STATISTICS IN MEDICINE, VOL. 16, 239–257 (1997)

MIXTURE MODELS FOR THE JOINT DISTRIBUTION OF REPEATED MEASURES AND EVENT TIMES

JOSEPH W. HOGAN

Center for Statistical Sciences, Box G-H, Brown University, Providence, RI 02912, U.S.A.

AND

NAN M. LAIRD

Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A.

SUMMARY

Many long-term clinical trials collect both a vector of repeated measurements and an event time on each subject; often, the two outcomes are dependent. One example is the use of surrogate markers to predict disease onset or survival. Another is longitudinal trials which have outcome-related dropout. We describe a mixture model for the joint distribution which accommodates incomplete repeated measures and right-censored event times, and provide methods for full maximum likelihood estimation. The methods are illustrated through analysis of data from a clinical trial for a new schizophrenia therapy; in the trial, dropout time is closely related to outcome, and the dropout process differs between treatments. The parameter estimates from the model are used to make a treatment comparison after adjusting for the effects of dropout. An added benefit of the analysis is that it permits using the repeated measures to increase efficiency of estimates of the event time distribution.

1. INTRODUCTION

Models for the joint distribution of longitudinal and event-time variables have a wide variety of applications in clinical trials and prospective studies. For example, disease progression in AIDS may be modelled using a longitudinally-measured surrogate marker such as CD4 cell count. Several authors have evaluated the merits of surrogate markers for AIDS using joint distribution models (see Self and Pawitan,[1] Tsiatis *et al.*,[2] DeGruttola and Tu).[3] A second class of applications for these models is longitudinal studies where subjects have different lengths of follow-up time, spend different amounts of time in compliance with assigned treatment, or otherwise cease to be measured on assigned protocol. Wu and Bailey,[4,5] Wu and Carroll,[6] Schluchter,[7] Mori *et al.*,[8] Diggle and Kenward,[9] and Little[10] propose models in which the mean longitudinal response – or some appropriate summary such as slope across time – can be estimated after accounting for the effects of early withdrawal from a study. Little[11] provides a comprehensive review of methods for longitudinal data with dropouts.

In a model for a joint distribution, no causal relationship is assumed; interpretation depends largely on the setting. For example, in a longitudinal study with attrition rates which depend on outcome, an investigator can use the model to make inference about the marginal distribution of the repeated outcomes after accounting for the effects of attrition. Alternatively, the repeated

CCC 0277–6715/97/030239–19

measures can be regarded as time-varying covariates measured with error when the outcome of interest is a survival or duration time.

In applied situations, the joint distribution of a vector $\mathbf{Y}$ of repeated outcomes and an event time D is commonly described by either a *selection model* or a *mixture model*. In selection models, time to event is conditioned on the longitudinal outcomes so that the joint density function $f_{Y,D}$ is modelled as $f_{D|Y} f_Y$ (here and throughout, we use f subscripted by a random variable to indicate a density function). A typical approach is to specify f_Y using a linear mixed effects model[12] and use the individual random slopes as explanatory variables in a proportional hazards model[13] for dropout time. Selection models often require specialized software for parameter estimation, and in general are computationally quite intensive.

In mixture models, the joint distribution is formed by conditioning the repeated measures on event time and mixing over f_D; that is, $f_{Y,D} = f_{Y|D} f_D$. Wu and Bailey,[4,5] for example, allow $\mathbf{Y}$ to depend on D through individual random effects which are estimated from a linear mixed model for $\mathbf{Y}$. Limitations of available mixture model approaches include assuming the dropout process is fully observed[4,5,10] and assuming a parametric form for the dropout distribution.[3,7]

Dawson and Lagakos[14] give a method for making treatment comparisons with longitudinal data which avoids specifying a model for the relationship between $\mathbf{Y}$ and D; the idea is to combine conditional test statistics across dropout times to make a single treatment comparison. Although the approach is conceptually similar to the mixture model, it makes the more limiting assumption that the relationship between outcome and event time is the same between treatment groups, and that dropout distribution is equal between treatments. A general review of several parametric and non-parametric methods for analysing longitudinal data with dropouts is provided by Wu *et al.*[15]

We propose a likelihood-based mixture-model approach to modelling the joint distribution of $\mathbf{Y}$ and D when observations on $\mathbf{Y}$ may be incomplete and when D is subject to independent right censoring. No parametric restriction is placed on the event-time distribution. The conditional distribution of $(\mathbf{Y}|D)$ is multivariate normal and follows a linear mixed effects model for longitudinal outcomes; the event time D is a covariate for the conditional mean of $\mathbf{Y}$. Because the event times may be right-censored, the model for $f_{Y|D}$ must handle incomplete covariates. Our approach to handling the censored covariate is similar to that proposed by Ibrahim[16] for incomplete covariates in the generalized linear regression model. An appealing feature of the mixed effects structure for $f_{Y|D}$ is that it allows individuals to have different numbers of observations and mistimed measurements.

We use our model to analyse data from a double-blind randomized clinical trial of adult schizophrenics; the trial is designed to compare a standard anti-psychotic medication to a new one. The repeated outcome $\mathbf{Y}$ is level of psychosis as measured using the Brief Psychiatric Rating Scale (BPRS).[17] Each subject is scheduled for weekly evaluation for six weeks (except at the fifth week), and a comparison of patient outcome at the sixth week is used to assess treatment effect. Over the course of the study, patients who perform poorly under assigned treatment or who have serious side-effects are removed from the trial; further, the attrition rate may differ in the two treatment groups. In this case, removal from trial for reasons related to poor outcome constitutes the event at time D; we let removal due to side-effects denote censoring of the event time.

In the presence of outcome-related dropout, traditional data-analytic techniques may yield biased results of treatment effect. In the schizophrenia example, a 'completers-only' analysis, which compares the group means of those whose measurements are available at the sixth week, overstates the treatment benefit on each arm because only patients on whom the treatment is successful are included. Another approach is to fit a random effects growth curve model[12] to all available observations of BPRS score, and compare the model-based estimates of week-6 mean

score. Under this model, however, all missing data in the response is assumed to be missing at random (MAR);[18] thus, conditional on observed response prior to dropout, those who drop out early are assumed to come from the same distribution as those who drop out late or even complete the trial.

In many applications, such as those where the event is loss to follow-up or death, observations on $\mathbf{Y}$ cease after the event occurs; for the model we propose, this restriction is unnecessary. The distribution of the entire $\mathbf{Y}$-vector, including outcomes which occur before and after the event, is specified by the model. There are applications in which post-event outcomes are available; if, for example, the event time is duration of compliance with initial treatment or time to primary infection, investigators are sometimes able to continue measuring $\mathbf{Y}$ and may have scientific reasons for doing so. Although it is straightforward to include measurements on $\mathbf{Y}$ taken after time D, the interpretation of the mean response depends on what model is used and what data are included. This point is considered further in the discussion.

2. THE MODEL

Most designs for longitudinal data specify a fixed number of observations, say M, for each subject. Let $\mathbf{Y}_i^*$ be an $M \times 1$ vector of outcomes for the ith person, and let D_i be the elapsed time to some event which is related to the distribution of $\mathbf{Y}_i^*$. For example, if $\mathbf{Y}_i^*$ is clinical outcome, one type of event is withdrawal from trial due to lack of treatment efficacy. By contrast, patient withdrawal for reasons not related to outcome is not, by our definition, an 'event'; its occurrence censors our observation of the true event time. A detailed treatment of the observed event-time process appears at the end of this section and in Section 3.

The pairs $(\mathbf{Y}_i^*, D_i)$ are independently distributed according to the joint density function $f_{Y^*,D}$, which can be expressed as the mixture $f_{Y^*,D} = f_{Y^*|D} f_D$. We assume that conditional on D_i, $\mathbf{Y}_i^*$ is multivariate normal; although any structure can be given to the variance, we adopt the random effects structure of the linear mixed model.[12] No parametric form is assumed for the cumulative distribution function (CDF) F_D of D. Unconditionally, $\mathbf{Y}_i^*$ is a mixture of the conditional normal distributions.

In practice, it is common for observations on $\mathbf{Y}_i^*$ to be missing. Let

$$\mathbf{Y}_i = [Y_{i1}, Y_{i2}, \ldots, Y_{i,m_i}]'$$

represent the observed portion of $\mathbf{Y}_i^*$, where measurements are recorded at times $t_{i1}, \ldots, t_{i,m_i}$. The assumption that $\mathbf{Y}_i^*$ follows a linear mixed model given D_i – and hence that missing values in $\mathbf{Y}_i^*$ are MAR conditional on D_i – allows us to express the conditional piece of the mixture model in terms of $f_{Y|D}$ rather than $f_{Y^*|D}$. The model for $(\mathbf{Y}_i|D_i)$ is described in two stages. In the first stage

$$(\mathbf{Y}_i|\boldsymbol{\beta}_i, D_i) = \mathbf{Z}_i\boldsymbol{\beta}_i + \mathbf{e}_i$$

where $\mathbf{Z}_i$ is an $m_i \times P$ design matrix for the mean trajectory of $\mathbf{Y}_i$ over time, $\boldsymbol{\beta}_i$ is $P \times 1$ vector containing an individual's random intercept and time trends, and $\mathbf{e}_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}_{m_i}(\mathbf{0}, \sigma^2\mathbf{I}_i)$ is random error. The design matrix $\mathbf{Z}_i = \mathbf{Z}_i(D_i)$ may depend on event time. At the second stage, the regression

$$(\boldsymbol{\beta}_i|D_i) = \mathbf{W}_i\boldsymbol{\alpha} + \mathbf{b}_i$$

gives the mean of each $\boldsymbol{\beta}_i$ as a function of event time D_i and any other covariates of interest (such as treatment group and group–time interactions) through the $P \times R$ design matrix $\mathbf{W}_i = \mathbf{W}_i(D_i)$ and an R-vector $\boldsymbol{\alpha}$ of parameters for the mean. The $\mathbf{b}_i$ are i.i.d. P-dimensional error vectors which are independent of each $\mathbf{e}_j$ and are normally distributed with zero mean and variance matrix $\boldsymbol{\Phi}$.

