Biostatistics (2000), **1**, 4, *pp.* 465–480
Printed in Great Britain

Joint modelling of longitudinal measurements and event time data

ROBIN HENDERSON*, PETER DIGGLE, ANGELA DOBSON
Medical Statistics Unit, Lancaster University, LA1 4YF, UK
robin.henderson@lancaster.ac.uk

SUMMARY

This paper formulates a class of models for the joint behaviour of a sequence of longitudinal measurements and an associated sequence of event times, including single-event survival data. This class includes and extends a number of specific models which have been proposed recently, and, in the absence of association, reduces to separate models for the measurements and events based, respectively, on a normal linear model with correlated errors and a semi-parametric proportional hazards or intensity model with frailty.

Special cases of the model class are discussed in detail and an estimation procedure which allows the two components to be linked through a latent stochastic process is described. Methods are illustrated using results from a clinical trial into the treatment of schizophrenia.

Keywords: Biomarkers; Counting processes; Informative drop-out; Repeated measurements; Serial correlation; Survival.

1. INTRODUCTION

Many scientific investigations generate both *longitudinal measurement data*, with repeated measurements of a response variable at a number of time points, and *event history data*, in which times to recurrent or terminating events are recorded. A well-known example is in AIDS research in which a biomarker such as CD4 lymphocyte count is determined intermittently and its relationship with time to seroconversion or death is of interest (e.g. Pawitan and Self, 1993; Tsiatis *et al.*, 1995; Wulfsohn and Tsiatis, 1997). Another example, to be considered in detail later, is illustrated in Figure 1. These data arose as part of a clinical trial into drug therapy for schizophrenia patients, previously described by Diggle (1998). The upper plot shows the development over time of mean scores for each of three treatment groups on a particular measure of psychiatric disorder (Positive and Negative Symptom Scale, PANSS). Not all patients completed the trial however: survival curves in the lower plot show that a substantial proportion of each group of patients withdrew before completing the measurement schedule. It is not clear whether the apparent decrease in PANSS profiles relects a genuine change over time, or is an artefact caused by selective drop-out, with patients with high (worse) PANSS values being less likely to complete the trial.

Hogan and Laird (1997b) give an excellent review of models and methods for joint analysis of data of this type. As well as the AIDS studies cited earlier, other useful references include Faucett and Thomas (1996), Lavalley and Degrutolla (1996), Hogan and Laird (1997a), Faucett *et al.* (1998) and Finkelstein and Schoenfeld (1999). Most previous work has been based on specific applications, and clearly the

*To whom correspondence should be addressed

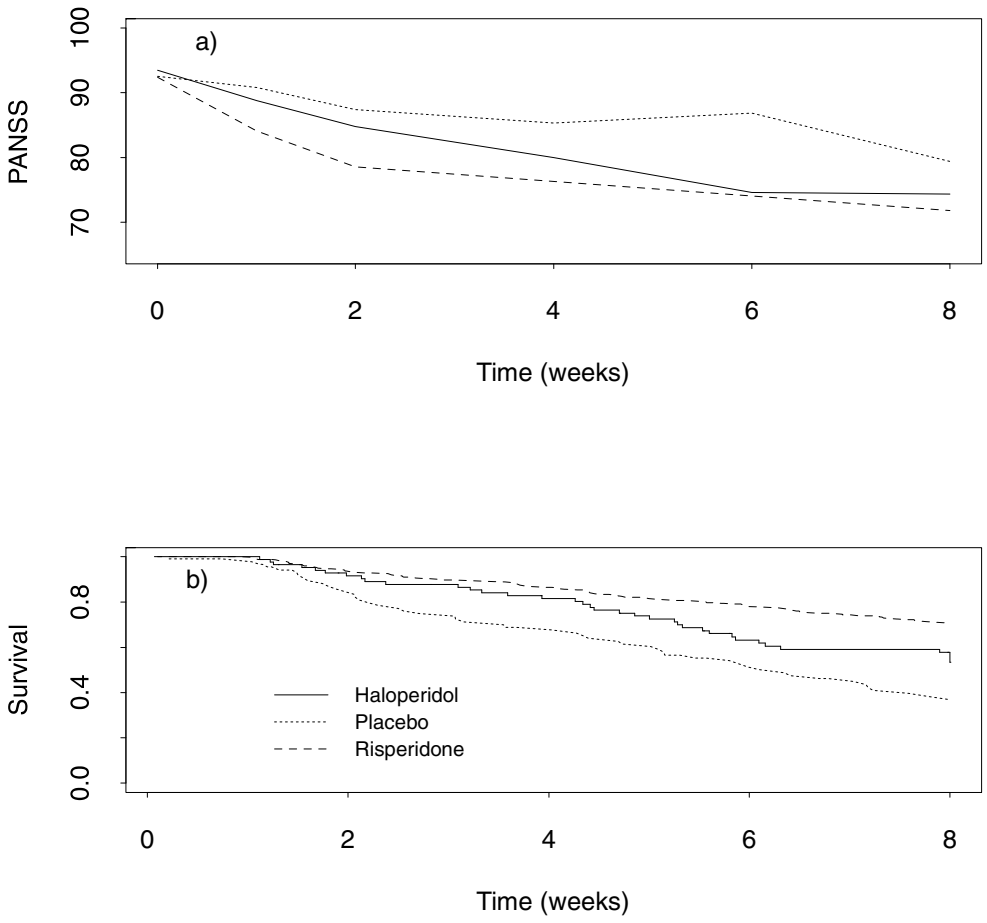

Fig. 1. Longitudinal and event time summaries for schizophrenia trial: (a) mean PANSS score and (b) survival curves for time to drop-out due to inadequate response.

statistical and scientific objectives of investigations of this kind will depend upon the the application of interest. In particular, the primary focus for inference may be on:

(a) adjustment of inferences about longitudinal measurements to allow for possibly outcome-dependent drop-out;
(b) the distribution of time to a terminating or recurrent event conditional on intermediate longitudinal measurements; or
(c) the joint evolution of the measurement and event time processes.

In analysing the PANSS data of Figure 1, for instance, Diggle (1998) had emphasis (a), with interest mainly in the effect of treatment on the longitudinal PANSS score. Objective (b) is more appropriate for survival analysis with a time-dependent covariate measured with error. In this work we will consider (c) to

be important, with equal interest in both longitudinal and event time components. However, the methods and models apply equally to (a) or (b).

