Biostatistics (2002), **3**, 1, *pp.* 33–50
Printed in Great Britain

Identification and efficacy of longitudinal markers for survival

ROBIN HENDERSON*
Medical Statistics Unit, Lancaster University, LA1 4YF, UK
Robin.Henderson@lancaster.ac.uk

PETER DIGGLE
Medical Statistics Unit, Lancaster University, LA1 4YF, UK
Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205-2179, USA

ANGELA DOBSON
Medical Statistics Unit, Lancaster University, LA1 4YF, UK

SUMMARY

Methods for the combined analysis of survival time and longitudinal biomarker data have been developed in recent years, with most emphasis on modelling and estimation. This paper focuses on the use of longitudinal marker trajectories as individual-level surrogates for survival. A score test for association which requires only standard methods for implementation is derived for the initial identification of candidate biomarkers. Methods for assessing efficacy of markers are discussed and a measure contrasting conditional and marginal distibutions is proposed. An application using prothrombin index as biomarker for survival of liver cirrhosis patients is included.

Keywords: Explained variation; Joint modelling; Residual lifetime distribution; Score test; Surrogates.

1. INTRODUCTION

Interest in methods for the combined analysis of longitudinal and survival time data has developed considerably in recent years (for example, De Stavola and Christensen, 1996; Hogan and Laird, 1997; Wulfsohn and Tsiatis, 1997; Boscardin *et al.*, 1998; Bycott and Taylor, 1998; Henderson *et al.*, 2000). These methods are appropriate when sequential measurements of a biomarker are made on each subject in a clinical trial but the sequence can be terminated early either through death of the patient or withdrawal from the trial for other reasons. Figure 1 summarizes the results of a typical illustration, a controlled trial into prednisone treatment of liver cirrhosis patients reported by Andersen *et al.* (1993, p. 19) and others. The upper plot shows smooth estimates of prothrombin index, a measure of liver function, and the lower plot shows Kaplan–Meier survival curves. The estimates in the upper plot are based on mean values over all patients still alive at each time point, which means that the observed trends may be due to the attrition of high-risk patients rather than a general increase in mean prothrombin index with time since diagnosis. Joint longitudinal and survival methods are needed to investigate the development of prothrombin index over time and the relationship between prothrombin and survival.

*To whom correspondence should be addressed

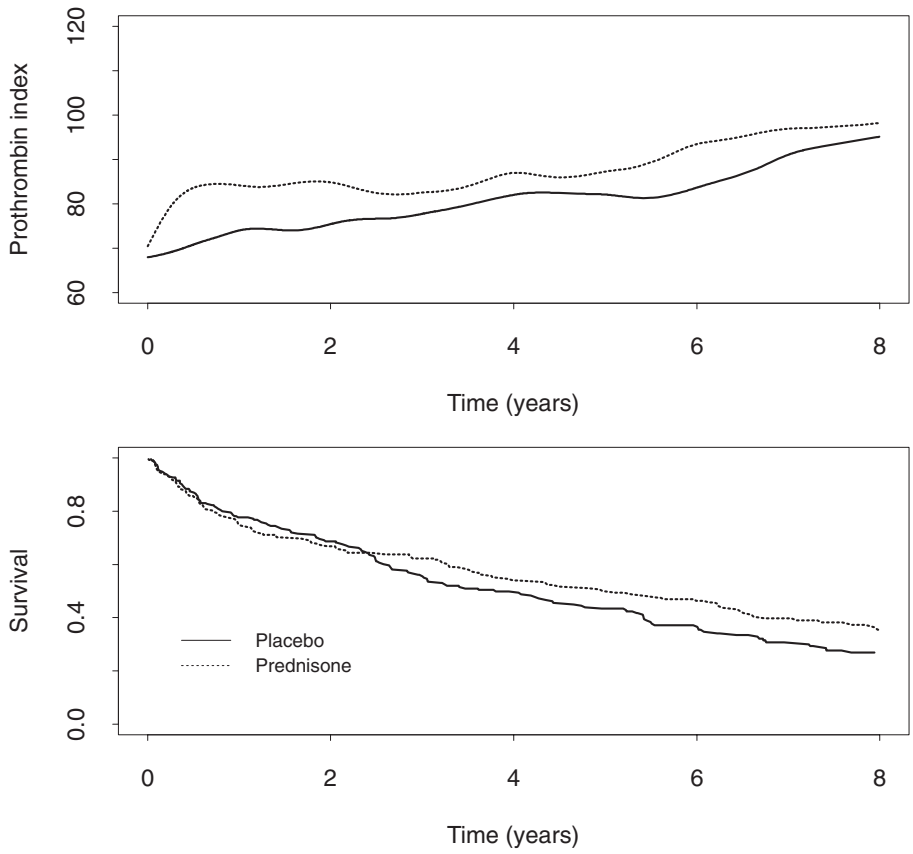

Fig. 1. Mean prothrombin index (upper plot) and patient survival (lower plot) for liver cirrhosis data.