The design matrix $\mathbf{W}_i$ does not require specifying $\boldsymbol{\beta}_i$ as a linear or non-linear function of dropout time; it arbitrarily describes the conditional mean of $\boldsymbol{\beta}_i$, given a particular realization of D. The two stages can be combined in the standard formulation for a linear mixed model,

$$\begin{aligned}(\mathbf{Y}_i|D_i) &= \mathbf{Z}_i(\boldsymbol{\beta}_i|D_i) + \mathbf{e}_i \\ &= \mathbf{Z}_i[\mathbf{W}_i\boldsymbol{\alpha} + \mathbf{b}_i] + \mathbf{e}_i \\ &= \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \end{aligned} \tag{1}$$

where $\mathbf{X}_i = \mathbf{X}_i(D_i) = \mathbf{Z}_i\mathbf{W}_i$ is the $m_i \times R$ design matrix for fixed effects, and $\mathbf{Z}_i$ is the random effects design described in the first stage. Independence of $\mathbf{b}_i$ and $\mathbf{e}_j$ for all pairs (i, j) gives $\text{var}\{\mathbf{Y}_i|D_i\} = \boldsymbol{\Sigma}_i(\boldsymbol{\phi}) = \mathbf{Z}_i\boldsymbol{\Phi}\mathbf{Z}_i' + \sigma^2\mathbf{I}_i$, where $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_Q]$ are the non-redundant variance parameters; in this notation, $\phi_Q = \sigma^2 = \text{var}\{e_{ik}\}$ $(k = 1, \ldots, m_i)$.

In the absence of any information on $\mathbf{Y}_i$, the Kaplan–Meier product-limit estimator[19] is the non-parametric ML estimator of F_D.[21] Moreover, the problem of estimating F_D non-parametrically in the presence of right censoring can be transformed into the simpler one of estimating multinomial probabilities with incomplete data (see Cox and Oakes,[22] pp. 175–176). When estimating parameters in the joint distribution, however, there is information about censored values of D_i in the observed $\mathbf{Y}_i$; the product-limit estimator is still a valid estimate of F_D, but inefficient and not ML. We use the multinomial formulation of F_D because it lends itself naturally to ML estimation for parameters in the mixture model of $f_{Y,D}$ and to use of the EM algorithm.

Let $\mathscr{D} = \{s_1, s_2, \ldots, s_L\}$ be the set of ordered event times, and let $\pi_l = \Pr\{D = s_l\}$ (with $\sum_{l=1}^{L}\pi_l = 1$). The parameter vector $\boldsymbol{\pi}$ represents the $L - 1$ non-redundant probabilities needed to construct the empirical CDF $\hat{F}_D$. Under this formulation, event time data for the ith subject consists of a multinomial vector of indicators

$$\boldsymbol{\delta}_i = (\delta_{i1}, \delta_{i2}, \ldots, \delta_{iL})$$

where $\delta_{il} = I\{D_i = s_l\}$ and $\Pr\{\delta_{il} = 1\} = \pi_l$. For typical applications, including our example in Section 4, separate parameter vectors are used for different groups within the study. For simplicity, we generally use one parameter vector $\boldsymbol{\pi}$ with the tacit assumption that each treatment group has a different event-time distribution.

Missing observations in $\boldsymbol{\delta}_i$ are induced by independent right censoring of event times. Events can often be classified into those where $\mathbf{Y}_i$ and D_i are related (such as lack of efficacy leading to treatment termination) and those where $\mathbf{Y}_i$ and D_i are not (such as patient withdrawal due to relocation). The former is regarded as an observed event time, and the latter as a censored event time. When a random variable C_i independently censors D_i from the right, we observe $\tilde{D}_i = \min\{D_i, C_i\}$ and an indicator $\xi_i = I\{D_i \leqslant C_i\}$. When $\tilde{D}_i$ is a censored event time, $\xi_i = 0$ and $\boldsymbol{\delta}_i = (0, 0, \ldots, 0, *, *, \ldots, *)$ is incomplete ($*$ denotes a missing observation). The final zero occurs at element $\max\{l: s_l \leqslant \tilde{D}_i\}$. For example, if $\mathscr{D} = \{1, 3, 8\}$ and $(\tilde{D}_i, \xi_i) = (4, 0)$, then $\boldsymbol{\delta}_i = (0, 0, *)$ in the presence of right censoring.

A potential problem in using the product-limit estimator on right-censored data is defining the estimator beyond the last observed time point when that observation is censored. In particular, at the end of many studies, those still remaining have their event times censored at the time T corresponding to close of study. We avoid this problem by reserving the final multinomial category for those who have not yet experienced an event at time T, effectively creating a separate category for trial completers.

3. ESTIMATION

Two approaches to parameter estimation are described below. The first gives consistent estimates which are efficient only in the complete data case, that is, no censoring in D. The second is an ML estimation procedure carried out with the EM algorithm, and makes the most efficient use of the data with censoring in D. Some further notation is useful for the discussion of estimation procedures. Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\pi})$ be the collection of non-redundant model parameters;

$$\mathscr{C} = \{(\mathbf{Y}_i, \boldsymbol{\delta}_i, \mathbf{X}_i, \mathbf{Z}_i):\ i = 1, \ \ldots, N\}$$

refers to the complete but perhaps partially observed data, and $\mathscr{O}$ is the subset of $\mathscr{C}$ which is observed. In the context of the mixed model (1), the observed $\mathbf{Y}_i$ is considered 'complete' data; thus only elements of $\boldsymbol{\delta}_i$ are subject to missingness. Respectively, $\mathscr{C}_i$ and $\mathscr{O}_i$ represent complete and observed data on the ith individual. We let the set $\mathscr{C}_{il} = \mathscr{C}_i(s_l) = (\mathbf{Y}_i, \boldsymbol{\delta}_{il}, \mathbf{X}_{il}, \mathbf{Z}_{il})$ denote an individual's complete data and covariates, setting $D_i = s_l$. The complete-data likelihood function is $\mathscr{L}(\boldsymbol{\theta}; \mathscr{C})$ and its log-likelihood is $\ell(\boldsymbol{\theta}; \mathscr{C}) = \log \mathscr{L}(\boldsymbol{\theta}; \mathscr{C})$. Observed-data counterparts are $\mathscr{L}_{\mathrm{o}}$ and ℓ_{o}.

3.1. ML Estimation when Event Times are Completely Observed

Using the random effects model (1) to specify $f_{Y|\delta}$ and a multinomial distribution for f_δ, the complete-data likelihood for the joint distribution $f_{Y,\delta}$ is

$$\begin{aligned}
\mathscr{L}(\boldsymbol{\theta}; \mathscr{C}) &= \prod_{i=1}^{N} \int_{\mathbb{R}^p} f_{Y|b,\delta}(\boldsymbol{\alpha}, \boldsymbol{\phi}; \mathscr{C}_i, \mathbf{b}_i) f_b(\boldsymbol{\phi}; \mathbf{b}_i) f_\delta(\boldsymbol{\pi}; \boldsymbol{\delta}_i)\, \mathrm{d}\mathbf{b}_i \\
&= \prod_{i=1}^{N} f_{Y|\delta}(\boldsymbol{\alpha}, \boldsymbol{\phi}; \mathscr{C}_i) f_\delta(\boldsymbol{\pi}; \boldsymbol{\delta}_i) \\
&= \prod_{i=1}^{N} \prod_{i=1}^{L} [f_{Y|\delta}(\boldsymbol{\alpha}, \boldsymbol{\phi}; \mathscr{C}_{il})\ \pi_l]^{\delta_{il}} \\
&= \left\{ \prod_{i=1}^{N} \prod_{l=1}^{L} [f_{Y|\delta}(\boldsymbol{\alpha}, \boldsymbol{\phi}; \mathscr{C}_i)]^{\delta_{il}} \right\} \left\{ \prod_{l=1}^{L} \pi_l^{\delta_{+l}} \right\}. \qquad (2)
\end{aligned}$$

Here, $f_{Y|b,\delta}$, f_b, and $f_{Y|\delta}$ are multivariate normal density functions and f_δ is the multinomial mass function $\prod_l \pi_l^{\delta_{il}}$; the indicator δ_{il} is equal to one when $D_i = s_l$, and equal to zero otherwise. In (2), $\delta_{+l} = \sum_{i=1}^{N} \delta_{il}$.

Because $\mathscr{L}(\boldsymbol{\theta}; \mathscr{C})$ factors over the parameters $(\boldsymbol{\alpha}, \boldsymbol{\phi})$ and $\boldsymbol{\pi}$, it may be maximized by separately maximizing $f_{Y|\delta}(\boldsymbol{\alpha}, \boldsymbol{\phi}; \mathscr{C})$ and $f_\delta(\boldsymbol{\pi}; \boldsymbol{\delta})$. Many commercially available software packages provide ML or restricted ML (REML) estimates for the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$; for the conditional mixed model, program 5V in BMDP,[23] Proc MIXED in SAS,[24] or the function LME in SPlus[25] give ML estimates for a wide variety of linear mixed models with normally-distributed random effects and residuals. The multinomial piece of the likelihood is maximized at the empirical CDF.