Our goal is to develop a flexible methodology for handling combined longitudinal and event history data, incorporating the most commonly used first-choice assumptions from both subject areas. Thus, for the longitudinal response we argue that any general model should be able to incorporate fixed effects, random effects, serial correlation and pure measurement error, whilst the event history or survival analysis should be based on a semi-parametric proportional hazards (or intensity) model with or without frailty terms. Another desirable feature for joint modelling is that in the absence of association between the longitudinal data and event times the joint analysis should recover the same results as would be obtained from separate analyses for each component.

A central feature of our modelling strategy is to postulate a latent bivariate Gaussian process $W(t) = \{W_1(t), W_2(t)\}$, and assume that the measurement and event processes are conditionally independent given $W(t)$ and covariates. Hence, association is described through the cross-correlation between $W_1(t)$ and $W_2(t)$. In Section 2 we describe the model and derive the likelihood for the combined data through a factorization akin to a random effects selection model. In Section 3 we assume that $W_1(t)$ and $W_2(t)$ can be specified through a linear random effects model, $W_k(t) = d_k(t)'U : k = 1, 2$, where U is multivariate Gaussian and the $d_k(t)$ are vectors of possibly time-varying explanatory variables. This model includes as a special case a model which has been considered previously by several authors, under which a Laird–Ware random effects model for $W_1(t)$ is combined with a proportionality assumption $W_2(t) \propto W_1(t)$. The extension allows us to consider situations in which the association between measurement and event processes is described in terms of particular components of variation, such as the intercept and/or slope of a subject-specific time trend, rather than the value of $W_1(t)$ alone. Additionally, random effects (or frailties) can be allowed to influence the event process independently of the measurement process. An exploitation maximization (EM) algorithm proposed by Wulfsohn and Tsiatis (1997) can be used for estimation and is illustrated in simulations. In Section 4 we consider an alternative model under which $W_1(t)$ and $W_2(t)$ include serially correlated stochastic components and describe an estimation procedure based on the likelihood factorization given in Section 2. In Section 5 we present the results of fitting a variety of models to the schizophrenia data introduced earlier in this section. Section 6 discusses possible extensions to the work.

2. The general model and its associated likelihood

2.1 *Notation*

Some of our notation is non-standard as there is occasional conflict between that commonly used in longitudinal data analysis and that which is familiar in event history analysis.

We consider a set of m subjects followed over an interval $[0, \tau)$. The ith subject provides a set of quantitative measurements $\{y_{ij} : j = 1, \ldots, n_i\}$ at times $\{t_{ij} : j = 1, \ldots, n_i\}$, together with realizations of a counting process $\{N_i(u) : 0 \leqslant u \leqslant \tau\}$ for the events and a predictable zero-one process $\{H_i(u) : 0 \leqslant u \leqslant \tau\}$ indicating whether the subject is at risk of experiencing an event at time u (usually denoted $Y_i(u)$ in survival analysis). The counting process $N_i(u)$ has jumps at times $\{u_{ij} : j = 1, \ldots, N_i(\tau)\}$, with $N_i(\tau)$ not more than one for survival data. We assume that the timing of the measurements $\{t_{ij}\}$ is non-informative in the sense that the decision to schedule a measurement is made independently of the measurement or counting processes. We also assume that the censoring of event times which might have occurred after the end of the study is non-informative, and that data from different subjects are generated by independent processes.

2.2 *Model formulation*

We propose to model the joint distribution of measurements and events for the ith subject via an unobserved or *latent* zero-mean bivariate Gaussian process, $W_i(t) = \{W_{1i}(t), W_{2i}(t)\}$, which is realized independently in different subjects. In particular, we assume that this latent process drives a pair of linked sub-models, which we call the *measurement* and *intensity* sub-models as follows.

(1) The sequence of measurements $y_{i1}, y_{i2}, \ldots$ at times $t_{i1}, t_{i2}, \ldots$ is determined by

$$Y_{ij} = \mu_i(t_{ij}) + W_{1i}(t_{ij}) + Z_{ij}, \tag{1}$$

where $\mu_i(t_{ij})$ is the mean response and $Z_{ij} \sim N(0, \sigma_z^2)$ is a sequence of mutually independent measurement errors. We assume that $\mu_i(t)$ can be described by a linear model

$$\mu_i(t) = x_{1i}(t)'\beta_1$$

in which the vectors $x_{1i}(t)$ and β_1 represent possibly time-varying explanatory variables and their corresponding regression coefficients, respectively.

(2) The event intensity process at time t is given by the semi-parametric multiplicative model

$$\lambda_i(t) = H_i(t)\alpha_0(t)\exp\{x_{2i}(t)'\beta_2 + W_{2i}(t)\}, \tag{2}$$

with the form of $\alpha_0(t)$ left unspecified. Vectors $x_{2i}(t)$ and β_2 may or may not have elements in common with $x_{1i}(t)$ and β_1.

We adopt the generic notation Y for the combined longitudinal data, N (from counting process terminology) for the combined event time data, X for covariate information, and W_1 and W_2 for the latent processes. Note that Y and N are conditionally independent, given X, W_1 and W_2. Without the conditioning, dependence between Y and N can arise in two ways: through the deterministic effects of common covariates X or through stochastic dependence between W_1 and W_2. We refer to the direct link between W_1 and W_2 as *latent association*. In the absence of latent association there is nothing to be gained by a joint analysis, unless the measurement and intensity sub-models have parameters in common.

The combination of equations (1) and (2) embraces a wide range of specific models which have been proposed for the separate analysis of continuous longitudinal measurements and survival outcomes. In particular, a flexible choice for $W_{1i}(t)$, combining suggestions in Laird and Ware (1982) and in Diggle (1988), would be

$$W_{1i}(t) = d_{1i}(t)'U_{1i} + V_{1i}(t). \tag{3}$$

In equation (3), $d_{1i}(t)$ is a vector of explanatory variables, $U_{1i} \sim MVN(0, \Sigma_1)$ is a corresponding vector of random effects and $V_{1i}(t)$ is a stationary Gaussian process with mean zero, variance σ_{v1}^2 and correlation function $r_1(u) = \text{cov}\{V_{1i}(t), V_{1i}(t-u)\}/\sigma_1^2$. Note that by assuming stationarity we are excluding the integrated Ornstein–Uhlenbeck process employed by Taylor *et al.* (1994) and Lavalley and Degrutolla (1996), although extension to this type of process is in principle straightforward. The process $W_{2i}(t)$ is specified in a similar way to $W_{1i}(t)$ and, in the absence of latent association, represents possibly time-dependent (log) Gaussian frailty (e.g. Yau and McGilchrist, 1998).