One motivation for interest in joint modelling methods is the potential to exploit the longitudinal marker as a surrogate for subsequent survival (Prentice, 1989; Satten and Logini, 1996; Tsiatis *et al.*, 1991) The recent literature on the use of markers as surrogate endpoints distinguishes two points of view. Prentice (1989) defined surrogacy in terms of the equivalence of hypothesis tests for treatment effects, using data on either the surrogate or the true endpoint. Buyse *et al.* (2000) refer to this as *trial-level* surrogacy, and distinguish it from *individual-level* surrogacy. For the liver cirrhosis example for instance, prothrombin index would be considered to be a useful trial-level surrogate if the average effect of prednisone treatment on survival could be determined more quickly through the average effect on prothrombin index. At the individual level prothrombin index would be a useful surrogate if the trajectory of irregularly observed values available at any time for a single subject provides useful prognostic information on subsequent survival of that person. Building on earlier work by Freedman *et al.* (1992) and Buyse and Molenberghs (1998), Buyse *et al.* (2000) consider individual-level surrogates to be valid if there is strong association between an individual subject's values of the marker and the endpoint of interest. Intuitively, one might well expect that a good trial-level surrogate would also turn out to be a good individual-level surrogate, and vice versa, but clearly there is no logical reason why this must be so. The resulting scope for controversy as to what constitutes a 'good' surrogate marker is well demonstrated by Begg and Leung (2000) and the associated discussion by N. E. Day, S. W. Duffy and R. L. Prentice.

In this paper, we focus on individual-level surrogacy for a survival endpoint. We begin by considering the selection of marker variables. Sometimes a number of candidate biomarkers may be available (for example, Einspahr *et al.*, 1997) and the initial problem is to assess each for association with survival time, so that those with most potential can be investigated further. However, joint modelling techniques are invariably heavily computer intensive and therefore it would be useful to have a relatively quick and simple screen for the initial identification of association, to indicate whether further joint analysis is likely to add materially to the scientific interpretation of the data. For this purpose, we develop a score test for association between survival time and biomarker values. The score test requires only separate analyses of the two components, using standard techniques.

We then consider the efficacy of individual-level markers. From this point of view, an effective surrogate is one for which the conditional residual lifetime distribution accounting for marker information at an interim time is more strongly concentrated around the actual, but as-yet unobserved, survival time than is the marginal residual lifetime distribution ignoring the marker data. In Section 4 we propose a modification of an explained variation measure suggested by Schemper and Henderson (2000), which can be used to measure predictive accuracy with and without the interim longitudinal marker information. In Section 5 we employ the suggested measure in exploring the effectiveness of prothrombin as an individual-level surrogate for liver cirrhosis survival and also investigate sensitivity to choice of both model and interim time.

Our model, notation and assumptions are given in Section 2. In Section 3 we describe the score test for association and in Section 4 discuss the use of longitudinal information at an interim time in estimating probability of survival to some future time, including the proposed measure of variation explained by the marker. The illustration using prothrombin index as a marker for survival of liver cirrhosis patients is considered in detail in Section 5, including a comparison of joint modelling and estimation methods with simpler techniques. Some final remarks in Section 6 close the paper.

2. Model and notation

We have survival time and longitudinal information for m subjects. Longitudinal measurements are obtained intermittently and we allow the possibility of different numbers and timing of measurements for different subjects. A Gaussian linear model is assumed for the response Y_t at time t:

$$Y_t = x_1(t)'\beta_1 + W(t) + Z_t, \tag{1}$$

where $x_1(t)$ is a p_1-vector of explanatory variables, $W(t)$ is the value at time t of an unobserved zero-mean Gaussian random process, and Z_t denotes zero-mean Gaussian measurement error. Within-subject correlation in responses arises through serial correlation in the $W(t)$ process, the measurement errors being assumed to be mutually independent.

Survival time is associated with the longitudinal response through the effect of the latent process $W(t)$ but is otherwise conditionally independent. A semiparametric proportional hazards model is assumed, with intensity

$$\lambda(t) = H(t)\alpha_0(t)\exp\{x_2(t)'\beta_2 + \gamma W(t)\}. \tag{2}$$

Here $H(t)$ is a predictable 0–1 at-risk process, $\alpha_0(t)$ is an unspecified baseline hazard, and $x_2(t)$ is a p_2-vector of observed explanatory variables. A generalization under which $\gamma W(t)$ is replaced with a second random process correlated with $W(t)$ is possible (Henderson *et al.*, 2000), but not considered further in this paper. Noninformative right censoring of survival time is allowed as usual. We use generic notation T, Y and W for survival time, longitudinal responses and latent process, respectively.