3.2. Obtaining Consistent Estimates when Event Times are Censored

Under independent right censoring, a consistent but not efficient estimate of $\boldsymbol{\pi}$ is obtained from the empirical CDF $\hat{F}_D$. The following two-step procedure allows the analyst to obtain a consistent estimate $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, and to obtain approximate standard errors for functions of $\tilde{\boldsymbol{\theta}}$. This procedure is not fully efficient; in particular, the amount of censoring in D is directly related to the loss of efficiency.

Step 1: Using data from the subset of subjects with complete $\boldsymbol{\delta}_i$ (that is, $\xi_i = 1$), estimate the mixed model parameters $(\boldsymbol{\alpha}, \boldsymbol{\phi})$. This gives consistent estimates $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\phi}})$ of $(\boldsymbol{\alpha}, \boldsymbol{\phi})$, under the assumption that conditional on D_i, censoring of D_i is unrelated to $\mathbf{Y}_i$.

Step 2: Using event time data on all subjects, estimate $\boldsymbol{\pi}$ with $\tilde{\boldsymbol{\pi}}$ using the product-limit estimator; $\tilde{\boldsymbol{\pi}}$ is consistent because the censoring time C_i is independent of $\mathbf{Y}_i$. It is not ML under our model with censored D_i because D_i and $\mathbf{Y}_i$ are dependent.

Consider a (vector valued) function $\mathbf{h}(\boldsymbol{\theta})$ such as $E_\theta[\mathbf{Y}]$, the mean profile from the marginal distribution of $\mathbf{Y}$. Although a robust procedure such as the bootstrap could be used to find standard errors of $\mathbf{h}(\tilde{\boldsymbol{\theta}})$, simple approximations may be found using the delta method. For large samples, $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\phi}}, \tilde{\boldsymbol{\pi}})$ is approximately normally distributed. Making the assumption that $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\phi}})$ is independent of $\tilde{\boldsymbol{\pi}}$, we have

$$\begin{bmatrix} \tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \tilde{\boldsymbol{\phi}} - \boldsymbol{\phi} \\ \tilde{\boldsymbol{\pi}} - \boldsymbol{\pi} \end{bmatrix} \overset{\text{apx.}}{\sim} \mathrm{N}(\mathbf{0}, \mathbf{V})$$

where

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_\alpha & \mathbf{V}_{\alpha\phi} & \mathbf{0} \\ \mathbf{V}_{\phi\alpha} & \mathbf{V}_\phi & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V}_\pi \end{bmatrix}.$$

An estimate $\tilde{\mathbf{V}}$ of $\mathbf{V}$ is obtained in pieces from output of the appropriate packages, and an estimate of var $\{\mathbf{h}(\tilde{\boldsymbol{\theta}})\}$ is $[\mathbf{J}(\tilde{\boldsymbol{\theta}})]' \tilde{\mathbf{V}} [\mathbf{J}(\tilde{\boldsymbol{\theta}})]$, where $\mathbf{J}(\boldsymbol{\theta}) = \partial \mathbf{h}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$. This approach is similar to that advocated by Wu and Bailey;[4,5] that is, to estimate the conditional distribution of $\mathbf{Y}$ given D, identify a parameter of interest in the marginal distribution of $\mathbf{Y}$, and average over $\hat{F}_{\mathrm{D}}$ to obtain an estimate of that parameter. The difference is that while Wu and Bailey require D to be completely observed, this *ad hoc* procedure allows censored observations in D but sacrifices some efficiency in doing so.

3.3. ML Estimation when Event Times are Censored: EM Algorithm

When some $\boldsymbol{\delta}_i$ are incomplete, full use of the data is made by obtaining ML estimates from the observed data likelihood function. This is a straightforward application of the generalized EM algorithm,[20] a widely-used iterative procedure for finding ML estimates from incomplete data. At each iteration, the algorithm updates the parameter estimate by maximizing the expected value of the complete-data log-likelihood as a function of $\boldsymbol{\theta}$, given the observed data and the current parameter estimate.

From (2), it is clear that the complete-data log-likelihood can be expressed as

$$\begin{aligned} \ell(\boldsymbol{\theta}; \mathscr{C}) &= \log \mathscr{L}(\boldsymbol{\theta}; \mathscr{C}) \\ &= \sum_{i=1}^{N} \sum_{l=1}^{L} \delta_{il} \ell_i(\boldsymbol{\theta}; \mathscr{C}_{il}), \end{aligned} \tag{3}$$

where $\ell(\boldsymbol{\theta}; \mathscr{C}_{il}) = \log[\pi_{il} f_{Y|\delta}(\boldsymbol{\alpha}, \boldsymbol{\phi}; \mathscr{C}_{il})]$. Recall that $\mathscr{C}_{il}$ represents an individual's observations with design matrices $\mathbf{X}_i = \mathbf{X}_{il}$ and $\mathbf{Z}_i = \mathbf{Z}_{il}$ configured according to $D_i = s_l$; contributions to the log-likelihood are made only for $\delta_{il} = 1$.

The objective function to be maximized at each iteration of the EM is $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$, the expected value of (3) given the observed data $\mathcal{O}$ and the current update $\boldsymbol{\theta}^{(r)}$. Because only $\boldsymbol{\delta}_i$ is subject to non-response as we define it,

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) &= E[\ell(\boldsymbol{\theta};\mathscr{C})|\mathcal{O},\boldsymbol{\theta}=\boldsymbol{\theta}(r)] \\ &= \sum_{i=1}^{N}\sum_{l=1}^{L} E[\delta_{il}|\mathcal{O}_i,\boldsymbol{\theta}^{(r)}]\ell_i(\boldsymbol{\theta};\mathscr{C}_{il}). \end{aligned}$$

For those with $\boldsymbol{\delta}_i$ completely observed, the conditional expectation $E[\delta_{il}|\mathcal{O}_i,\boldsymbol{\theta}^{(r)}]$ is simply δ_{il} (equal to either one or zero); for those with incomplete $\boldsymbol{\delta}_i$,

$$\begin{aligned} E[\delta_{il}\,|\,\mathcal{O}_i,\boldsymbol{\theta}^{(r)}] &= \Pr\{\delta_{il}=1\,|\,\mathcal{O}_i,\boldsymbol{\theta}=\boldsymbol{\theta}^{(r)}\} \\ &= \frac{\pi_l^{(r)}|\boldsymbol{\Sigma}_{il}^{(r)}|^{-1/2}\exp\{-\tfrac{1}{2}\mathbf{e}_{il}^{\prime(r)}[\boldsymbol{\Sigma}_{il}^{(r)}]^{-1}\mathbf{e}_{il}^{(r)}\}I\{s_l>\tilde{D}_i\}}{\sum_{k=1}^{L}\pi_k^{(r)}|\boldsymbol{\Sigma}_{ik}^{(r)}|^{-1/2}\exp\{-\tfrac{1}{2}\mathbf{e}_{ik}^{\prime(r)}[\boldsymbol{\Sigma}_{ik}^{(r)}]^{-1}\mathbf{e}_{ik}^{(r)}\}I\{s_k>\tilde{D}_i\}} \end{aligned}$$

where $\mathbf{e}_{il}^{(r)}=\mathbf{Y}_i-\mathbf{X}_{il}\boldsymbol{\alpha}^{(r)}$ and $\boldsymbol{\Sigma}_{il}^{(r)}=\boldsymbol{\Sigma}_{il}(\boldsymbol{\phi}^{(r)})$. To simplify notation, let

$$\begin{aligned} \omega_{il}^{(r)} &= E[\delta_{il}|\mathcal{O}_i,\boldsymbol{\theta}^{(r)}] \\ &= \xi_i\delta_{il}+(1-\xi_i)\Pr\{\delta_{il}=1|\mathcal{O}_i,\boldsymbol{\theta}=\boldsymbol{\theta}^{(r)}\}. \end{aligned}$$

Thus, Q has the convenient representation

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})=\sum_{i=1}^{N}\sum_{l=1}^{L}\omega_{il}^{(r)}\ell_i(\boldsymbol{\theta};\mathscr{C}_{il})$$

which exposes it as a linear function of complete-data log-likelihood terms. Complete cases contribute only one term to Q, while incomplete cases make a separate, weighted contribution for every potential event time beyond $\tilde{D}_i$. The sum of weights for each individual is equal to one, and even though Q has more terms than the complete data log-likelihood, the sum of the weights at each iteration of the EM is equal to the number of subjects, that is

$$\sum_{i=1}^{N}\sum_{l=1}^{L}\omega_{il}^{(r)}=N,\quad r=0,1,2,3,\ldots.$$

When maximization of the complete-data log-likelihood is straightforward, so is maximization of $Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(r)})$. In our situation, Q still factors over $(\boldsymbol{\alpha},\boldsymbol{\phi})$ and $\boldsymbol{\pi}$. Finding updates $(\boldsymbol{\alpha}^{(r+1)},\boldsymbol{\phi}^{(r+1)})$ in the M step amounts to maximizing the likelihood for a mixed model fit to observations with non-zero contributions to Q. These observations are treated as independent, but their log-likelihood contributions in the conditional model are weighted with the $\omega_{il}^{(r)}$. The update for $\boldsymbol{\pi}$ is

$$\pi_l^{(r+1)}=N^{-1}\sum_{i=1}^{N}\omega_{il}^{(r)},\ l=1,\ldots,L.$$

Ibrahim[16] describes this procedure in the context of univariate generalized linear models with missing discrete covariates. By assuming that the distribution of D_i is discrete with positive mass at the observed event times, D_i can be viewed as a discrete covariate which is only partially observed under censoring. In principle then, any model for the conditional distribution of $(\mathbf{Y}|D)$ which has a tractable complete-data log-likelihood can be used for the conditional distribution $f_{Y|\delta}$. The result is a relatively easy implementation of the EM algorithm for obtaining ML

estimates. For the linear mixed model, BMDP appears to allow weighted contributions to the Q function – which is a log-likelihood for (1), the conditional piece of the model – through the FREQ option in program 5V. An undocumented programming loophole allows non-integer values for case frequency in 5V, which evidently weight the contributions to the log-likelihood. For each analysis we performed, it provided the same estimates and standard errors as a Fisher scoring algorithm written and implemented in SAS/IML.[26]