2.3 *The likelihood function*

The marginal distribution of the observed measurements Y is easily obtained. It is convenient to factorize the likelihood for the observed data as the product of this marginal distribution and the

conditional distribution of the events, N, given the observed values of Y. A complete factorization of the likelihood would include a third term, the conditional distribution of the number and timings of observed measurements given Y and N. However, our assumption that the measurement schedule is non-informative implies that this third term can be ignored.

Let θ denote the combined vector of unknown parameters. We define $\mathcal{W}_{2i}$ to be the complete path of W_{2i} over the interval $[0, \tau)$ and let $\mathcal{W}_2$ be the collation of these paths over all subjects. Conditional on $\mathcal{W}_2$, the event history data are independent of the measurements Y and we can write the likelihood $L = L(\theta, Y, N)$ as

$$L = L_Y \times L_{N|Y} = L_Y(\theta, Y) \times E_{\mathcal{W}_2|Y}\{L_{N|W_2}(\theta, N \mid \mathcal{W}_2)\}, \tag{4}$$

in which $L_Y(\theta, Y)$ is of standard form corresponding to the marginal multivariate normal distribution of Y. The conditional likelihood for the event data, $L_{N|W_2}(\theta, N \mid \mathcal{W}_2)$, captures any likelihood contribution arising from the achieved numbers of longitudinal measurements before any drop-out or failure. With $A_0(t) = \int_0^t \alpha_0(u)du$ denoting the cumulative baseline intensity, we can write $L_{N|W_2}$ (Andersen *et al.*, 1993, p. 482) as

$$L_{N|W_2}(\theta, N \mid \mathcal{W}_2) = \prod_i \left\{ \left(\prod_t [\exp\{x_{2i}(t)'\beta_2 + W_{2i}(t)\}\alpha_0(t)]^{\Delta N_i(t)} \right) \times \exp\left[- \int_0^\tau H_i(t) \exp\{x_{2i}(t)'\beta_2 + W_{2i}(t)\} dA_0(t) \right] \right\}. \tag{5}$$

Determination of L apparently requires an expectation with respect to the distribution of the infinite dimensional process $\mathcal{W}_2$, given the longitudinal measurements Y. However, this is avoided under the semi-parametric approach because the non-parametric estimator of baseline intensity is zero except at observed event times $\mathcal{U} = \{u_{ij}\}$. Hence, the expectation need only be taken over a finite number of variables $W_2 = \{\mathcal{W}_2(u) : u \in \mathcal{U}\}$. Note that the joint distribution of Y and W_2 is multivariate normal with easily derived covariance structure (e.g. Mardia *et al.*, 1979).

3. Linear random effects models and EM estimation

Tsiatis *et al.* (1995), Faucett and Thomas (1996), and Wulfsohn and Tsiatis (1997) all assume a Laird and Ware (1982) linear random effects model $W_{1i}(t) = U_{1i} + U_{2i}t$, in conjunction with a proportionality assumption $W_{2i}(t) = \gamma W_{1i}(t)$, where (U_{1i}, U_{2i}) are subject-specific bivariate Gaussian random effects. A natural extension allows the random slope and intercept to have different effects on the event process, and to add a time-constant frailty term. Thus, dropping the subscript i we again assume

$$W_1(t) = U_1 + U_2 t,$$

where (U_1, U_2) are zero-mean bivariate Gaussian variables with respective variances σ_1^2 and σ_2^2, and correlation coefficient ρ. However, for $W_2(t)$ we now propose a specification

$$W_2(t) = \gamma_1 U_1 + \gamma_2 U_2 + \gamma_3(U_1 + U_2 t) + U_3, \tag{6}$$

where $U_3 \sim N(0, \sigma_3^2)$ is independent of (U_1, U_2). In this model the parameters γ_1, γ_2 and γ_3 measure the association induced through the intercept, slope and current W_1 value, respectively, whilst U_3 models frailty orthogonal to the measurement process.

The EM estimation algorithm described by Wulfsohn and Tsiatis (1997) can easily be extended to this model and is our preferred method of estimation. The procedure involves iterating between the following two steps until convergence is achieved.

(1) *E-step*. Consider the random effects $U = (U_1, U_2, U_3)'$ as missing data. We determine the expected values conditional on the observed data (Y, N) of all functions $h(U)$ appearing in the logarithm of the complete data likelihood $L(\theta, Y, N, U)$, using current parameter estimates. Wulfsohn and Tsiatis show that the conditional expectations can be written as

$$E[h(U)|Y, N] = \left\{ \int h(U) f(N|U) f(U|Y) dU \right\} / f(N|Y) \tag{7}$$

with

$$f(N|Y) = \int f(N|U) f(U|Y) dU. \tag{8}$$

Here $f(N|U)$ is the contribution (of the ith individual) to the event-time component (5) of the complete-data likelihood and $f(U|Y)$ is the Gaussian conditional distribution of the random effects given the longitudinal data. Since U is low dimensional the integrals can be approximated using Gauss–Hermite quadrature, which can also be used to evaluate the final log-likelihood, $\log(L(\theta, Y, N)) = \log(E_U[L(\theta, Y, N, U)|Y, N])$.

(2) *M-step*. We maximize the complete data log-likelihood with each function $h(U)$ replaced by its corresponding expectation. A one-step approximation based on equation (5) can be used for the event time regression parameters β_2 and γ given the current $A_0(t)$. In turn, $A_0(t)$ can be estimated using the usual Breslow estimator given β_2, γ and the appropriate expected functions of U (Andersen *et al.*, 1993, p. 483). Closed form estimators are available for the measurement parameters (β_1, σ_z^2) and random effects parameters $(\sigma_1^2, \sigma_2^2, \rho, \sigma_3^2)$, again given the appropriate expectations of functions of U. Full details are provided by Wulfsohn and Tsiatis (1997).

In Table 1 we summarize the results of a simulation study into the performance of the estimation procedure, illustrating in particular the effects of ignoring association between Y and N. We considered two sample sizes, $m = 250$ and $m = 500$, and for each we simulated data under two scenarios: without latent association (all $\gamma = 0$, upper part) and with latent association (not all $\gamma = 0$, lower part). When latent association was present both the intercept and current W_1 value were taken to affect event times ($\gamma_1\gamma_3 \neq 0$) but not the slope alone ($\gamma_2 = 0$). The longitudinal model was taken to be

$$Y_t = \beta_{11} + \beta_{12}t + \beta_{13}X + U_1 + U_t \times t + Z_t$$

with $X \sim N(0, 1)$ and $n = 6$ measurements scheduled at times 0, 0.5, 1, 1.5, 2 and 3 units. Event times were generated from the model

$$\alpha(t) = \alpha_0(t) \exp\{\beta_{21}X + \gamma_1 U_1 + \gamma_2 U_2 + \gamma_3(U_1 + U_t \times t) + U_3\}$$

with Weibull baseline $\alpha_0(t)$ chosen to give about 40% drop-out by time 1, and 70% drop-out by time 3. No parametric assumptions on baseline hazard were made during estimation. A final truncation time of $\tau = 4$ was used.