A variety of estimation procedures are possible for this class of model, depending in part upon the assumptions required for $W(t)$, but including MCMC (Faucett and Thomas, 1996), EM (Wulfsohn and Tsiatis, 1997), and a combined simplex/EM procedure (Henderson *et al.*, 2000). We do not consider estimation explicitly in this work.

3. A SCORE TEST FOR ASSOCIATION

This test is based on separate analyses of the two components Y and T under the null hypothesis H_0 : $\gamma = 0$. Under this assumption Y is simply multivariate Gaussian and T follows a standard proportional hazards model.

With subscript i indexing subject, the score test statistic (Appendix A) is

$$U = \sum_{i=1}^{m} \int_0^{\tau} E_{W|Y}\{W_i(t)\}\, \mathrm{d}\hat{M}_i(t), \tag{3}$$

where the expectation is with respect to the conditional distribution of the latent process given the observed measurements, and

$$\hat{M}_i(t) = N_i(t) - \hat{\Lambda}_i(t) = N_i(t) - \int_0^t H_i(u) \exp\{x_{2i}(u)'\hat{\beta}_2\}\, \mathrm{d}\hat{A}_0(u)$$

is the estimated martingale process obtained from an analysis of the survival data. As usual, N_i is the counting process for subject i and $\hat{A}_0$ is the nonparametric estimator of the cumulative baseline hazard

$$\hat{A}_0(u) = \int_0^u \frac{J(s)}{\sum_{i=1}^m H_i(s) \exp\{x_{2i}(s)'\hat{\beta}_2\}}\, \mathrm{d}N(s),$$

where $N(s) = \sum N_i(s)$ and $J(s) = I\big(\sum H_i(s) > 0\big)$. Since $\hat{A}_0$ has jumps at event times only, $E_{W|Y}\{W_i(t)\}$ is required for a finite number of values of t only. Note that the unconditional expectation of each $\{W_i(t)\}$ term is zero for all t and the power of the test derives from the information which Y provides about W. Also, whilst (3) is a formal score test statistic, additionally it has a wider intuitive interpretation as follows. The term $E_{W|Y}\{W_i(t)\}$ estimates the value of $W_i(t)$; the term $\mathrm{d}\hat{M}_i(t)$ is the residual between observed and expected event-intensities for subject i at time t. The test statistic is then the covariance, integrated over time and summed over subjects, between these two empirical quantities.

The variance of U can be consistently estimated either by

$$V_1 = \sum_{i=1}^{m} \int_0^{\tau} [E_{W|Y}\{W_i(t)\}]^2\, \mathrm{d}\hat{\Lambda}_i(t)$$

obtained from the martingale formulation, or more directly by

$$V_2 = \sum_{i=1}^{m} \left[\int_0^{\tau} E_{W|Y}\{W_i^2(t)\}\, \mathrm{d}\Lambda_i(t) - \int_0^{\tau} \int_0^{\tau} \mathrm{cov}_{W|Y}\{W_i(t), W_i(s)\}\, \mathrm{d}M_i(t)\, \mathrm{d}M_i(s) \right].$$

In either case $U/V^{1/2}$ is asymptotically $N(0, 1)$ under the usual regularity conditions. In small samples an adjustment to the variance estimate to allow for the replacement of parameters with estimates may be useful, as sketched in the appendix. All terms required for the calculation of U and both variance estimates can be obtained from standard software packages.

Table 1. *Power of score test for $W(t)$ model with random intercept (variance σ_1^2) and stationary Gaussian process (variance σ_v^2 and lag one correlation $\rho(1)$). Overall lag one covariance is $\sigma_1^2 + \sigma_v^2\rho(1)$. Longitudinal and survival data are linked through the association parameter γ, with independence at $\gamma = 0$*

σ_1^2	σ_v^2	$\rho(1)$	$\sigma_1^2 + \sigma_v^2\rho(1)$	$\gamma = 0$	$\gamma = 0.1$	$\gamma = 0.25$	$\gamma = 0.5$
0.5	0.5	0.5	0.75	0.06	0.23	0.70	1.00
		0.05	0.525	0.05	0.08	0.55	1.00
0.8	0.2	0.5	0.90	0.06	0.26	0.86	1.00
		0.05	0.81	0.06	0.18	0.79	1.00
0.2	0.8	0.5	0.60	0.04	0.16	0.68	0.98
		0.05	0.24	0.03	0.08	0.21	0.77

We have conducted several simulation studies to check the adequacy of the asymptotic approximations to the null distribution of the proposed test, to compare the alternative variance estimators, and to give some indication of power to detect association between Y and T. A small sample of simulation results is shown in Table 1. For these we considered a single group of subjects with target measurement times 0, 1, 2 and 3 units. The longitudinal measurement process Y had mean response $E[Y(t)] = 5+t$, measurement error variance $\sigma_z^2 = 0.25$ and latent process $W(t) = U_1 + V(t)$. Here U_1 is an $N(0, \sigma_1^2)$ random effect and $V(t)$ is a stationary Gaussian process with variance σ_v^2 and correlation function $\rho(s) = \text{corr}(V(t), V(t+s)) = \exp(-|s|/\phi)$, independent of U_1.