3.4. Variance Estimation and the EM Algorithm

Several authors have addressed the issue of estimating $\text{var}\{\hat{\boldsymbol{\theta}}\}$ when $\hat{\boldsymbol{\theta}}$ is the ML estimate obtained via the EM algorithm (for example, Louis[27] and Meilijson).[28] In order to keep terminology and notation straight, what follows is a brief synopsis of large-sample results when $\hat{\boldsymbol{\theta}}$ is the ML estimate obtained from complete data (a more thorough treatment appears in Meilijson[28]). Let $\mathbf{S}(\boldsymbol{\theta}; \mathscr{C}) = \partial \ell / \partial \boldsymbol{\theta}$ be the score (gradient) vector for the complete-data log-likelihood, and $\mathbf{H}(\boldsymbol{\theta}; \mathscr{C}) = \partial^2 \ell / \partial \boldsymbol{\theta}^2$ its Hessian (curvature) matrix. The *observed information matrix* is $\mathbf{J}(\boldsymbol{\theta}; \mathscr{C}) = -\mathbf{H}(\boldsymbol{\theta}; \mathscr{C})$, and the *expected* or *Fisher information matrix* is

$$\mathbf{I}(\boldsymbol{\theta}) = E[\mathbf{J}(\boldsymbol{\theta}; \mathscr{C})] = E[\mathbf{S}(\boldsymbol{\theta}; \mathscr{C})\mathbf{S}(\boldsymbol{\theta}; \mathscr{C})'].$$

When the distribution function of the complete data is from the regular exponential family, the last two expressions are identical. Asymptotically, $\hat{\boldsymbol{\theta}}$ is distributed as $\mathrm{N}(\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta}))$. Two consistent estimators of the Fisher information – and hence of $\text{var}\{\hat{\boldsymbol{\theta}}\}$ – are $\mathbf{I}(\hat{\boldsymbol{\theta}})$ and $\mathbf{J}(\hat{\boldsymbol{\theta}}; \mathscr{C})$; the latter is generally easier to compute and is advocated by Efron and Hinkley.[29] A third consistent estimator of $\mathbf{I}(\boldsymbol{\theta})$ is the *empirical Fisher information* described by Meilijson.[28] The sample covariance matrix of the individual scores $\mathbf{s}_i(\boldsymbol{\theta}; \mathscr{C}_i)$ is

$$\hat{\mathbf{I}}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{s}_i(\boldsymbol{\theta}; \mathscr{C}_i)\, \mathbf{s}_i(\boldsymbol{\theta}; \mathscr{C}_i)' - \frac{1}{N^2} \mathbf{S}(\boldsymbol{\theta}; \mathscr{C})\mathbf{S}(\boldsymbol{\theta}; \mathscr{C})'.$$

When evaluated at $\hat{\boldsymbol{\theta}}$, this reduces to the empirical Fisher information

$$\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{s}_i(\hat{\boldsymbol{\theta}}; \mathscr{C}_i)\mathbf{s}_i(\hat{\boldsymbol{\theta}}; \mathscr{C}_i)'$$

which is consistent for $\mathbf{I}(\boldsymbol{\theta})$ but not necessarily efficient.[28]

When data are incomplete, identifying the *observed data log-likelihood function* $\ell_o(\boldsymbol{\theta}; \mathscr{O})$ – and thus its first and second derivatives $\mathbf{S}_o(\boldsymbol{\theta}; \mathscr{O})$ and $\mathbf{H}_o(\boldsymbol{\theta}; \mathscr{O})$ – is often complicated or impossible; this is precisely the reason for using the EM algorithm. For cases in which the score and Hessian are obtainable from the *complete data log-likelihood*, Louis[27] shows that the observed information matrix $\mathbf{J}_o$, which corresponds to the observed data log-likelihood function $\ell_o(\boldsymbol{\theta}; \mathscr{O})$, satisfies

$$\mathbf{J}_o(\boldsymbol{\theta}; \mathscr{O}) = -E[\mathbf{H}(\boldsymbol{\theta}; \mathscr{C}) | \mathscr{O}] - E[\mathbf{S}(\boldsymbol{\theta}; \mathscr{C})^{\otimes 2} | \mathscr{O}] + \{E[\mathbf{S}(\boldsymbol{\theta}; \mathscr{C}) | \mathscr{O}]\}^{\otimes 2} \tag{4}$$

where $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}'$. When evaluated at the final iteration of the EM algorithm, the last term in (4) is zero. One difficulty in using $\mathbf{J}_o$ is that second derivatives are often difficult to obtain, even from the complete-data log-likelihood. Louis[27] also demonstrates, however, that for any $\boldsymbol{\theta}_0$ in the parameter space,

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta} | \boldsymbol{\theta}_0)|_{\theta = \theta_0} &= E[\mathbf{S}(\boldsymbol{\theta}_0; \mathscr{C}) | \mathscr{O}] \\ &= \mathbf{S}_o(\boldsymbol{\theta}_0; \mathscr{O}). \end{aligned} \tag{5}$$

This says that the derivative of the Q function, the expected complete-data log-likelihood, is equal to the conditional expectation of the complete-data score given observed data, which is in turn the score function from the observed data log-likelihood. The expectation in (5) is generally easy to calculate. Meilijson[28] points out that (5) holds for individual score functions $\mathbf{s}_{oi}(\boldsymbol{\theta}; \mathcal{O}_i)$ as well, so that $\hat{\mathbf{I}}_o(\hat{\boldsymbol{\theta}}; \mathcal{O}_i)$ is consistent for the Fisher information of the observed data log-likelihood.

An obvious advantage to using the empirical Fisher information is that it requires only the score vector for each subject. In mixture models of the joint distribution of $\mathbf{Y}$ and D in which $f_{Y|\delta}$ is multivariate normal and D can be given the multinomial formulation described above, $\mathbf{S}(\boldsymbol{\theta}; \mathscr{C})$ has components

$$\mathbf{S}_\alpha = \sum_{i=1}^{N} \mathbf{X}_i' \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\alpha})$$

$$\mathbf{S}_{\phi_q} = \frac{1}{2} \sum_{i=1}^{N} \operatorname{trace} \{\boldsymbol{\Sigma}_i^{-1} (\mathbf{e}_i \mathbf{e}_i' - \boldsymbol{\Sigma}_i) \boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_{iq}\} \quad (q = 1, \ldots, Q)$$

$$\mathbf{S}_{\pi_l} = \sum_{i=1}^{N} \left[\frac{\delta_{il}}{\pi_l} - \frac{\delta_{iL}}{\pi_L} \right] \quad (l = 1, \ldots, L-1)$$

where $\mathbf{S}_\theta = \partial \ell(\boldsymbol{\theta}; \mathscr{C}) / \partial \boldsymbol{\theta}$ and $\dot{\boldsymbol{\Sigma}}_{iq} = \partial \boldsymbol{\Sigma}_i / \partial \phi_q$.[31] At convergence, $\hat{\omega}_{il} = \Pr\{\delta_{il} = 1 | \mathcal{O}, \hat{\boldsymbol{\theta}}\}$, and

$$\mathbf{s}_{oi}(\hat{\boldsymbol{\theta}}; \mathcal{O}_i) = \sum_{l=1}^{L} \hat{\omega}_{il} \mathbf{s}_i(\hat{\boldsymbol{\theta}}; \mathscr{C}_{il}).$$

Inverting the matrix $\sum_{i=1}^{N} \mathbf{s}_{oi}(\hat{\boldsymbol{\theta}}; \mathcal{O}_i) \mathbf{s}_{oi}(\hat{\boldsymbol{\theta}}; \mathcal{O}_i)'$ gives a consistant estimate of $\operatorname{var}\{\hat{\boldsymbol{\theta}}\}$.

When the sample size is not large enough for the empirical Fisher information to give reliable results, the observed information matrix – evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ – should be used for variance estimation; the Appendix contains formulae for the required second derivative matrices. A reasonable way to check whether standard errors estimated via empirical Fisher information are tenable is to compare them to standard errors obtained when using the *ad hoc* strategy for consistent estimation outlined above. Increased efficiency should be evident when comparing ML estimates from the mixed model to those obtained using the *ad hoc* procedure.

4. EXAMPLE: SCHIZOPHRENIA DATA

4.1. Background and Data Summary

A recent trial to compare various doses of a new therapy (NT) for schizophrenia against a standard therapy (ST) enrolled 242 patients and followed their progress for six weeks. Three doses – low, medium and high – of NT were administered; we compare the 61 patients on medium dose NT to the 63 on ST in order to illustrate our methods; the four-arm comparison which includes ST and the three NT groups is a straightforward extension.

