Table 1 shows parameter estimates for the component CD4 distributions. Those who experience the event beyond year 3 have slightly higher mean square root CD4 at baseline (30.9 vs 29.6) but substantially different slope (-1.3 vs -2.4 units change per 24 weeks; 95% confidence interval for the difference is (0.5, 1.6)). Estimated intercept variance is about twice as high for those who progress prior to year 3, and estimated slope variance is more than three times greater. Within-subject variance also is higher among those progressing early (15.7 vs 11.7).

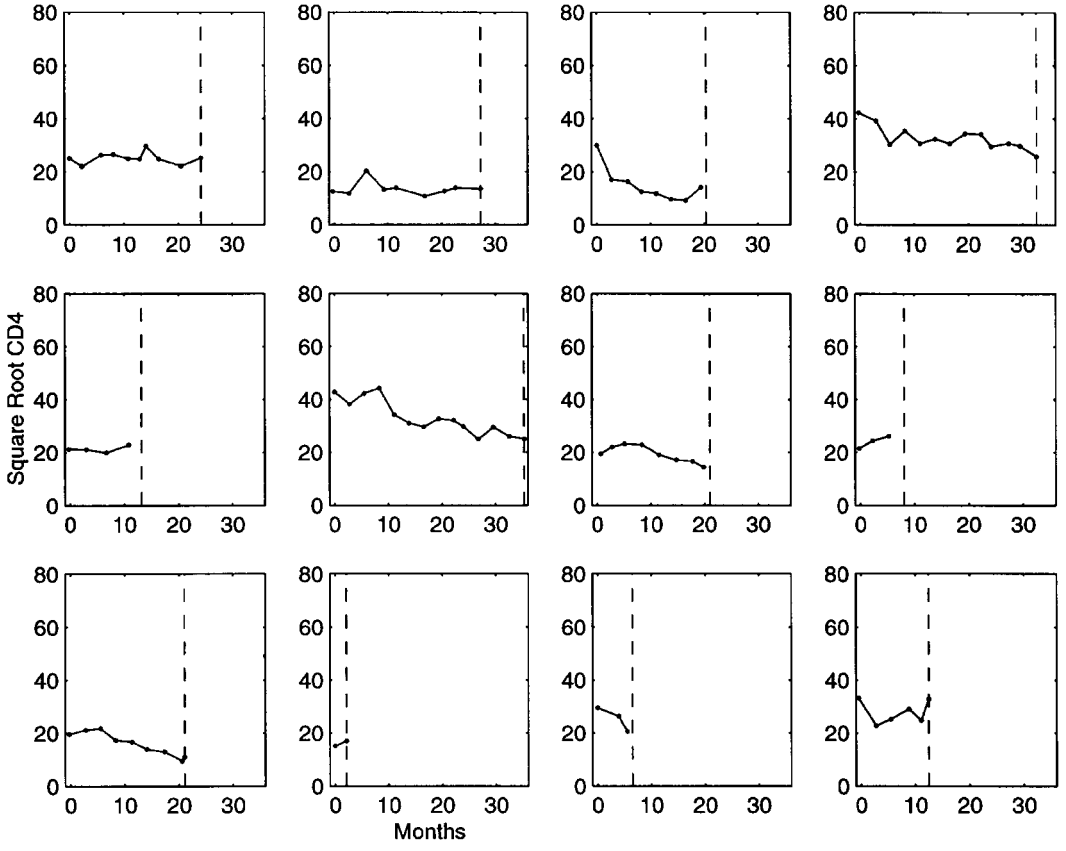

Figure 3 Longitudinal square root CD4 data for 12 patients lost to follow-up (censored) prior to year 3. Vertical hashed line indicates censoring time.