The variances of the random intercept, σ_1^2, and of the stationary Gaussian process, σ_v^2, were chosen to sum to one in the simulations, thus giving an $N(0, 1)$ marginal distribution for the latent process $W(t)$ at all t. This helps in the interpretation of the scale parameter γ, which can be considered crudely as the effect of an omitted standard normal covariate. Two values of ϕ were used, to give lag 1 correlation in the $V(t)$ process of either 0.5 or 0.05. The resulting covariance between successive response measurements, $\sigma_1^2 + \sigma_v^2\rho(1)$, is included in the tabulated results to aid interpretation.

Survival times were generated according to (2), using Weibull baseline survival with survivor function $S_0(t) = \exp(-0.1t^2)$, and a single time-constant standard normal covariate for each subject with associated regression coefficient $\beta_2 = 1$. Censoring was taken at a maximum follow-up time $\tau = 5$ only, leading to around 15–20% censored observations in the simulations reported here. In additional simulations we found that varying τ over a range of values greater than the final measurement time had relatively little effect.

Table 1 gives the estimated power for a nominal 5% test of $H_0 : \gamma = 0$ for four true values of γ, based in each case on 200 simulations with $m = 250$ subjects. Parameter values required for the calculation of the test statistic were replaced with estimates obtained from separate analyses of the two components, with a semiparametric approach for the survival times, not requiring knowledge of the Weibull baseline. Results are given only for the martingale-based variance estimator V_1 and with correction for parameter estimation: results without correction or with the more complicated V_2 were very similar.

Results in Table 1 and from other simulations show that: attained test sizes are close to nominal values based on the asymptotic $N(0, 1)$ distribution; power increases as γ moves away from zero, as expected; for fixed γ, power also increases with the correlation between successive longitudinal measurements, again as expected because the observed measurements from each subject then provide more information on the unobserved value of W_1 at event times; there was little difference between V_1 and V_2 as variance estimators, and little effect of adjustment for parameter estimation. Since V_1 and V_2 perform similarly,

the simpler V_1 can be recommended, and on the evidence of our results the correction term does not seem necessary. Further information on the simulation study is available on request to the corresponding author.

4. LONGITUDINAL MARKERS FOR SURVIVAL

4.1 *Interim survival estimation*

Now suppose a longitudinal marker has been identified and a joint model has been fitted to sample data. In subsequent practice we may wish to use the model to make inferences about patient survival to some future time τ_2 given all information available at some earlier time τ_1. Letting Y_{01} denote all longitudinal measurements obtained on the subject over the interval $[0, \tau_1]$, we assume there is interest in

$$S(\tau_2 \mid \tau_1, Y_{01}) = P(T > \tau_2 \mid T > \tau_1, Y_{01}).$$

Evaluation of the conditional probability of surviving to τ_2 involves taking expectations with respect to the unobserved latent process $W(t)$. Strictly, this is a continuous-time process but under the semiparametric approach the estimated baseline hazard is zero except at observed event times, and for this reason the expectation only involves the values of $W(t)$ at a finite number of time points, namely the measurement times and the event times. Let W_{01} and W_{02} denote the corresponding vectors of values of $W(t)$ at measurement or event times within the intervals $[0, \tau_1]$ and $[0, \tau_2]$, respectively. Using $f(\cdot)$ to denote density and exploiting the conditional independence of survival time and longitudinal measurements given the latent process, we then have

$$\begin{aligned} S(\tau_2 \mid \tau_1, Y_{01}) &= \int P(T > \tau_2 | T > \tau_1, W_{02}) f(W_{02} | T > \tau_1, Y_{01}) \, \mathrm{d}W_{02} \\ &= \frac{\int P(T > \tau_2 | T > \tau_1, W_{02}) P(T > \tau_1 | W_{02}) f(W_{02} | Y_{01}) \, \mathrm{d}W_{02}}{P(T > \tau_1 | Y_{01})} \\ &= \frac{\int P(T > \tau_2 | W_{02}) f(Y_{01} | W_{02}) f(W_{02}) \, \mathrm{d}W_{02}}{\int P(T > \tau_1 | W_{02}) f(Y_{01} | W_{02}) f(W_{02}) \, \mathrm{d}W_{02}} \\ &= \frac{\int P(T > \tau_2 | W_{02}) f(Y_{01} | W_{01}) f(W_{02}) \, \mathrm{d}W_{02}}{\int P(T > \tau_1 | W_{01}) f(Y_{01} | W_{01}) f(W_{01}) \, \mathrm{d}W_{01}}. \end{aligned} \tag{4}$$