We analysed each simulated data set in two ways: firstly with separate analyses of measurements Y and event times N, ignoring any latent association, and secondly using the joint approach and an EM implementation of maximum likelihood estimation as outlined above. In roughly 5% of simulations the algorithm failed to converge, usually as a result of difficulty in estimating one or more of the variance parameters for U. Results based in each case on 100 successful completions of the algorithm are given in Table 1. They can be summarized as follows.

Table 1. *Simulation results*

(a) No latent association

Parameter	True	$m = 250$				$m = 500$			
		Separate		Joint		Separate		Joint	
		Mean	SE	Mean	SE	Mean	SE	Mean	SE
β_{11}	0	−0.002	0.005	0.000	0.005	0.004	0.004	0.004	0.004
β_{12}	1	1.009	0.008	1.016	0.009	0.994	0.006	0.991	0.006
β_{13}	1	0.998	0.006	0.997	0.006	0.996	0.004	0.996	0.004
σ_1^2	0.5	0.509	0.007	0.511	0.007	0.494	0.004	0.494	0.004
σ_2^2	1	1.008	0.013	1.011	0.014	1.004	0.008	1.006	0.008
ρ	0	0.003	0.009	−0.001	0.010	−0.014	0.007	−0.014	0.007
σ_z^2	0.25	0.249	0.002	0.249	0.002	0.252	0.001	0.252	0.001
β_{21}	1	0.966	0.010	0.976	0.010	0.968	0.008	0.986	0.011
σ_3^2	0.25	0.140	0.011	0.138	0.011	0.190	0.014	0.248	0.038
γ_1	0			0.000	0.017			0.003	0.011
γ_2	0			0.002	0.018			−0.012	0.014
γ_3	0			0.001	0.010			0.002	0.007

(b) Latent association

Parameter	True	$m = 250$				$m = 500$			
		Separate		Joint		Separate		Joint	
		Mean	SE	Mean	SE	Mean	SE	Mean	SE
β_{11}	0	0.043	0.006	0.004	0.006	0.037	0.004	−0.002	0.004
β_{12}	1	0.620	0.008	0.945	0.010	0.640	0.005	0.973	0.007
β_{13}	1	0.979	0.005	1.000	0.005	0.978	0.004	0.999	0.004
σ_1^2	0.5	0.496	0.006	0.502	0.006	0.492	0.004	0.501	0.004
σ_2^2	1	0.736	0.012	0.989	0.015	0.733	0.009	0.990	0.011
ρ	0	−0.023	0.011	−0.015	0.010	−0.011	0.007	−0.008	0.007
σ_z^2	0.25	0.259	0.002	0.250	0.002	0.260	0.001	0.250	0.001
β_{21}	1	0.789	0.013	0.969	0.014	0.843	0.012	0.982	0.011
σ_3^2	0.25	1.414	0.051	0.245	0.018	1.723	0.041	0.283	0.025
γ_1	−1.5			−1.505	0.039			−1.520	0.033
γ_2	0			−0.088	0.033			−0.103	0.026
γ_3	2			1.986	0.035			2.019	0.031

- Results for separate and joint analyses are all good if there is no latent association, except for severe underestimation of frailty variance σ_3^2, especially at the lower sample size. This finding is consistent with the results of Neilsen *et al.* (1992), who found similar downward bias in estimated frailty variances in gamma frailty models.
- Severe bias occurs for some parameters when there is ignored latent association, notably underestimation of the time trend (β_{12}) because with positive γ there is greater drop-out of high responders and the observed time trend is attenuated. Similarly, the estimate for the random slope variance σ_2^2 is negatively biased.

- In contrast, the frailty variance σ_3^2 is markedly overestimated. This is as expected, since the frailty effect is confounded with the ignored random effects.
- Overall, results under the joint modelling approach are good, although we note a small (but significant) bias in $\hat{\gamma}_2$, whose true value is zero. This indicates some difficulty in disentangling the separate effects.
- In the presence of latent association, the additional information available from the longitudinal data operates, in effect, as extra covariate information which helps overcome problems in frailty variance estimation. This may be because frailty is unidentifiable in semi-parametric models without covariates, with identification becoming easier as the number of important covariates increases.

4. STOCHASTIC PROCESS MODELS AND AN ALTERNATIVE ESTIMATION PROCEDURE

A natural extension to the Laird–Ware model for $W_1(t)$ adds the possibility of a serially correlated component, as in equation (3). Thus we now consider

$$W_1(t) = U_1 + U_2 t + V(t)$$

where U_1 and U_2 are as previously described and $V(t)$ is a stationary Gaussian process, independent of $(U_1, U2)$, with $V(t) \sim N(0, \sigma_v^2)$ and $\mathrm{Corr}(V(t), V(t-u)) = r(u; \phi)$. Typically we might take

$$r(u; \phi) = \exp(-|u|^{\nu}/\phi)$$

for fixed ν, say 0.5, 1 or 2, and ϕ to be estimated. The process $V(t)$ thus describes local deviations whereas the term $U_2 t$ represents a sustained trend. The second process $W_2(t)$ may be assumed to be proportional to $W_1(t)$ or may again be broken into components and augmented by frailty $U_3 \sim N(0, \sigma_3^2)$, as in the previous section.

For each subject, let $U = (U_1, U_2, U_3, V_1, V_2, \ldots)'$ be the unobserved random effects, where $V_1, V_2, \ldots$ are the values of $V(t)$ at the distinct event and measurement times required to form the contribution from this subject to the complete data likelihood $L(\theta, Y, N, U)$. Depending on the models for W_1 and W_2, only a subset of elements of U may be required. In particular, models with a term $V(t)$ replacing the random linear trend $U_2 t$ are useful for data with long sequences of measurements on each subject.

Since U and Y are jointly Gaussian, the conditional distribution of U given Y can be derived easily, and thus in principle the Wulfsohn and Tsiatis EM algorithm of the previous section can be applied. However, the extension to incorporate the serially correlated process leads to practical difficulties in implementation and we suggest an alternative estimation procedure, which involves two modifications of the Wulfsohn and Tsiatis technique.