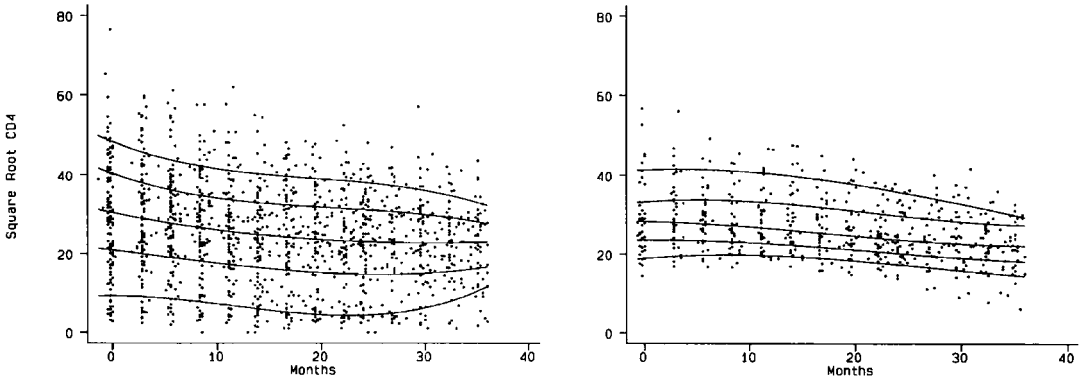

Figure 4 Longitudinal square root CD4 data for subjects who experience disease progression prior to year 3 (left panel) and after year 3 (right panel). From top to bottom, each band represents 90th, 75th, 50th, 25th and 10th percentile, obtained using quantile regressions with cubic time trends.

Table 1 REML parameter estimates (bootstrap standard error) for square root CD4 distribution and three-year survival probability from 208 subjects in the high-dose arm of ACTG128

Parameter	Clinical progression	
	Before year 3	After year 3
Mean intercept	29.6 (1.18)	30.9 (1.41)
Mean slope (per 24 weeks)	−2.4 (0.19)	−1.3 (0.19)
var(b_0)	181.0 (22.63)	98.4 (20.51)
var(b_1)	4.9 (1.38)	1.5 (0.55)
corr(b_0, b_1)	−0.52 (0.08)	−0.66 (0.11)
Within-subject variance	15.7 (1.47)	11.7 (1.85)
Three-year survival probability		
REML	0.341 (0.0363)	
Kaplan–Meier	0.355 (0.0373)	

The mixture model REML estimate of three-year event-free probability is 0.341, with bootstrap standard error 0.0363, representing a 6% gain in efficiency over the KM estimator. The two-step method yields survival estimate 0.340; interestingly, the standard error of the two-step estimator calculated from the identical bootstrap samples is 0.0356. The discrepancy is a small one and it should be kept in mind that the standard errors themselves are estimated. The variance component estimates from the two-step estimator indicate larger differences in component CD4 distributions than those from the REML estimator; for the respective component distributions, the estimates of intercept variance are 188.4 and 65.5, and of slope variance 4.6 and 0.9.

The REML estimate differs only slightly from the KM (0.341 vs 0.355). Larger discrepancies may indicate either a dependent censoring process or a misspecified model for $f(\mathbf{y}|u)$. We take up these issues in more detail in the discussion. Although the gain in efficiency is modest, recall that the censoring rate is only 0.20. Simulation studies in the next section indicate that larger efficiency gains are possible when censoring is heavier, even when using $\tilde{\boldsymbol{\theta}}$.

6 Simulation study

In order to study properties of mixture model estimators under different circumstances, we carried out a small simulation study which is designed to mimic clinical trial designs like ACTG128. We consider the problem of estimating the probability of remaining event-free beyond the last time at which survival is followed up, and generate data under two general scenarios. In the first, follow-up on both the repeated measures and the survival process is $T = \tilde{T} = 10$ months; in the second, the repeated measures are taken for $\tilde{T} = 10$ months but subjects are followed on survival for $T = 30$ months. In both cases, our estimand is $S(T)$, the probability of being without disease progression at the end of follow-up on survival.

Each subject has 11 potential outcomes y_{ij}, taken at times $\mathbf{t} = (0, 1, 2, \ldots, 10)$. To be conservative, we assume that repeated measures cease to be observed beyond $u_i^* = \min(u_i, c_i)$. We generate both failure time and censoring time from a log-normal distribution, and vary nominal censoring rates from 30 to 70%; censoring times are generated independent of survival times. Relative efficiency calculations are based on

500 realizations of the KM estimator $\hat{S}_{\mathrm{K}}(T)$ and the mixture model two-step estimator $\hat{S}_{\mathrm{M}}(T)$ calculated from samples of size 200. The estimated relative efficiency is $\widehat{\mathrm{RE}} = \widehat{\mathrm{var}}\hat{S}_{\mathrm{K}}(T)/\widehat{\mathrm{var}}\hat{S}_{\mathrm{M}}(T)$, where $\widehat{\mathrm{var}}$ represents sample variance.