We will be interested in comparing this conditional probability with the corresponding marginal value ignoring any information in Y_{01}:

$$S(\tau_2 \mid \tau_1) = \frac{\int P(T > \tau_2 | W_{02}) f(W_{02}) \, \mathrm{d}W_{02}}{\int P(T > \tau_1 | W_{01}) f(W_{01}) \, \mathrm{d}W_{01}}. \tag{5}$$

The term $f(Y_{01}|W_{01})$ in the integrands in (4) is a weighting which reflects the relevant information in Y_{01}. At one extreme there may be so much measurement error in the longitudinal data that Y_{01} provides no information about the latent process. In that case $f(Y_{01}|W_{01})$ is constant for all W_{01} and (4) reduces to (5). At the other extreme, if it is possible to completely determine the true value, W_{02}^0 say, of W_{02} from Y_{01}, then $f(Y_{01}|W_{01})f(W_{02})$ and $f(Y_{01}|W_{01})f(W_{01})$ are zero except at $W_{02} = W_{02}^0$. Then we have maximum information from Y_{01} and

$$S(\tau_2 \mid \tau_1, Y_{01}) = S(\tau_2 \mid \tau_1, W_{02}^0). \tag{6}$$

In practice (4) will be intermediate between (5) and (6).

4.2 *Measuring marker effectiveness*

To judge the effectiveness of the marker in helping predict survival to the endpoint τ_2, we need a way of comparing each of (4) and (5) with the actual survival time. We propose a modification of a measure of explained variation for semiparametric survival models suggested by Schemper and Henderson (2000). This is based on mean absolute deviation between observed and estimated survival status and can be adjusted to either a single time τ_2 or an interval $[\tau_1, \tau_2]$.

Define $S_i^0(t)$ to be the value at time t of the observed *survivor process* for individual i, which takes the value one if the individual was known to be alive at t, zero if the individual died before t, and is undefined if the survival time for the individual was censored before t. Then, if the marker is effective we would expect to find, on average, relatively small absolute deviations between $S_i^0(\tau_2)$ and the corresponding estimates $S_i(\tau_2 \mid \tau_1, Y_{01,i})$ from (4), where the subscript i identifies an individual subject. Estimation is complicated when there is censoring, as simply deleting cases with undefined $S_i^0(\tau_2)$ would lead to bias in the observed mean absolute deviation as an estimator of the underlying population quantity. However, a slight adjustment of the estimator of Schemper and Henderson (2000), so as to condition on prior survival to τ_1, leads to the unbiased estimator

$$\begin{aligned} M_Y(\tau_1, \tau_2) = \frac{1}{r(\tau_1)} \sum_{i:t_i \geqslant \tau_1} [& I(t_i \geqslant \tau_2)(1 - S_i(\tau_2 \mid \tau_1, Y_{01,i})) + \delta_i I(t_i < \tau_2) S_i(\tau_2 \mid \tau_1, Y_{01,i}) \\ & + (1 - \delta_i) I(t_i < \tau_2)\{(1 - S_i(\tau_2 \mid \tau_1, Y_{01,i}))S_i(\tau_2 \mid t_i, Y_{01,i}) \\ & + S_i(\tau_2 \mid \tau_1, Y_{01,i})(1 - S_i(\tau_2 \mid t_i, Y_{01,i}))\}], \end{aligned} \tag{7}$$

where $r(\tau_1)$ is the number of subjects at risk at τ_1 and δ_i is an indicator of censoring ($\delta_i = 0$) or observed failure ($\delta_i = 1$). The first term in the expression for $M_Y(\tau_1, \tau_2)$ is the contribution of those known still to be alive at t ($S_i^0(\tau_2) = 1$) and the second is the contribution of those known to have died ($S_i^0(\tau_2) = 0$). Patients with censored follow-up before τ_2 are allocated to $S_i^0(\tau_2) = 1$ or $S_i^0(\tau_2) = 0$ in accordance with the respective conditional probabilities, leading to the final two terms.

A second estimator $M(\tau_1, \tau_2)$ can be defined in a similar way to $M_Y(\tau_1, \tau_2)$ to measure predictive accuracy without knowledge of Y_{01}, by using in (7) the marginal values (5) in place of the conditional probabilities (4). A comparison between $M_Y(\tau_1, \tau_2)$ and $M(\tau_1, \tau_2)$ measures the improvement in predictive accuracy obtained through marker information; for example, we might use a relative measure,

$$R_M(\tau_1, \tau_2) = 1 - M_Y(\tau_1, \tau_2)/M(\tau_1, \tau_2).$$

In principle, variance estimates for the statistics $M(\tau_1, \tau_2)$ and $M_Y(\tau_1, \tau_2)$ can be obtained and used for formal inference. We have not attempted this as our intention is that they should be used and interpreted as informal summary measures, rather like the familiar R^2 from linear regression.