Schizophrenia status was assessed using the Brief Psychiatric Rating Scale, or BPRS.[17] Scores range from 0 to 108, with higher scores indicating more severe symptoms. Assessments were carried out at randomization (week zero) and at weeks 1, 2, 3, 4 and 6. Patients were hospitalized for the first four weeks, with discharge permitted thereafter. The primary objective of the study is to compare the change from baseline to week 6 in the four groups.

For various reasons, including those directly related to drug efficacy, several patients drop out of the trial before week 6. An analysis of week-6 scores which ignores dependence of the repeated outcomes on the dropout process runs the risk of producing biased results. We model the joint

Table I. Break down of completers and dropouts, by treatment group, for schizophrenia trial

Treatment	Completers	Dropouts by reason			
		Adverse experience	Lack of efficacy	Other reason	Total
New	40	1	7	13	61
Standard	34	12	11	6	63

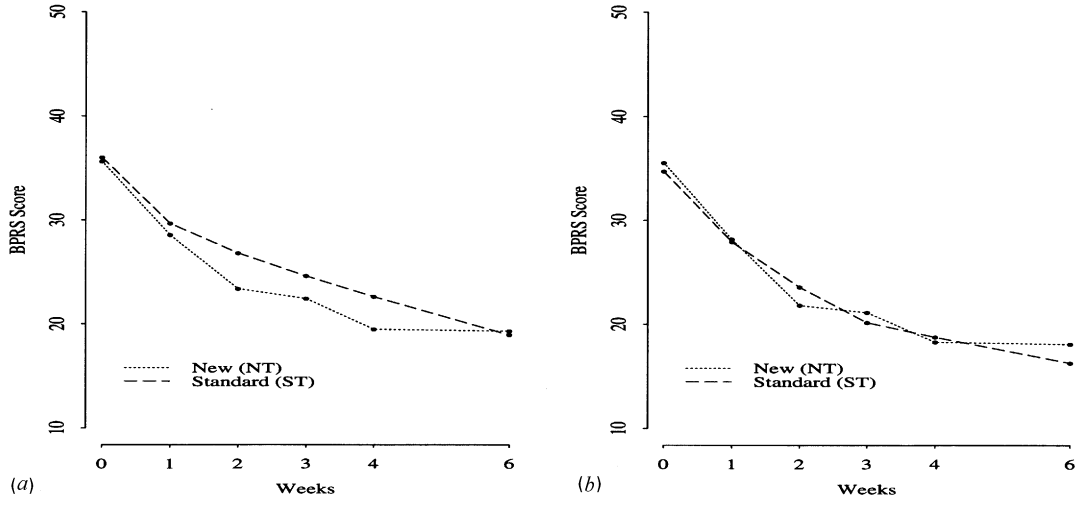

Figure 1. (*a*) Sample means for BPRS score from all available data on each treatment, including those whose dropout is treatment related (informative). (*b*) Sample means from those who either complete protocol or are non-informative dropouts

distribution of dropout time and repeated BPRS score, and integrate over the dropout distribution in order to obtain adjusted week-6 estimates of mean BPRS score. The study protocol called for the BPRS to be administered at the time of removal from study, but not subsequently; hence, a few patients who 'drop out' have BPRS scores at all six occasions.

Regarding the dropout process, patients fall into one of four categories: (1) completed protocol, (2) dropped out due to adverse experience (for example, side-effects); (3) dropped out due to lack of treatment effect, and (4) dropped out for other reasons (including patient improvement). We consider (3) to be outcome-related, treating dropout times from this category as *informative* ($\xi = 1$). Reasons (2) and (4) are not considered outcome-related, and corresponding dropout times are treated as censored observations of true dropout time. It is possible, of course, to make other choices about which dropout times are informative. Of the 21 patients who drop out on NT, 7 are for lack of efficacy (33 per cent); on the ST arm, 11 of 39 dropouts leave because of lack of efficacy (28 per cent). Refer to Table I for a summary of observation patterns in the two groups.

Figure 1(*a*) shows trends in BPRS score using all available data, including the informative dropouts; it suggests that NT outperforms ST during weeks 1 through 4, but this result could be biased by differential dropout in the two treatment arms. Figure 1(*b*) is a plot of sample means using only those who complete the trial or drop out for non-outcome-related reasons; this suggests no treatment difference. Figure 2 shows sample means for each treatment group,

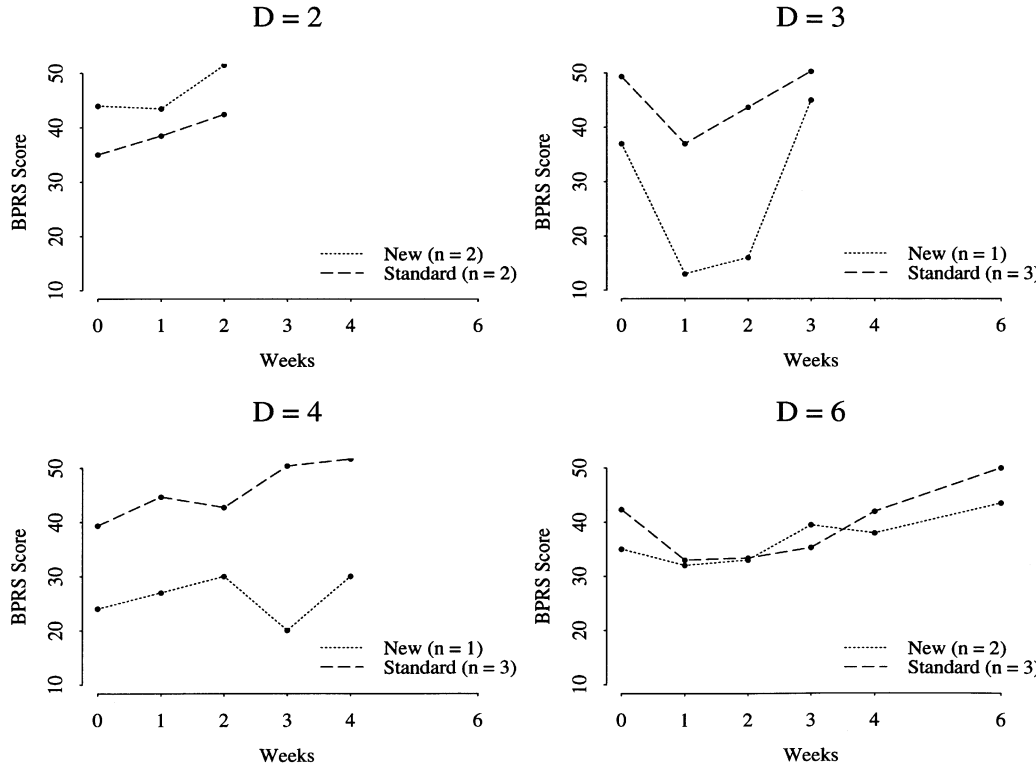

Figure 2. Sample means for each treatment group, stratified by time of treatment-related (informative) dropout. One subject on NT dropped out after a baseline BPRS score of 35

stratified by week of informative dropout. In contrast with those who complete the trial, informative dropouts exhibit rising BPRS scores over time. Figure 2 suggests that outcome and time to dropout are related, and that this relationship should be taken into account if inferences on mean BPRS are to be made.

4.2. Choosing a Model

Because there are a maximum of only six event-time outcomes in each treament group (five times for dropout plus a category for completers), we take the pattern-mixture approach[10] to modelling the schizophrenia data, which calls for a separate regression model to be fit for each incompleteness pattern (in this case, dropout time). The model we begin with has separate regression coefficients for each dropout time and each treatment group, but variance components which are common across dropout times and treatment groups. Strata are combined for a more parsimonious model, and lack of fit is checked with the chi-square approximation to the log-likelihood-ratio statistic. Models are fit using an EM algorithm written in SAS/IML.[26]

Figures 1 and 2 suggest that BPRS follows a quadratic curve over time, with potentially different curves for each dropout time and for each treatment group. We begin by fitting a 'full' model which describes a different second-degree curve for each dropout/treatment combination. The exception is those who drop out at week 2; their BPRS is assumed to be linear over time. One

Table II. Estimates of time trend by treatment and dropout time, from reduced model. Because orthogonal polynomials are used, the intercept corresponds to estimated mean BPRS score after $2\frac{2}{3}$ weeks. Standard errors in parentheses

Week of dropout	Treatment	Number observed	Intercept	Linear trend	Quadratic trend
0, 2		5	48·7 (2·87)	4·22 (0·91)	
3, 4	New	2	36·4 (2·06)	†	0·72 (0·07)
	Standard	6	48·8 (2·67)	†	‡
6		5	*	2·12 (0·66)	‡
Complete		74	23·5 (0·67)	−2·92 (0·16)	‡
Non-informative dropouts	New	14			
	Standard	18			

* Dropouts at weeks 3 and 4 (NT group) are pooled with dropouts at week 6 (NT and ST) to estimate a common intercept
† Dropouts at weeks 0, 2, 3 and 4 have a common linear trend
‡ Completers, along with dropouts at weeks 3, 4 and 6 have a common quadratic trend

subject in the NT group, who had only a baseline measurement, is included in the week-2-dropout stratum for the model. This model also assumes between-subject variability in the intercept, linear trend and quadratic trend. For efficient parameter estimation, orthogonal polynomials are used to describe time trends. In total, the full model has 28 regression parameters ($\boldsymbol{\alpha}$), seven variance parameters ($\boldsymbol{\phi}$), and nine distinct parameters to describe dropout in both treatment groups ($\boldsymbol{\pi}_0, \boldsymbol{\pi}_1$).