The first modification is in the numerical approximation to the integrals required at the E-step. Once $V(t)$ is incorporated, the dimensionality of U can be quite large, making Gauss–Hermite quadrature prohibitively time consuming. For instance, under a simple model with $W_1(t) = V(t)$ and $W_2(t) = \gamma W_1(t)$, one of the calculations required for each subject is the expectation of

$$\exp\left[-\int_0^{\tau} H(t) \exp\{\gamma V(t)\} dA_0(t)\right].$$

If the subject is at risk through d distinct failure times this expectation is with respect to a d-dimensional vector (of $V(t)$ at each failure time, conditional on Y). Evaluation of p^d terms in p-point quadrature is then necessary, which is prohibitive for realistic d. Instead, we suggest a Monte Carlo integration method with antithetic simulation for variance reduction. The steps involved for any required function $E[h(U)|Y, N]$ are as follows.

(1) Writing $U|Y \sim N(\mu_y, \Sigma_y)$, find the Choleski decomposition C_y of Σ, so $C_y C_y' = \Sigma_y$.
(2) Generate a number M of antithetic pairs

$$U_+ = \mu_y + C_y Z \qquad U_- = \mu_y - C_y Z$$

where $Z \sim N(0, I)$ is of the same dimension as U.
(3) Estimate the integrals in equations (7) and (8) by the simulation sample means of $h(U) f(N|U)$ and $f(N|U)$ over the $2M$ realizations of U.

The value of M can be tuned to balance numerical accuracy with computational speed. Negative correlation between U_+ and U_- leads to a smaller variance in the sample mean values of $h(U) f(N|U)$ and $f(N|U)$ than would be obtained from $2M$ independent simulations of U.

The second modification to the procedure exploits the likelihood factorization (4), and essentially treats only W_2 as missing data rather than both W_1 and W_2. This reduces sensitivity to error in numerical integration and overcomes some of the convergence difficulties arising from attempting to estimate variance properties of W_1 from missing data. The procedure combines estimation of variance parameters by a search technique based on evaluating equation (4), and estimation of the remaining regression and hazard parameters by a conditional EM procedure. The variance parameters are a subset, ξ say, of $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_v^2, \phi, \sigma_z^2)'$, depending on the model in use. The procedure is then as follows.

(1) Initialize ξ by separate analyses of Y and N using standard methods.
(2) For given ξ use an EM algorithm as in Section 3 to estimate $(\beta_1, \beta_2, \{A_0(t)\}, \gamma)$, using either antithetic Monte Carlo simulation or quadrature for numerical integration, as required.
(3) Evaluate the logarithm of $L = L_Y \times L_{N|Y}$ (4), using the marginal distribution of Y and noting that each subject's contribution to $L_{N|Y}$ is given by equation (8) and is evaluated as part of the EM procedure at Step 2.
(4) Update ξ and return to Step 2, using a simplex algorithm (Nelder and Mead, 1965) to search for the value of ξ which maximizes equation (4).

In our experience, the inner EM algorithm is quick and accurate with only a relatively low value of M required at the Monte Carlo E-step. A higher value is needed for accurate evaluation of the likelihood contribution once the inner algorithm has converged. This leads to the outer simplex search being relatively slow, though some gain in speed can be achieved by using a fairly low value of M initially, while the estimates are some distance from the maximum, with a higher value for a more refined search at the later stages. Time required for the Cholesky decomposition and simulations at the E-step can be reduced if required by discretizing the timescale, to reduce the number of distinct event times and hence the dimension of U.

5. Example: a clinical trial with informative drop-out

We now return to the illustrative example introduced briefly in Section 1. Data are available from 523 patients, randomly allocated amongst the following six treatments: placebo, haloperidol 20 mg and risperidone at dose levels 2 mg, 6 mg, 10 mg and 16 mg. Haloperidol is regarded as a standard therapy. Risperidone is described as 'a novel chemical compound with useful pharmacological characteristics, as has been demonstrated in *in vitro* and *in vivo* experiments'. The primary response variable was the total score obtained on the PANSS, a measure of psychiatric disorder. In an earlier analysis of these data, Diggle (1998) combined the four risperidone groups into one, and for comparability we do the same. The resulting numbers of patients randomized to placebo, haloperidol and risperidone treatments were 88, 87 and 348.

The study design specified that the PANSS score should be taken at weeks −1, 0, 1, 2, 4, 6 and 8, where −1 refers to selection into the trial and 0 to baseline. The week between selection and baseline was used to establish a stable regime of medication for each patient, and in the analysis of the data we shall exclude the week −1 measurements. Of the 523 patients, 270 were identified as drop-outs and 183 of these gave as the reason for drop-out 'inadequate response'. In our analysis we shall treat drop-out due to inadequate response as a potentially informative event, and drop-out for other reasons as a censored follow-up time. Exact drop-out time was not recorded during the trial, the only information being on the first missed observation time. For this analysis we imputed each drop-out time from a uniform distribution over the appropriate interval between last observed and first missed measurement times. Results were not sensitive to imputation.

Figure 1(a) shows the observed mean response as a function of time within each treatment group, i.e. each average is over those patients who have not yet dropped out. All three groups had a decreasing mean response, perhaps at a slower rate towards the end of the study. The overall reduction in mean response within each active treatment group is very roughly from between 90 and 95 to around 75. This appears close to the criterion for clinical improvement, which was stated, in advance of the trial, to be 'a reduction of 20% in the mean PANSS scores'. The decrease in the placebo group was smaller overall. However, at each time-point these observed means are, necessarily, calculated only from those subjects who have not yet dropped out of the study. Figure 1(b) shows estimated survival curves for time to drop-out due to inadequate response: large differences between groups are evident, with the highest drop-out rate in the placebo group and the lowest in the risperidone group.

In the analyses reported here we assume a saturated model for the mean PANSS response, with a distinct element of β_1 for each treatment and measurement time combination. We have also analysed the data assuming a quadratic time trend for each treatment, which gave very similar results and is not reported here. For the drop-out model we assume time constant treatment effects, with the standard treatment haloperidol group as baseline and thus one entry in β_2 for each of the placebo and risperidone groups. We used the method of Section 4 for estimation for all models, with an initial analysis using $M = 500$ simulations for all expectations, then once the approximate estimates were obtained a more refined analysis with $M = 1000$ simulations at the E-step within the inner EM algorithm but $M = 2000$ simulations for likelihood evaluation. This led to a standard deviation of around 0.02 for Monte Carlo error in the log expected likelihoods.