Instead of measuring mean absolute deviation at a single time τ_2, we may prefer to form an average over an interval $[\tau_1, \tau_2]$. Schemper and Henderson (2000) suggest an integration of the absolute deviation between observed survivor processes and estimated survivor curves, with weights in proportion to failure density. Their statistic can be again adapted to provide an unbiased estimator,

$$D_Y(\tau_1, \tau_2) = \frac{\sum_{i:\tau_1 \leqslant t_i \leqslant \tau_2} \delta_i \{\hat{G}(\tau_1)/\hat{G}(t_i)\} M_Y(\tau_1, t_i)}{\sum_{i:\tau_1 \leqslant t_i \leqslant \tau_2} \delta_i \{\hat{G}(\tau_1)/\hat{G}(t_i)\}}.$$

Here $\hat{G}(\cdot)$ is the Kaplan–Meier estimator of the censoring time distribution, which is used to compensate for the loss of censored cases. Again, a similar estimator $D(\tau_1, \tau_2)$ can be defined to measure

predictive accuracy without Y_{01}, by replacing $M_Y(\cdot, \cdot)$ with $M(\cdot, \cdot)$. These values may be used directly or to construct a relative measure

$$R_D(\tau_1, \tau_2) = 1 - D_Y(\tau_1, \tau_2)/D(\tau_1, \tau_2).$$

All measures are invariant to monotonic transformations of the time scale and are consistent under random censoring of the underlying population quantities. See Schemper and Henderson (2000) for further information.

Choice of fixed point (M) or interval (D) measures will be application-specific. Sometimes there may be a clinically recognized fixed survival time, such as one-year graft survival being used as a measure of the effectiveness of a renal transplant, in which case M-measures should be used. On the other hand, the D-measures present a more complete summary rather than a snapshot and can also be defined for complete follow-up by letting τ_2 increase indefinitely.

Finally, two cautionary notes are worth mentioning, whichever method of measuring marker effectiveness is selected. Firstly, explained variation measures in survival tend to be low, particularly under proportional hazards models, even when there are strong covariate/marker effects. Secondly, having conditional measures M_Y and D_Y close to their marginal counterparts M and D indicates that there is little to be gained on average from knowledge of Y, but this can mask the potential for quite useful information on individual subjects with relatively extreme longitudinal trajectories, as is illustrated in the following section.

5. APPLICATION TO LIVER CIRRHOSIS TRIAL

5.1 *Modelling*

We consider the liver cirrhosis trial introduced in Section 1 and previously described by Andersen *et al.* (1993, p. 19) and others. Data are available for $m = 488$ patients, randomly allocated at diagnosis to prednisone (251) or placebo (237) and followed until death or end of study, some 12 years after the first patients were recruited. A number of variables were recorded at entry and throughout the study, though here we concentrate on just two: treatment and repeated *prothrombin index* measurements. The latter, which might be considered as a marker for disease progression, forms our longitudinal component Y. Measurements were obtained at entry and then scheduled for 3, 6, 12 months and annually thereafter, though the achieved times and numbers varied considerably between patients, with up to 17 values for some.

The upper plot of Figure 1 shows smooth estimates of mean prothrombin index for surviving patients in each of the groups, truncated at 8 years follow-up since after that time relatively few patients remained at risk and prothrombin data are sparse. Smoothing is necessary because measurement times differed between patients. The smooth profiles mask considerable variability between patients: four illustrative cases will be presented later. In both groups there is an increase in average prothrombin over time from means around 70 ('abnormal', Andersen *et al.* (1993, p. 33)) at entry to close to the 'normal' 100 after 8 years follow-up. There is an initial steep increase between baseline and 3 month measurement for the prednisone treated group, which is not the case in the placebo group. This and the later continued increase may in part be a consequence of diagnosis and subsequent entry into the study often being made following hospitalization due to exacerbation of symptoms (Christensen *et al.*, 1986). However, mean profiles based on all patients at risk may be misleading because death of high-risk patients with lower prothrombin values would also lead to an apparent increase over time in mean prothrombin level amongst patients still alive. The survivor curves in the lower plot of Figure 1 indicate that only some 30% of patients survive 8 years from entry, with slightly better prognosis for prednisone-treated patients than for those given placebo only.

We assume the overall model of Section 2 and consider three different submodels for the latent process $W(t)$:

$$\begin{aligned}
&\text{A: } W(t) = U_1 && U_1 \sim N(0, \sigma_1^2)\\
&\text{B: } W(t) = U_1 + U_2 \times t && U_1 \sim N(0, \sigma_1^2),\ U_2 \sim N(0, \sigma_2^2),\\
& && \text{Corr}(U_1, U_2) = \rho\\
&\text{C: } W(t) = U_1 + V(t) && U_1 \sim N(0, \sigma_1^2),\ V(t) \sim N(0, \sigma_v^2),\\
& && \text{Corr}(V(t), V(t+s)) = \exp(-|s|/\phi).
\end{aligned}$$

The mean prothrombin response Y is assumed to be linear in time, with separate slope and intercept for each group. To describe the sharp increase in the prednisone group in mean response between time zero and the first measurement time we include an indicator variable for measurements at time zero. For the survival component we assume a proportional hazards model with a time-constant treatment effect.