Because of the small number of dropouts in each treatment group at each time of dropout, some of the $\boldsymbol{\alpha}$ parameters are poorly determined; combining observations across strata is a feasible way to reduce the number of regression parameters needed to describe the data. For example, Figure 2 suggests that treatment groups can be combined to make one estimated curve for those with $D = 6$. Several reduced models were compared to the full model, with lack of fit determined by likelihood ratio test. A more parsimonious model, with eight regression parameters, is summarized in Table II and Figure 3. The apparent over-prediction in the NT arm for dropouts at weeks 3 or 4 results from a common mean being fit to that stratum and those who drop out at week 6. The chi-square statistic used to approximate the likelihood ratio test for lack of fit gives $\chi^2_{20} = 28{\cdot}2$, with $p = 0{\cdot}104$.

The effect of dropout on BPRS trends is seen by comparing intercepts and linear and quadratic coefficient estimates among the dropout strata (Table II). Early withdrawal for treatment-related reasons is related to higher estimated BPRS: for example, those who complete the study have the lowest intercept (23·5) and a negative linear trend (−2·9 points per week); early dropouts (week 2 or before) have the highest intercept (48·7) and a linear increase of about 4 points per week. Because orthogonal polynomials are used to model time trends, the intercept corresponds to mean score after $2\frac{2}{3}$ weeks. To test for presence of informative dropout, a model with four regression parameters (intercept, main effect of treatment, linear and quadratic trends) was compared to the reduced model; the likelihood ratio test gives $\chi^2_4 = 87{\cdot}4$.

Figure 3 compares model-based estimates to observed means in each dropout/treatment stratum. For all dropout times except 3 and 4, BPRS scores have a common time trend across treatment group; for those dropping out at week 3 or 4, the treatment difference apparently depends upon baseline differences. For a given dropout time, no significant treatment-by-time-trend interaction is evident.

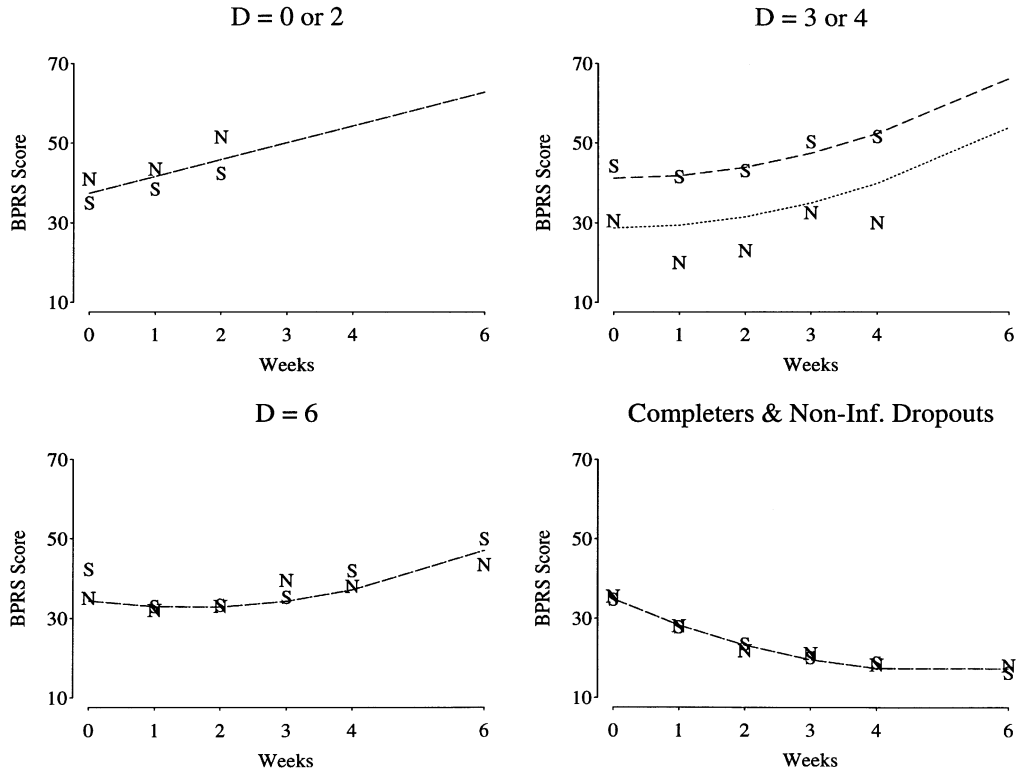

Figure 3. Model-based estimates and observed means, by time of informative dropout and by treatment. 'N' is new treatment, 'S' is standard treatment. A common time trend between treatments is assumed in each stratum except dropout at week 3 or 4

The graphs in Figure 3 represent the estimated conditional mean BPRS scores, given dropout time. In order to make inference about the unconditional means, it is necessary to account for the different rates of dropout between treatments. Table III contains estimated dropout probabilities and standard errors for each treatment, using both the Kaplan–Meier and mixture model estimates; the estimated completion rate is 85 per cent on new therapy compared with 75 per cent on the standard therapy. Although this difference is not statistically significant ($|Z| = 1{\cdot}34$, computed from mixture model), estimates of BPRS which are adjusted for differential dropout rates are affected. By averaging model-based estimates of conditional mean BPRS over the treatment-specific dropout probabilities, we obtain an estimated profile of mean BPRS scores over the course of the study (Figure 4).

The effect of different dropout rates is seen by comparing Figure 4, which suggests a treatment difference, with the bottom right panel of Figure 3 (completers and non-informative dropouts only), which does not. The profiles in Figure 4 represent the average BPRS scores which would have been observed had the patients been maintained on their initial treatment; this also assumes that the on-treatment trajectories would continue to hold beyond dropout time had the patients remained on treatment. In this setting, where drug therapies are not uniformly effective for all patients, it is at least plausible to assume that the informative dropouts continue to get worse even if they remain on treatment. The profiles in Figure 4 reflect the fact that fewer of the patients who

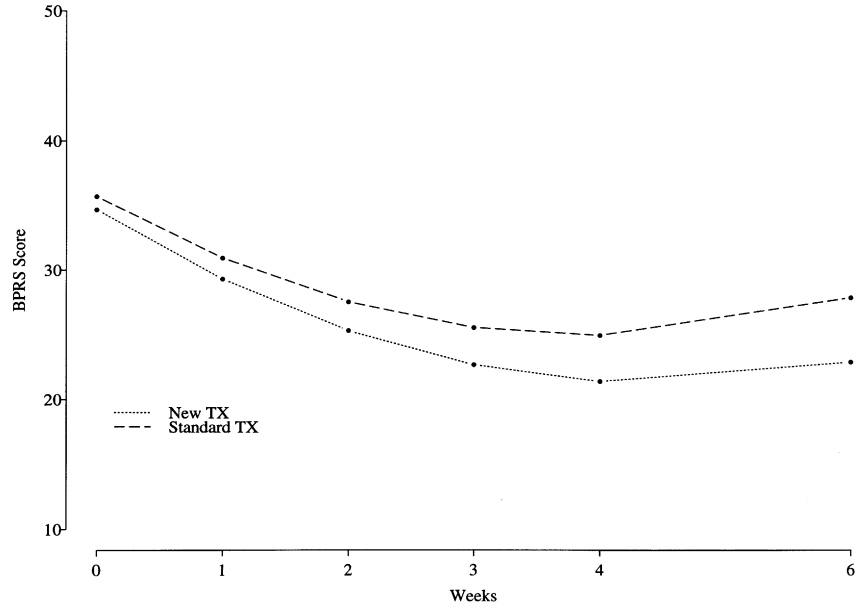

Figure 4. Model-based estimate of profile of mean BPRS scores, averaging the conditional means from Figure 3 over the dropout probabilities in Table III. The difference at week 6 is attributable to differential completion rate between treatments

are randomized to standard therapy stand to benefit from that therapy, something which cannot be discerned by analyzing completers only. Conditional on a specified duration of treatment, new therapy appears to have the same efficacy as standard therapy; however, a higher proportion remain on new therapy. Comparing net effects therefore favours the new therapy. The dropout-adjusted estimates of mean BPRS are 22·9 for new therapy and 27·9 for the standard therapy, a difference of 5·0 points (S.E. 3·17) with 95 per cent confidence interval $(-1{\cdot}2, 11{\cdot}2)$. Standard error is approximated using the delta method (see Appendix). A test for significant difference in net benefit against a two-sided alternative gives $|Z| = 1{\cdot}57$, which provides weak evidence ($p = 0{\cdot}112$) in favour of the new treatment for schizophrenia. Note that no treatment-by-time interaction appears in the conditional model of $f_{Y|\delta}$, but such an interaction appears in Figure 4. This results from differential rates of informative dropout between treatment groups.

It is worthwhile to compare mixed model results to standard approaches which ignore dropout. When a linear mixed model (treatment effect, linear and quadratic trends, and treatment-by-time-trend interactions) is fit to all available repeated measures under the assumption that missingness is MAR in the marginal distribution of $\mathbf{Y}_i$, estimated week-6 means are 20·8 (NT) and 23·9 (ST). Downward bias is evident, especially on the ST arm where informative dropout is more prevalent. The estimated difference is 3·1, with 95 per cent confidence interval $(-3{\cdot}0, 9{\cdot}3)$. Fitting the same model to only those who complete the trial actually favours the standard therapy; estimated week-6 BPRS means are 18·3 (NT) and 17·2 (ST), a difference of $-1{\cdot}1$ with 95 per cent confidence limits $(-6{\cdot}3, 4{\cdot}1)$. This result is consistent with the absence of treatment effect for completers in the reduced mixed model (Table II).