For our simulation study, applying REML estimation via the EM algorithm across such a wide variety of data sets proved to be time consuming and sometimes unstable, especially in cases with heavy censoring. Instead, we estimated relative efficiencies using: (i) the two-step procedure and (ii) known parameter values. Under independent censoring, both the two-step and REML estimates are consistent for the survival distribution, so the relative efficiencies under the two-step estimator can be regarded as lower bounds of relative efficiencies realized under REML (or ML). Simulations were carried out using Matlab, Version 5.0 (MathWorks, Inc, Natick, MA, USA); estimates of $\boldsymbol{\theta}$ for the two-step procedure were calculated via REML using Proc Mixed in SAS (SAS Institute, Cary, NC, USA).

Data were generated as follows:

1) Log $u_i \sim N(3,1)$, where u_i is the underlying survival time.
2) Log $c_i \sim N(\mu,1)$ is the log censoring time. Nominal censoring rates of 30%, 50% and 70% correspond to $\mu = 3.74$, 3, and 2.26, respectively.
3) Both u_i and c_i are rounded to the nearest integer before finding $\tilde{u}_i = \min(u_i, c_i)$ and $\xi_i = I\{u_i \leq c_i\}$. The rounding is done to mimic actual data records.
4) If $\tilde{u}_i > T$, then $\tilde{u}_i$ is set equal to T and ξ_i is set to 1. In the multinomial specification of the survival distribution, this places all those who are observed to be event free at T into one category.
5) For $k \in \{0,1\}$, $\beta_{ki} = \Delta_k I\{u_i > T\} + b_{ki}$, where $(b_{0i}, b_{1i})^{\mathrm{T}} \sim N(\mathbf{0}, \boldsymbol{\Phi}(u_i))$ is bivariate normal.
6) $y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + e_{ij}$, where $e_{ij} \sim N(0,\sigma^2)$, $\sigma^2 = 1/4$ and $\mathbf{t}_i = (0,1,2,\ldots,\min\{\tilde{u}_i, 10\})^{\mathrm{T}}$. Under this model, $S(10) \doteq 0.76$ and $S(30) \doteq 0.34$.

Simulation results are presented in Table 2. For each model configuration, two numbers are listed; the first is the estimated RE based on estimating $S(T)$ from $\tilde{\boldsymbol{\theta}}$, and the second based on estimating $S(T)$ from the true, underlying value of $\boldsymbol{\theta}$.

Consider first the case where the random effects distribution differs only by a shift in the mean, but where the variances do not change. For this scenario, $\phi_0 = \mathrm{var}(b_0) = 1$, $\phi_1 = \mathrm{var}(b_1) = 1/100$, $\sigma^2 = 1/4$, and the intercept is taken to be independent of the slope. We assume that those who survive past T differ from those who do not in the intercept only and then in both intercept and slope. For each parameter, the difference is one standard deviation. When the component distributions differ only in the intercept, efficiency gains are negligible if $\boldsymbol{\theta}$ is being estimated (rows 1–3 of Table 2). When survival beyond T depends on both intercept and slope, however, up to 21% gain in efficiency over the KM estimator can be realized (rows 4–6). Noteworthy here is that when $T = 30$, the disparity in efficiency gains between using known and estimated values of $\boldsymbol{\theta}$ is greater under heavy censoring. Missingness in $\mathbf{y}_i$ induced by heavy censoring adds variability to $\tilde{\boldsymbol{\theta}}$ and hence to the mixture model estimate of the survival distribution.