We first analysed the survival and prothrombin data separately, leading to maximized log likelihoods.

	Prothrombin	Survival	Combined
A	−13 300.655	−1829.509	−15 130.164
B	−13 231.718	−1829.509	−15 061.227
C	−13 213.531	−1829.509	−15 043.040

Standardized score statistics $U/V^{1/2}$ using the martingale variance formulation are −9.07, −10.23 and −8.01 for models A, B and C respectively, providing very strong evidence of association between survival time T and prothrombin index Y. With the interpretation of the score statistic as a covariance between the latent process and the martingale residuals between observed and expected numbers of events, the negative signs indicate that high prothrombin is associated with long-term survival (fewer events than expected).

Parameter estimates, standard errors and maximized log likelihoods obtained under joint analyses are given in Table 2. We used the method described by Henderson *et al.* (2000) for fitting all models, with standard errors obtained by Monte Carlo methods based on re-estimation from simulated data. For each of models A, B and C there is strong evidence of association between Y and T, with large increases in log likelihood in comparison with the values obtained under separate analyses, and negative estimates of γ confirming that high prothromobin levels are associated with reduced risk.

To assess model adequacy we compared the observed data with simulations generated under the fitted models. Details are omitted except to report that although the log-likelihood values strongly favour models B and C over model A, all three models gave good agreement between observed and simulated mean prothrombin profiles for patients still at risk, and between observed and simulated Kaplan–Meier survival plots, for each treatment group. However, there was a bigger difference between models with respect to estimated underlying prothrombin profiles for hypothetical dropout-free populations (Figure 2). As expected, these all fall below the observed profiles as time increases, because of the loss of the observed data from high-risk patients with low prothrombin levels. Note, however, the more pronounced difference between observed and underlying dropout-free means under model B than under the other models. This type of pattern was also seen for the schizophrenia data analysed by Henderson *et al.* (2000), where there was also a particularly large difference between observed and estimated underlying profiles when a random slope term was allowed in the latent process model. Sensitivity to modelling assumptions when the data are incomplete is a topic which requires further research beyond the scope of this paper.

Table 2. *Liver cirrhosis trial results for three different $W(t)$ models: random intercept (A); random intercept and random slope (B); and random intercept and stationary Gaussian process (C)*

	A : $W(t) = U_1$		B : $W(t) = U_1 + U_2 t$		C : $W(t) = U_1 + V(t)$	
	Est	SE	Est	SE	Est	SE
Prothrombin β_1						
Constant	69.813	1.391	72.111	1.442	69.282	0.453
Treatment, P	10.806	2.024	11.425	2.111	11.944	0.607
Time, t	1.714	0.224	−0.463	0.521	1.156	0.193
$P \times t$	−1.073	0.344	−0.758	0.710	−0.890	0.155
Time = 0, B	−0.265	1.442	−2.011	1.423	−0.629	0.995
$P \times B$	−11.214	1.954	−11.225	2.112	−12.313	1.430
Survival β_2						
Treatment	−0.037	0.153	−0.075	0.146	−0.185	0.099
Latent association						
γ	−0.037	0.004	−0.045	0.004	−0.025	0.001
Random effects						
σ_1^2	383.259	29.471	335.729	27.335	280.237	37.905
σ_2^2			20.686	2.541		
ρ			0.066	0.097		
σ_v^2					237.128	29.267
ϕ					1.915	0.517
Noise						
σ_z^2	334.629	8.872	289.388	9.242	210.105	19.679
Log likelihoods						
Prothrombin	−13302.174		−13240.952		−13223.483	
Survival	−1777.879		−1751.225		−1760.576	
Combined	−15080.053		−14992.177		−14984.059	

5.2 *Prothrombin as marker for survival*

Having fitted an adequate joint model and shown strong association between prothrombin and survival, we now investigate the use of prothrombin information as a marker for survival. For illustrative purposes we assume the interim time τ_1 is 3 years and the endpoint τ_2 is either 4.5 or 6 years. To be clinically useful τ_1 should not be too close to τ_2, but on the other hand should be large enough to allow accrual of marker information.

Table 3 gives the actual and relative estimated mean absolute deviations between true survival status and estimated survival probabilities at the endpoint τ_2 and averaged over the interval $[\tau_1, \tau_2]$, for each of the latent association models A, B and C. Also shown are the corresponding values using two versions of a simpler approach whereby the prothrombin measurement is used as a time-dependent covariate in a straightforward proportional hazards model for the data. For model D we used as covariate at time t the mean value of all prothrombin measurements obtained up to that time, and for model E we used only the most recent measurement. In both cases there was a highly significant effect of the covariate, with regression coefficients (and standard errors) −0.032(0.003) and −0.035(0.003): high prothrombin index is associated with low hazard, as expected.