Table 2 summarizes the results of fitting a variety of models for the latent processes W_1 and W_2. We begin (Model I) with a simple random intercept model for W_1, no random effects allowed in the drop-out model and hence no association. There is no improvement in fit (measured by the likelihood value) when frailty orthogonal to W_1 is allowed (not shown) with the estimated frailty variance lying on the boundary of zero. However, once latent association is allowed there is a substantial improvement in combined likelihood (Model II). The frailty result here is worth a comment: the initial analysis without latent association indicated no frailty in the drop-out component, which might be taken to suggest that there are no unmeasured covariates which could influence drop-out. Yet from Model II there is a clear association with the PANSS score. We suspect this result is caused by the known difficulty of estimating frailty effects with a semi-parametric baseline when there are few covariates.

Diggle (1998) combined a random intercept with a stationary Gaussian process $V(t)$ in his model for PANSS scores, assuming correlation function

$$\mathrm{Corr}(V(t), V(t-u)) = \exp(-|u|^2/\phi).$$

In incorporating this component we discretized drop-out times to integer values for the initial estimation, but used the original imputed times for final estimation and likelihood evaluation. Inclusion of $V(t)$ leads to substantial increases in marginal likelihood for the longitudinal component (Model III) and combined likelihood when the standard association model $W_2(t) \propto W_1(t)$ is assumed (Model IV). A further large

Table 2. *Log maximized likelihoods for schizophrenia data*

	$W_1(t)$	$W_2(t)$	$\log L_Y$	$\log L_{N\|Y}$	$\log L$
Intercept only					
I	U_1	0	−10251.85	−1228.55	−11480.40
II	U_1	$\gamma W_1(t)$	−10252.58	−1181.13	−11433.72
Intercept +SGP					
III	$U_1 + V(t)$	0	−10126.66	−1228.55	−11355.21
IV	$U_1 + V(t)$	$\gamma W_1(t)$	−10132.55	−1146.14	−11278.69
V	$U_1 + V(t)$	$\gamma_1 U_1 + \gamma_2 V(t)$	−10139.79	−1107.61	−11247.40
Intercept +slope					
VI	$U_1 + U_t t$	0	−10127.31	−1228.55	−11355.86
VII	$U_1 + U_t t$	$\gamma W_1(t)$	−10133.76	−1137.41	−11271.17
VIII	$U_1 + U_t t$	$\gamma W_1(t) + U_3$	−10135.99	−1132.90	−11268.88
IX	$U_1 + U_t t$	$\gamma_1 U_1 + \gamma_2 U_2 + \gamma_3 W_1(t)$	−10147.75	−1096.05	−11243.80
X	$U_1 + U_t t$	$\gamma_2 U_2 + \gamma_3 W_1(t)$	−10148.42	−1095.60	−11244.03

increase in overall likelihood occurs when the two components of $W_1(t) = U_1 + V(t)$ are allowed to have separate effects on the drop-out process (Model V). Note that there is considerable sacrifice in the marginal likelihood L_Y for the measurements in order to improve the drop-out and overall components $L_{N|Y}$ and L.

A standard Laird–Ware random slope and intercept model apparently fits the marginal PANSS distribution almost as well as the Diggle intercept and Gaussian process model (Model VI), with once more strong evidence of association between the longitudinal and drop-out components (Model VII). Inclusion of a frailty term U_3 leads to improved likelihood (Model VIII). Of the models considered this is the only instance where we found frailty to have a non-negligible effect. When the model is extended to a full linear random effects model for $W_2(t)$ (Model IX) there is another substantial increase in likelihood, with a large drop in L_Y more than compensated by an increase in the second likelihood component $L_{N|Y}$. There is almost no loss in likelihood in removing from this model the separate effect on drop-out of the random intercept term U_1 (Model X). On the basis of this likelihood analysis, this is our final model for these data.

Thus, under Model X drop-out appears to be affected separately by two latent factors: the current value of $W_1(t)$ and the steepness of the trajectory. Since high PANSS indicates poor condition, both of these conclusions are clinically reasonable: patients with either poor or rapidly declining mental health have increased risk of drop-out due to inadequate response. Parameter values for this model are given in Table 3 together with the corresponding estimates when no association between PANSS and drop-out is assumed, the two components being analysed separately. Note the reduction in the random slope variance σ_2^2 under separate analyses, and the attenuation of estimated treatment effects on the drop-out process, both consistent with the simulation results in Table 1. Standard errors in Table 3 were obtained by a Monte Carlo method, refitting Model X to 100 simulated data sets generated using parameter values taken from the original analysis. In order to complete the re-estimations within a reasonable time we used the smaller value of $M = 500$ in the Monte Carlo likelihood evaluation and stopped each inner EM algorithm as soon as an iteration caused a decrease in estimated likelihood. We accepted the value of $\xi = (\sigma_1^2, \sigma_2^2, \rho)$

Table 3. *Parameter estimates and standard errors for final joint model with and without latent association*

	Var(U_1)	Var(U_2)	Corr(U_1, U_2)	Var(Z)	Placebo	Resp.	U_2	$W_1(t)$
	σ_1^2	σ_2^2	ρ	σ_z^2	β_{21}	β_{21}	γ_2	γ_3
Joint	283.37	12.59	0.06	100.24	0.779	−0.884	0.349	0.042
(X)	(18.17)	(1.31)	(0.02)	(3.95)	(0.344)	(0.322)	(0.051)	(0.007)
Separate	275.92	7.12	0.01	106.10	0.480	−0.508	0	0
(VI)	(21.42)	(0.81)	(0.08)	(4.18)	(0.218)	(0.196)	–	–

which gave the maximum estimated likelihood after 30 iterations of the outer simplex. Our experience is that the estimated ξ is quickly very stable but Monte Carlo noise in the likelihood function evaluation prevents the simplex procedure from indicating convergence when M is small.

Estimated values of β_1 under Model X are shown in Figure 2(a), with $\pm$ two standard errors. These estimate the hypothetical drop-out-free population PANSS profiles, and are higher than the observed profiles (also shown) as a result of the tendency for patients with high scores to drop out due to inadequate response. Mean values for haloperidol-treated patients are lower (and thus better) than for the placebo group, as expected, but higher than the mean values for the resperidone group. A similar though less pronounced pattern is seen if the same model is assumed for the measurements but there is no latent association with drop-out (Model VI, Figure 2(b)). The general estimated drop-out-free pattern in Figures 2(a) and 2(b), of a late increase in PANSS after an inital fall, occurs to some extent for all the Laird–Ware intercept plus slope models we considered (Models VI–X) but for none of the other models. To illustrate, Figures 2(c) and 2(d) show estimated values of β_1 under the apparently best fitting intercept-only and intercept plus SGP models (Models II and V).