Table III. Estimates and standard errors for probability of remaining on protocol at a given week, by treatment group. The Kaplan–Meier (KM) estimator uses only information about dropout times, and the mixture model (MM) estimates incorporate information from BPRS scores· Efficiency comparisons are made using variance ratios

Treatment	Estimator	Week				
		0	2	3	4	6
New	KM	0·9836 (0·0163)	0·9465 (0·0301)	0·9264 (0·0356)	0·9048 (0·0408)	0·8617 (0·0489)
	MM	0·9836 (0·0106)	0·9439 (0·0257)	0·9239 (0·0291)	0·8991 (0·0369)	0·8531 (0·0411)
	Variance ratio	2·37	1·37	1·50	1·22	1·42
Standard	KM		0·9649 (0·0244)	0·9082 (0·0392)	0·8462 (0·0503)	0·7794 (0·0593)
	MM		0·9623 (0·0202)	0·9011 (0·0357)	0·8337 (0·0465)	0·7500 (0·0570)
	Variance ratio		1·45	1·20	1·17	1·08

4.3. Efficiency

An added feature of a model for the joint distribution of repeated outcomes and an event time is increased efficiency in estimating event-time distribution; that is, when the repeated outcomes and event times are dependent, estimation of the event-time distribution uses the extra information provided by the repeated outcomes. Although primary interest for the schizophrenia example is in mean BPRS scores, the dropout probability estimates from the likelihood-based model are more efficient than estimates obtained via Kaplan–Meier. No extra distributional assumptions are placed on the D_i, but the information in BPRS score is used to increase efficiency. Table III compares on-treatment probability estimates and their standard errors from the Kaplan-Meier estimator alone and from the likelihood model which makes use of information in the BPRS scores; variance ratios indicate that using the likelihood model results in moderate efficiency gains (between 8 per cent to 50 per cent, with a notable exception of 137 per cent in the dropout category with only one observation).

Just as the repeated measures allow more precise estimation of the event time distribution, the censored event times provide extra information with which to estimate the regression parameters. Section 3 gives a method for consistent estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$ using only non-censored dropout times; Table IV shows a comparison of standard errors for these estimates against the full likelihood model, which incorporates information from censored event times. The estimates are roughly the same, which is expected because both are consistent for the true values $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$; the only difference is inclusion of non-informative dropouts in the mixture model. For the mean parameters, the largest gains in efficiency are seen in strata which are sparsely populated with informative dropouts (means estimated by $\alpha_1, \alpha_3, \alpha_4$). Including the censored subjects reduces variability in the variance parameter estimates, with the most benefit going to the estimated intercept variance ϕ_1 and within-subject residual variance ϕ_7.

Table IV. Comparing parameter estimates and standard errors between two methods: (i) Estimating $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$ using only non-censored observations (NCO), and (ii) estimating the same parameters with all data from the mixture model (MM)

Parameter	Interpretation	Estimate		Standard error	
		MM	NCO	MM	NCO
α_0	Mean at week 2·67 (completers)	23·4	24·0	0·67	1·23
α_1	Difference in mean for $D = 0, 2$	25·2	24·4	3·08	5·34
α_2	Linear trend, $D = 0, 2, 3, 4$	4·2	3·9	0·92	1·01
α_3	Difference in mean for $D = 3, 4, 6$	13·0	12·8	2·37	4·01
α_4	TX difference, $D = 3, 4$	12·4	13·9	3·01	5·84
α_5	Quadratic trend	0·7	0·8	0·07	0·08
α_6	Linear trend, $D = 6$	2·1	1·8	0·66	0·75
α_7	Linear trend, completers	−2·9	−3·0	0·16	0·21
ϕ_1	Var (int)	109·8	108·6	10·34	16·91
ϕ_2	Cov (int, lin)	4·6	3·4	1·76	2·16
ϕ_3	Var (lin)	2·2	2·1	0·40	0·54
ϕ_4	Cov (int, quad)	−3·3	−2·8	0·63	0·89
ϕ_5	Cov (lin, quad)	−0·66	−0·57	0·12	0·16
ϕ_6	Var (quad)	0·24	0·18	0·07	0·09
ϕ_7	Residual variance	30·4	29·9	1·14	2·63

5. DISCUSSION

We have described a general likelihood-based mixture-model framework for modelling the joint distribution of repeated outcomes and event times. Our approach accommodates missing data in the repeated measures and right censoring of event times, and incorporates non-ignorable missingness in the marginal distribution of the repeated measures. ML estimation with the EM algorithm amounts to weighting the log-likelihood contributions of those with censored D_i across the potential realizations of D_i. The potential failure times are drawn from the collection of those times already observed.

Although the example discussed in Section 4 uses dropout due to lack of effect as the outcome-related event, the model does not require that repeated outcomes cease to be observed following the event; in fact, it describes the distribution of the entire **Y**-vector. For example, a patient may have different pre-event and post-event time trends. This is plausible when an investigator desires an *intention-to-treat* analysis, and measurements are taken on patients during and after compliance with assigned therapy. In this case, D_i represents length of time on assigned treatment (see Hogan and Laird[30]).

A potential limitation to this approach when applied to longitudinal data with informative dropout is the required modelling of time trends as a function of dropout time; this is an inherently paradoxical situation. Efficiency is gained for estimating the joint distribution parameters when subjects accrue repeated measures *and* experience events (that is, leave the study); from the standpoint of making inference about the marginal distribution of **Y**, however, a high attrition rate is undesirable. Our approach is useful because it allows the dropout process to be modelled in different ways. A comprehensive data analysis would compare results of several dependence structures between outcome and dropout time to assess sensitivity to model specification.

The mixture model approach has potential for improving efficiency in survival curve estimation. Our data analysis suggests that the mixture-model ML estimate of the event-time

distribution, which makes use of information from the repeated measures, is more efficient than the Kaplan–Meier estimator alone. Future investigations might examine (i) how gains in efficiency are related to underlying model parameters, and (ii) how efficiency is affected by misspecification of the relationship between the repeated measures and the event time.

Finally, we remark that mixture models for joint distribution of outcome and event time can be used in any situation where a parametric form for the mean of the repeated outcomes can be specified, such as with binary or categorical responses.

APPENDIX I: SECOND DERIVATIVES FOR CALCULATING OBSERVED INFORMATION

Variance estimation which uses the observed information matrix (4) from the observed data log-likelihood requires computing second derivatives of the observed data log-likelihood. Louis[27] shows that these are obtained by differentiating $\mathcal{Q}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}})$ at convergence of the EM algorithm. Jennrich and Schluchter[31] provide second derivatives of the complete-data log-likelihood for the general linear mixed model with arbitrary covariance structure; because Q is just a linear combination of complete-data log-likelihood terms, its second derivatives follow straightaway.

Let $\mathbf{H}_{\theta\theta} = \partial^2 \ell(\boldsymbol{\theta}; \mathscr{C})/\partial \boldsymbol{\theta}^2$ be the second derivative of the complete-data log-likelihood, $\mathbf{x}_{ir}$ the rth row of the design matrix $\mathbf{X}_i$, and $\mathbf{e}_i = \mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\alpha}$ an individual's residual vector. Derivatives of $\boldsymbol{\Sigma}_i(\boldsymbol{\phi})$ with respect to elements of $\boldsymbol{\phi}$ are denoted by $\dot{\boldsymbol{\Sigma}}_{iq} = \partial \boldsymbol{\Sigma}_i(\boldsymbol{\phi})/\partial \phi_q$ and $\ddot{\boldsymbol{\Sigma}}_{iqr} = \partial^2 \boldsymbol{\Sigma}_i(\boldsymbol{\phi})/\partial \phi_q \partial \phi_r$. Using this notation, sub-matrices of $\mathbf{H}_{\theta\theta}$ are

$$\mathbf{H}_{\alpha\alpha} = -\sum_{i=1}^{N} \mathbf{X}_i' \boldsymbol{\Sigma}_i^{-} \mathbf{X}_i$$

$$\mathbf{H}_{\alpha_p \phi_p} = -\sum_{i=1}^{N} \mathbf{x}_{ip}' \boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_{iq} \boldsymbol{\Sigma}_i^{-1} \mathbf{e}_i \quad (p = 1, \ldots, R \text{ and } q = 1, \ldots, Q)$$

$$\mathbf{H}_{\phi_q \phi_r} = -\frac{1}{2} \sum_{i=1}^{N} \operatorname{trace}\{\boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_{iq} \boldsymbol{\Sigma}_i^{-1} (2\mathbf{e}_i \mathbf{e}_i' - \boldsymbol{\Sigma}_i) \boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_{ir}\} + \frac{1}{2} \sum_{i=1}^{N} \operatorname{trace}\{\boldsymbol{\Sigma}_i^{-1} (\mathbf{e}_i \mathbf{e}_i' - \boldsymbol{\Sigma}_i) \boldsymbol{\Sigma}_i^{-1} \ddot{\boldsymbol{\Sigma}}_{iqr}\},$$
$$(q, r = 1, \ldots, Q)$$

$$\mathbf{H}_{\pi_l \pi_l} = -\sum_{i=1}^{N} \left[\frac{\delta_{il}}{\pi_l^2} + \frac{\delta_{iL}}{\pi_L^2} \right] \quad (l = 1, \ldots, L-1),$$

$$\mathbf{H}_{\pi_k \pi_l} = \sum_{i=1}^{n} \frac{\delta_{iL}}{\pi_L^2} \quad (k \neq l \text{ and } k, l = 1, \ldots, L-1).$$

Mixed partial derivatives involving $\boldsymbol{\pi}$ and either $\boldsymbol{\alpha}$ or $\boldsymbol{\phi}$ are zero.