Table 2 Simulation-based estimated relative efficiencies (RE) of KM estimator compared to the mixture model estimator of survival for 10 and 30 months follow-up on survival and 10 months follow-up on repeated measures. The first number listed is RE using the two-step estimation method where $\boldsymbol{\theta}$ is estimated from observed failures only; the second number uses the true θ. For each simulation, data for 200 subjects are replicated 500 times

Variance parameters						Mean parameters*		Nominal censoring rate	Relative efficiency (KM:MM) for estimating $S(T)$	
$U \le T$			$U > T$							
ϕ_0	ϕ_1	σ^2	ϕ_0	ϕ_1	σ^2	Δ_0	Δ_1		$T = 10$	$T = 30$
1	0.01	0.25	1	0.01	0.25	1	0	0.3	1.00, 1.01	1.04, 1.06
								0.5	0.98, 0.99	1.00, 1.10
								0.7	1.02, 1.05	0.96, 1.28
						1	0.1	0.3	1.01, 1.04	1.05, 1.07
								0.5	1.04, 1.04	1.17, 1.27
								0.7	1.08, 1.12	1.21, 1.54
1	0.01	0.5	1	0.01	0.5	1	0.1	0.5	1.04, 1.05	1.16, 1.29
								0.7	1.05, 1.06	1.08, 1.47
		1			1			0.5	1.02, 1.04	1.16, 1.23
								0.7	1.09, 1.11	1.23, 1.46
1	0.01	0.25	0.25	0.0025	0.25	0	0	0.3	1.00, 1.00	1.07, 1.09
								0.5	1.00, 1.01	1.03, 1.14
								0.7	1.04, 1.07	1.07, 1.24
						1	0	0.3	1.02, 1.02	1.06, 1.09
								0.5	1.06, 1.06	1.16, 1.29
								0.7	1.07, 1.13	1.17, 1.59
						1	0.1	0.3	1.04, 1.04	1.17, 1.18
								0.5	1.05 ,1.06	1.17, 1.46
								0.7	1.08, 1.16	1.37, 1.94
						0.5	0.05	0.3	1.01, 1.01	1.11, 1.12
								0.5	1.03, 1.04	1.11, 1.21
								0.7	1.09, 1.12	1.10, 1.39

*$\Delta_k = E(\beta_k | U > T) - E(\beta_k | U \le T)$ for $k = 0, 1$.

The middle section of the table shows the effect of increasing within-subject error by factors of two and four when the intercept and slope differ by one standard deviation each. The increase has surprisingly little effect on relative efficiency, regardless of whether known or estimated values of $\boldsymbol{\theta}$ are being used.

The final section of Table 2 presents estimates of relative efficiency for situations where the component distributions differ in scale. We decrease $\boldsymbol{\Phi}$ by a factor of four (reducing standard deviation of the intercept and slope each by one half) for those who are event free at T, reflecting a more stable repeated measures distribution for the event-free subjects. When the component distributions differ only in the scale and not in the mean, efficiency gains are modest and slightly greater than the intercept-change-only scenario (up to 7% efficiency gains when using estimated $\boldsymbol{\theta}$). When coupled with a change in scale, changes in the intercept and slope give more sizable efficiency gains, with estimated RE ranging from 1.02 to 1.37 when using $\tilde{\boldsymbol{\theta}}$ and from 1.02 to 1.94 when using the true $\boldsymbol{\theta}$.

It appears that the potential for recovering information is greater when follow-up on survival is longer than on the repeated measures, except perhaps in the case of heavy censoring. As mentioned earlier, interim analyses are situations in which a considerable amount of censoring is attributable to late entry (and is thus assumed independent of survival), and where the mixture model survival estimator is potentially useful. Even for a completed study with rather light censoring (such as ACTG128), modest efficiency gains can be made if the two component random effects distributions are substantially different. Moderate increases in within-subject error variances apparently do not have substantial effects on potential efficiency gains.

7 Discussion

We have given a method for incorporating auxiliary longitudinal covariates into a survival estimator by modelling their random effects distribution as a mixture over the survival distribution. A parametric model is used for component distributions in the mixture, but the marginal survival distribution function is left unspecified. Our focus is on making gains in efficiency for point estimates of survival under the assumption that censoring is independent of survival, conditional on the longitudinal covariates.

The simulation studies indicate that relative to the KM estimator, the survival estimator derived from the mixture model is more efficient when censoring is heavy or when follow-up on survival is longer than that on the repeated measures. The mixture model may prove useful in situations such as interim analyses or drawing inference after long-term follow-up on survival.