We investigated model adequacy by comparing data simulated under the final Model X with that observed in the trial. Figure 3 illustrates some of our findings, based on 100 simulations of samples of 88, 87 and 348 subjects in the three treatment groups, as in the trial itself. The upper plot in Figure 3 compares the observed mean PANSS scores amongst patients still involved in the study with the corresponding simulation means, and the lower plot compares observed and simulated survival curves. The simulated values are close to those observed in all cases. A similar plot (not shown) based on the intercept-only model with latent association (Model II) also shows very good agreement, and there is reasonable agreement also (but larger standard errors) for the best intercept plus SGP model (V), except for some overestimation of drop-out rate for the haloperidol group (again not shown). Other plots based only on PANSS profiles for subjects who complete the trial also show good agreement between observed and simulated data for all models.

6. DISCUSSION

We have presented a general approach to the joint analysis of longitudinal and event history data, building on previously described procedures, and assuming 'standard' models for the two components separately. The modelling strategy is based on a specification of two linked Gaussian random processes, $W_1(t)$ and $W_2(t)$, with correlation between $W_1(t)$ and $W_2(t)$ inducing stochastic dependence between the measurement and event processes. Depending on the precise specification of the W_j, there can be difficulties relating to identifiability and sensitivity.

As given in Section 2, there are no particular restrictions on the latent Gaussian processes $W_1(t)$ and $W_2(t)$ or their cross-correlation. Clearly, without further assumptions identifiability problems may occur

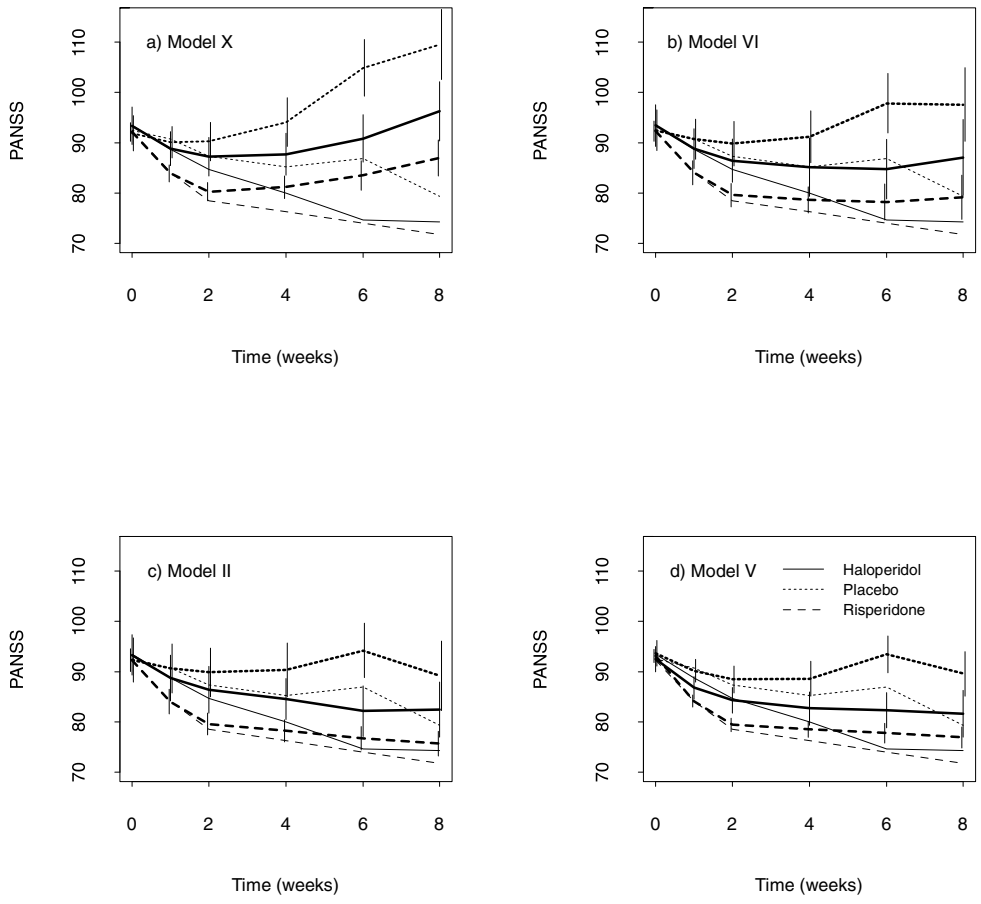

Fig. 2. Observed PANSS scores (fine lines) and hypothetical drop-out-free estimates (thick lines) under various models, with ± two standard errors. Refer to Table 2 and text for model descriptions.

under the semi-parametric intensity model for event times. In Sections 3–5 these were avoided by directly linking $W_2(t)$ to components of $W_1(t)$, apart from the possibility of a time-constant frailty term. More generally, any bivariate, zero-mean Gaussian process $W(t) = \{W_1(t), W_2(t)\}$ has a unique representation of the form

$$W_2(t) = \gamma(t) W_1(t) + W_{2|1}(t)$$

in which $W_1(t)$ and $W_{2|1}(t)$ are orthogonal, zero-mean Gaussian processes. This formulation conveniently separates two different kinds of extension to the special case of proportional latent processes: replacement of the parameter γ by a function of time, $\gamma(t)$; and introduction of a time-dependent frailty component which is independent of the measurement process. There are unresolved identifiability questions relating to both types of extension.