A random effects variance structure simplifies $\mathbf{H}_{\phi_q \phi_r}$ because $\boldsymbol{\Sigma}_i(\boldsymbol{\phi})$ is a linear function of the Q variance parameters; in particular, the second derivative $\ddot{\boldsymbol{\Sigma}}_i$ is zero. Dropping the subscript i, recall that $\boldsymbol{\Sigma}(\boldsymbol{\phi}) = \mathbf{Z}\boldsymbol{\Phi}\mathbf{Z}' + \phi_Q \mathbf{I}$, where $\boldsymbol{\Phi}$ is a $P \times P$ symmetric matrix composed of $\phi_1, \ldots, \phi_{Q-1}$ and $\mathbf{Z}$ is an $m \times P$ design matrix with elements z_{kl}. Let ϕ_{pq} represent the p, q element of $\boldsymbol{\Phi}$, so that $\phi_{pq} = \phi_{qp}$. Then the i, j element of $\boldsymbol{\Sigma}$ is

$$\sigma_{ij} = \sum_{p=1}^{P} \sum_{q=1}^{P} z_{ip} \phi_{pq} z_{jq} + \phi_Q I\{i = j\}$$

which is a linear combination of $\boldsymbol{\phi}$.

APPENDIX II: DELTA METHOD APPROXIMATIONS

The curves in Figure 4 represent estimated unconditional mean profiles of $\mathbf{Y}$, and are weighted averages of conditional mean profiles over the dropout times. Because the weighted averages are non-linear functions of the parameters, the delta method is used to approximate standard errors. We illustrate here how to approximate the variance-covariance matrix of an estimated treatment mean profile once an estimate of $\mathrm{var}\{\hat{\boldsymbol{\theta}}\}$ is found.

Let $\mathbf{X}_{gl}$ represent the design matrix for a subject in treatment group g who experiences an event at time s_{gl}. In our example, $g = 1, 2$ represents (1) standard and (2) new therapy, $\mathscr{D}_1 = \{2, 3, 4, 6, C\}$, and $\mathscr{D}_2 = \{0, 2, 3, 4, 6, C\}$, where C denotes protocol completion. The design on time includes all scheduled measurement times. In terms of $\boldsymbol{\theta}$, the expected profile mean for group g is

$$\mathbf{h}_g(\boldsymbol{\alpha}, \boldsymbol{\pi}_g) = \sum_{l=1}^{L_g} \pi_{gl} \mathbf{X}_{gl} \boldsymbol{\alpha}.$$

Let $\mathbf{V}_g$ be the sub-matrix of variances and covariances of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\pi}}_g$. Then the Jacobian $\mathbf{J}_g$ for group g consists of

$$\partial \mathbf{h}/\partial \boldsymbol{\alpha} = \overline{\mathbf{X}}_g$$

$$\partial \mathbf{h}/\partial \pi_g = [(\mathbf{X}_{g1} - \mathbf{X}_{L_g})\boldsymbol{\alpha}(\mathbf{X}_{g2} - \mathbf{X}_{L_g})\boldsymbol{\alpha} \ldots (\mathbf{X}_{g, L_g - 1} - \mathbf{X}_{L_g})\boldsymbol{\alpha}]$$

and the variance-covariance matrix of the group's estimated mean profile is approximated by evaluating $\mathbf{J}_g \mathbf{V}_g \mathbf{J}_g'$ at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. We calculated a standard error on the unconditional week-6 difference by assuming independence between treatment groups.

ACKNOWLEDGEMENTS

Support for this work was provided by National Institutes of Health through grants GM29745 and MH17119. The authors gratefully acknowledge Andrea Rotnitzky, Daniel Scharfstein, Kirstin Schulz and two anonymous referees for helpful comments on the manuscript.

REFERENCES

1. Self, S. and Pawitan, Y. 'Modeling a marker of disease progression and onset of disease', *in* (Jewell, N., Dietz, K. and Farewell, V. (eds), *AIDS Epidemiology*: *Methodological Issues*, Birkhäuser, Boston, 1992.
2. Tsiatis, A. A., DeGruttola, V. and Wulfsohn, M. S. '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*, **90**, 27–37 (1994).
3. DeGruttola, V. and Tu, X. M. 'Modeling progression of CD4-lymphocyte count and its relationship to survival time', *Biometrics*, **50**, 1003–1014 (1994).
4. Wu, M. C. and Bailey, K. R. 'Analysing changes in the presence of informative right censoring caused by death and withdrawal', *Statistics in Medicine*, **7**, 337–346 (1988).
5. Wu, M. C. and Bailey, K. R. 'Estimation and comparison of changes in the presence of informative right censoring: conditional linear model', *Biometrics*, **45**, 939–955 (1989). Correction: **46**, 88 (1990).
6. Wu, M. C. and Carroll, R. J. 'Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process', *Biometrics*, **44**, 175–188 (1988).
7. Schluchter, M. D. 'Methods for the analysis of informatively censored longitudinal data', *Statistics in Medicine*, **11**, 1861–1870 (1992).
8. Mori, M., Woodworth, G. G. and Woolson, R. F. 'Application of empirical Bayes inference to estimation of rate of change in the presence of informative right censoring', *Statistics in Medicine*, **11**, 621–631 (1992).
9. Diggle, P. and Kenward M. G. 'Informative drop-out in longitudinal data analysis', *Applied Statistics*, **43**, 49–93 (1994).

10. Little, R. J. A. 'Pattern-mixture models for multivariate incomplete data', *Journal of the American Statistical Association*, **88**, 125–134 (1993).
11. Little, R. J. A. 'Modeling the drop-out mechanism in repeated-measures studies', *Journal of the American Statistical Association*, **90**, 1112–1121 (1995).
12. Laird, N. M. and Ware, J. H. 'Random-effects models for longitudinal data', *Biometrics*, **38**, 963–974 (1982).
13. Cox, D. R. 'Regression models and life tables', *Journal of the Royal Statistical Society*, *Series B*, **34**, 187–220 (1972).
14. Dawson, J. D. and Lagakos, S. W. 'Size and power of two-sample tests of repeated measures data', *Biometrics*, **49**, 1022–1032 (1993).
15. Wu, M. C., Hunsberger, S. and Zucker, D. 'Testing for differences in changes in the presence of censoring: parametric and non-parametric methods', *Statistics in Medicine*, **13**, 635–646 (1994).
16. Ibrahim, J. G. 'Incomplete data in generalized linear models', *Journal of the American Statistical Association*, **85**, 765–769 (1990).
17. Overall, J. E. and Gorham, D. R. 'The Brief Psychiatric Rating Scale (BPRS): recent developments in ascertainment and scaling', *Psychopharmacology Bulletin*, **21**, 97–99 (1988).
18. Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data*, Wiley, New York, 1987.
19. Kaplan, E. L. and Meier, P. 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association*, **53**, 457–481 (1958).
20. Dempster, A. P., Laird, N. M. and Rubin, D. B. 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society*, *Series B*, **39**, 1–22 (1977).
21. Johansen, S. 'The product limit estimator as maximum likelihood estimator', *Scandinavian Journal of Statistics*, **5**, 195–199 (1978).
22. Cox, D. R. and Oakes, D. *Analysis of Survival Data*, Chapman and Hall, London, 1984.
23. Schluchter, M. D. '5V: Unbalanced repeated measures models with structured covariance matrices', *in* Dixon, W. J. (ed), *BMDP Statistical Software Manual*, Vol. 2, University of California Press, Berkeley, 1990, pp. 1207–1244, 1322–1327.
24. SAS Institute Inc. SAS Technical Report P-229, *SAS/STAT Software*: *Changes and Enhancements*, *Release* 6.07, Cary NC, SAS Institute Inc., Cary, North Carolina, 1992.
25. Pinheiro, J. C. and Bates, D. M. 'Mixed effects models and classes for S and SPlus', Technical Report 89, Department of Biostatistics, University of Wisconsin, Madison, 1995.
26. SAS Institute Inc. *SAS/IML Software*: *Usage and Reference*, *Version 6*, *First Edition*, SAS Institute Inc., Cary, North Carolina, 1989.
27. Louis, T. A. 'Finding the observed information matrix when using the EM algorithm', *Journal of the Royal Statistical Society*, *Series B*, **44**, 226–233 (1982).
28. Meilijson, I. 'A fast improvement to the EM algorithm on its own terms', *Journal of the Royal Statistical Society*, *series B*, **51**, 127–138 (1989).
29. Efron, B. and Hinkley, D. V. 'Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information', *Biometrika*, **65**, 457–488 (1978).
30. Hogan, J. W. and Laird, N. M. 'Intention-to-treat analysis for incomplete repeated measures which depend on event times', *Biometrics*, in press, 1996.
31. Jennrich, R. I. and Schluchter, M. D. 'Unbalanced repeated-measures models with structured covariance matrices', *Biometrics*, **42**, 805–820 (1986).