Our approach has several attractive features. First, modelling the repeated measures distribution as a two component mixture is conceptually simple and requires little in the way of complicated assumptions about the association between $\mathbf{y}$ and u. Second, the model allows one to take advantage of differences in both location and scale among those who fail before a prespecified time and those who do not; indeed, even a difference in scale alone distinguishes subjects well enough to give modest gains in efficiency. Third, the *ad hoc* method which uses failures to estimate the association parameters gives reasonable recovery of information under the assumption that censoring is independent of survival. When the component distributions arise from

the random effects model (1), consistent estimates of $\boldsymbol{\theta}$ are easily obtained using SAS Proc Mixed or similar software. Our approach accounts for both within- and between-subject variations in the longitudinal covariates, and is similar in spirit to the work of Tsiatis *et al*.,[3] Self and Pawitan[13] and Wulfsohn and Tsiatis;[14] by contrast, Murray and Tsiatis[6] assume that longitudinal covariates are measured without error.

There are notable limitations to applying these methods; of primary concern is correct specification of $f(\mathbf{y}|u)$. Misspecification can result in bias and, in a testing situation, possible increases in type I error. Fortunately, the data analyst can take advantage of standard model checking techniques for generalized linear models. As with the KM estimator, untestable assumptions about the censoring process must be made when implementing either the full ML or the two-step estimation. ML (or REML) estimates are valid under the assumption that censoring is independent of survival conditional on the repeated measures and that $f(\mathbf{y}|u)$ is correctly specified. In situations where censoring depends on u_i only through $\mathbf{y}_i$, ML estimates will be valid and KM estimates will be biased. When using the two-step procedure which subsets on the observed failures to estimate $\boldsymbol{\theta}$, survival estimates generally are valid only under the assumption that censoring is unconditionally independent of survival (the same assumption required for using the KM estimator). For example, if censoring depends on survival, but there is little or no association between survival and the repeated measures, then censoring acts as a selection mechanism and induces bias in $\tilde{\boldsymbol{\theta}}$ (and not in the KM estimator). Although this paper is primarily concerned with efficiency, identifying possible sources of bias is important when applying the mixture model.

Issues which remain to be studied include application to hypothesis testing, a detailed study of the properties of information-based estimates of variance, effects of model misspecification, and the utility of the model to overcome bias induced by dependent censoring. Important new methodology is being developed for fitting random effects models to repeated categorical data[44,45,46] and to multivariate longitudinal data[47], indicating a wide variety of potential applications within our general framework.

Acknowledgements

The authors thank the Pediatric AIDS Clinical Trials Group for providing the data from Protocol 128 (collected under NIAID contract N01-AI 95030), Dr Jane Lindsey for her assistance in preparing the data for analysis, and an anonymous referee for thoughtful and valuable comments on the initial manuscript. This work was partially supported by grant GM29745 from the United States National Institutes of Health.

References

1 Little RJA. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* 1993; **88**: 125–34.

2 Heyting A, Tolbloom JTBM, Essers JGA. Statistical handling of drop-outs in longitudinal clinical trials. *Statistics in Medicine* 1992; **11**: 2043–61.

3 Tsiatis AA, De Gruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error: applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* 1995; **90**: 27–37.