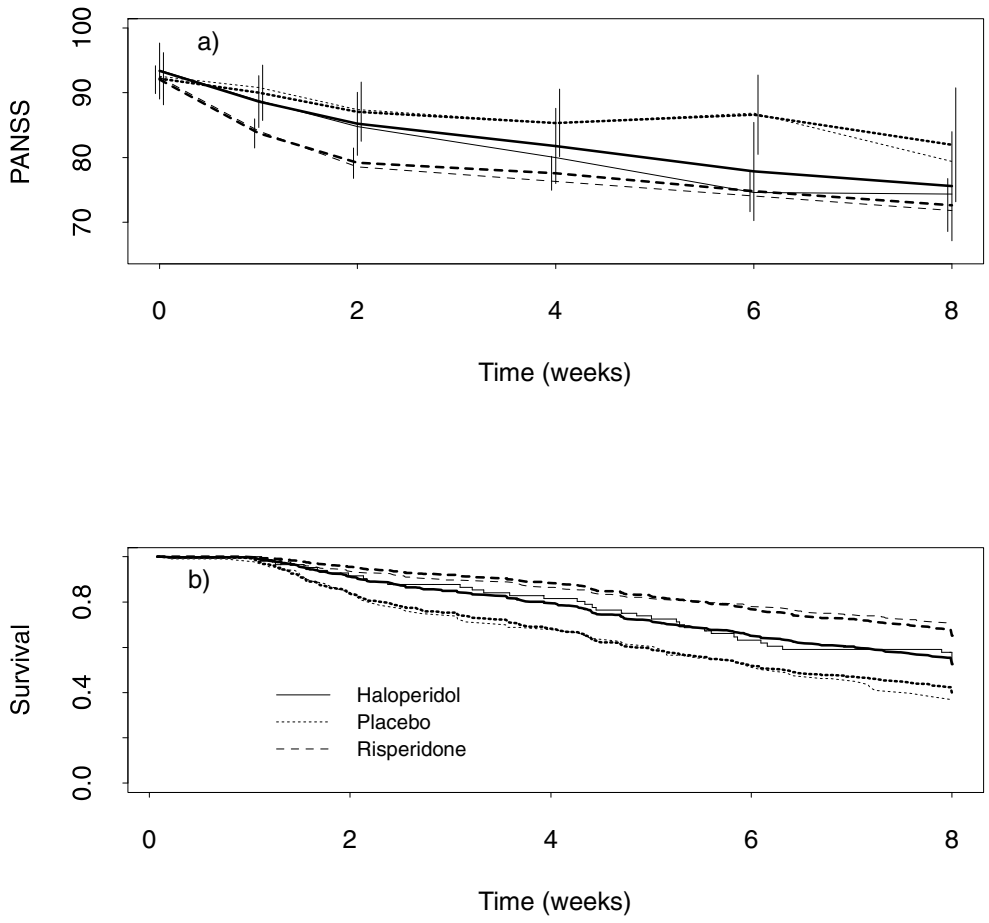

Fig. 3. Assessment of model adequacy: (a) observed PANSS scores (fine lines) and estimated values conditional on no drop-out (thick lines), under Model X, with ± two standard errors, and (b) observed and estimated survival curves for time to drop-out.

Firstly, with continuous failure time data some form of parametric assumption or smoothness criterion will almost certainly be required for $\gamma(t)$, which relates to possible non-constancy in time of the association between the measurement and event processes. Secondly, time-dependent frailty $W_{2|1}(t)$ orthogonal to the measurement process is not identifiable from single-event survival data with time constant covariates, and it is not clear what assumptions are required on covariate or event processes to ensure identifiability even under the restriction $\gamma(t) \equiv 0$. Thus, in analysing data with survival outcomes we recommend the assumption that $W_{2|1}(t) = U_3$, a time-constant univariate (log) Gaussian frailty term as in the models of Sections 3–5, until identifiability issues are resolved.

Turning to sensitivity, as our capability to fit ever more complex models increases, the questions of what, exactly, the data can tell us, and which terms are necessary in practice also become important. For

the schizophrenia example of Section 5 we found that several models, sometimes with quite different likelihood values, gave apparently good fits to the observable data, conditional on patients not dropping out, but quite different results for the unobserved (and hypothetical) drop-out-free population profiles. This is an area for further work, as is the subject of diagnostic assessment of fitted models. We have used standard sub-models for measurements and event times, which are of proven generic value. However, we acknowledge that, in particular applications, close attention to model adequacy is advisable.

Finally, we note that this general class of models may be useful in different areas of application. The focus for the analysis of the schizophrenia trial data was to make inferences about the effects of the different treatments on the mean PANSS profiles over time, whilst adjusting for treatment-dependent and outcome-dependent drop-out. In other work in progress, we are using the same class of models in a problem concerning surrogate markers, where the focus is on using longitudinal measurements to improve prediction of survival prognosis.

ACKNOWLEDGEMENTS

This research has been supported in part through funding from the National Institute of Mental Health, grant number R01 MH56639, and in part by the Medical Research Council of the UK through the award of a research studentship to Angela Dobson.

The authors are grateful to Jane Xu and Scott Zeger for helpful discussions, Peter Ouyang for providing the PANSS data on behalf of Janssen Research Foundation, and the editor and reviewers for their comments on an earlier version.

REFERENCES

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. AND KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

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

DIGGLE, P. J. (1998). Dealing with missing values in longitudinal studies. In Everitt, B. S. and Dunn, G. (eds), *Recent Advances in the Statistical Analysis of Medical Data*, London: Arnold, pp. 203–228.

FAUCETT, C. L. AND THOMAS, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistical Methods in Medical Research* **15**, 1663–1686.

FAUCETT, C. L., SCHENKER, N. AND ELASHOFF, R. M. (1998). Analysis of censored survival data with intermittently observed time-dependent binary covariates. *Journal of the American Statistical Association* **93**, 427–437.

FINKELSTEIN, D. M. AND SCHOENFELD, D. A. (1999). Combining mortality and longitudinal measures in clinical trials. *Statistical Methods in Medical Research* **18**, 1341–1354.

HOGAN, J. W. AND LAIRD, N. M. (1997a). Mixture models for the joint distribution of repeated measures and event times. *Statistical Methods in Medical Research* **16**, 239–257.

HOGAN, J. W. AND LAIRD, N. M. (1997b). Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistical Methods in Medical Research* **16**, 259–272.

LAIRD, N. M. AND WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

LAVALLEY, M. P. AND DEGRUTOLLA, V. (1996). Models for empirical Bayes estimators of longitudinal CD4 counts. *Statistical Methods in Medical Research* **15**, 2289–2305.

MARDIA, K. V., KENT, J. T. AND BIBBY, J. M. (1979). *Multivariate Analysis*. London: Academic Press.

NEILSEN, G. G., GILL, R. D., ANDERSEN, P. K. AND SØRENSEN, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics* **19**, 25–43.

NELDER, J. A. AND MEAD, R. (1965). A simplex method for function minimisation. *Computer Journal* **7**, 308–313.

PAWITAN, Y. AND SELF, S. (1993). Modelling disease marker processes in AIDS. *Journal of the American Statistical Association* **88**, 719–726.

TAYLOR, J. M. G., CUMBERLAND, W. G. AND SY, J. P. (1994). A stochastic model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association* **89**, 727–736.

TSIATIS, A. A., DEGRUTTOLA, V. AND WULFSOHN, M. S. (1995). Modelling 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.

WULFSOHN, M. S. AND TSIATIS, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–339.

YAU, K. K. W. AND MCGILCHRIST, C. A. (1998). ML and REML estimation in survival analysis with time dependent correlated frailty. *Statistical Methods in Medical Research* **17**, 1201–1213.

[*Received November 22, 1999; revised June 7, 2000; accepted for publication June 27, 2000*]