4 Fleming TR, Prentice RL, Pepe MS, Glidden D. Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine* 1994; **13**: 955–68.
5 Malani HM. A modification of the redistribution to the right algorithm using disease markers. *Biometrika* 1995; **82**: 515–26.
6 Murray S, Tsiatis AA. Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics* 1996; **32**: 137–51.
7 Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**: 963–74.
8 Crowder MJ, Hand DJ. *Analysis of repeated measures.* London: Chapman & Hall, 1990.
9 Diggle PJ. An approach to the analysis of repeated measurements. *Biometrics* 1988; **44**: 959–71.
10 Goldstein H. *Multilevel models in educational and social research*. London: Griffin, 1987.
11 Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 1988; **44**: 175–88.
12 De Gruttola V, Tu XM. Modeling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics* 1994; **50**: 1003–14.
13 Self S, Pawitan Y. Modeling a marker of disease progression and onset of desease. In: Jewell N, Dietz K, Farewell V eds. *AIDS epidemiology: methodological issues*. Boston, MA: Birkhäuser, 1992.
14 Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics* 1997; **53**: 330–39.
15 Little RJA. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 1995; **90**: 1112–21.
16 Hogan JW, Laird NM. Model-based approaches to analyzing incomplete longitudinal and failure time data. *Statistics in Medicine* 1997; **16**: 259–72.
17 Heckman JJ. Sample selection bias as a specification error. *Econometrica* 1979; **47**: 153–61.
18 Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 1977; **B39**: 1–22.
19 Faucett C, Thomas D. Simultaneously modeling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine* 1996; **15**: 1663–85.
20 Schluchter MD. Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine* 1992; **11**: 1861–70.
21 Wu MC, Bailey KR. Analysing changes in the presence of informative right censoring caused by death and withdrawal. *Statistics in Medicine* 1988; **7**: 337–46.
22 Wu MC, Bailey KR. Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics* 1989; **45**: 939–55.
23 Hogan JW, Laird NM. Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* 1997; **16**: 239–58.
24 Redner RA, Walker HF. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* 1984; **26**: 195–202.
25 Cox DR. A remark on censoring and surrogate response variables. *Journal of the Royal Statistical Society* 1983; **B45**: 391–93.
26 Lagakos SW. Using auxiliary variables for improved estimates of survival time. *Biometrics* 1977; **33**: 399–404.
27 Finkelstein DM, Schoenfeld DA. Analysing survival in the presence of an auxiliary variable. *Statistics in Medicine* 1994; **13**: 1747–54.
28 Gray RJ. A kernel method for incorporating information on disease progression in the analysis of survival. *Biometrika* 1994; **81**: 527–39.
29 Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**: 457–81.
30 Murray S, Tsiatis AA. Using auxiliary time-dependent covariates to recover information in nonparametric testing with censored data. *Journal of the American Statistical Association*, 1997 (in press).
31 Pepe MS, Fleming TR. Weighted Kaplan–Meier statistics: a class of distance tests for censored survival data. *Biometrics* 1989; **45**: 497–507.
32 Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In: Jewell N, Dietz K, Farewell V eds. *AIDS epidemiology: methodological issues*. Boston, MA: Birkhäuser, 1992.
33 Slud EV, Korn EL. Semiparametric two-sample tests in clinical trials with a post-randomisation response indicator. *Biometrika* 1997; **84**: 221–30.
34 Jewell NP, Kalbfleisch JD. Marker processes in survival analysis. *Lifetime Data Analysis* 1996; **2**: 15–29.

35 Cox DR, Oakes D. *Analysis of survival data*. London: Chapman & Hall, 1984.
36 Little RJA, Rubin DB. *Statistical analysis with missing data.* New York: John Wiley, 1987.
37 Lancaster T, Intrator O. Panel data with survival: hospitalization of HIV patients. *Journal of the American Statistical Association* 1997 (in press).
38 Laird N, Lange N, Stram D. Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of the American Statistical Association* 1987; **82**: 97–105.
39 Louis TA. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1982; **44**: 226–33.
40 Pinheiro JC, Bates DM. Mixed effects models methods and classes for S and Splus. Technical report, Department of Biostatistics, University of Wisconsin at Madison, 1995.
41 Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1986; **1**: 54–75.
42 Brady MT, McGrath N, Brouwers P *et al*. (Pediatric AIDS Clinical Trials Group). Randomized study of the tolerance and efficacy of high- versus low-dose zidovudine in human immunodeficiency virus-infected children with mild to moderate symptoms (AIDS Clinical Trials Group 128). *Journal of Infectious Diseases* 1996; **173**: 1097–106.
43 Akiake H. A new look at statistical model identification. *IEEE Transactions on Automatic Control* 1974; **AC-19**: 716–23.
44 Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association,* 1993; **88**: 9–25.
45 McCulloch CE. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* 1997; **92**: 162–70.
46 Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics* 1994; **50**: 933–44.
47 Sy JP, Taylor JMG, Cumberland WG. A stochastic model for the analysis of bivariate longitudinal AIDS data. *Biometrics* 1997; **53**: 542–55.