Table 3 shows that the five models lead to similar values for all of the summary measures considered

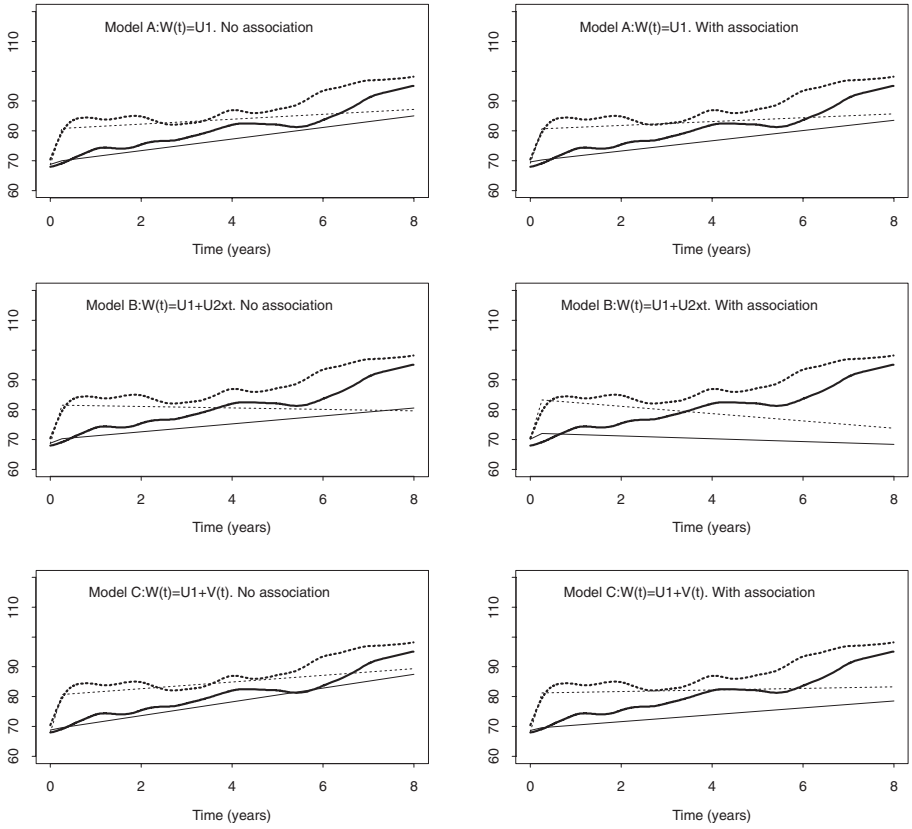

Fig. 2. Observed (bold) and hypothetical dropout-free means (fine) for placebo (solid) and prednisone (broken) groups.

and that, as expected, prediction to the nearer horizon 4.5 is more accurate than to the further horizon 6. Whilst knowledge of the longitudinal data Y_{01} does lead to improved prediction accuracy, because each value of $M_Y(\cdot,\cdot)$ is smaller than the corresponding value of $M(\cdot,\cdot)$, in relative terms the improvements are modest. Despite a highly statistically significant association between prothrombin and survival, there seems to be little prognostic capability in knowledge of the prothrombin trajectory for the majority of patients. This is a common feature of proportional hazards survival models (Schemper and Henderson, 2000).

The similarity of the mean absolute deviations masks some quite large differences between the results from different models for individual patients and the potential for prothrombin to be a useful marker for patients with the more extreme values. Figure 3 illustrates this by comparing $S(\tau_2|\tau_1, Y_{01})$ at $\tau_1 = 3$ and $\tau_2 = 6$ for the five different models, with model A taken as reference. There can be some substantial differences between estimates under models A and B, with quite large scatter about the $1:1$ line. There is less scatter when model C is compared with model A, though estimates under the former are attentuated towards the marginal estimate, ignoring the information in the longitudinal data. This reflects the acceptance under model C of further unpredictable variation in $W(t)$ over $[\tau_1, \tau_2]$ no matter how much information is obtained up to τ_1, because of the effect of the stationary component $V(t)$ in the model for $W(t)$. Model D, which uses the mean prothrombin value obtained to date as a time-dependent covariate

Table 3. *Measures of marker effectiveness for liver cirrhosis data. Estimation at $\tau_1 = 3$ years and prognosis to $\tau_2 = 4.5$ or 6 years. Values M and M_Y estimate mean absolute deviation at endpoint τ_2 between observed and fitted survival, values D and D_Y average over (τ_1, τ_2). Relative measures are $R_M = 1 - M/M_Y$ and $R_D = 1 - D/D_Y$. Models A–E are described in the text*

	Conditional on $T > \tau_1$ and Y_{01}			
	$M_Y(3, 4.5)$	$M_Y(3, 6)$	$D_Y(3, 4.5)$	$D_Y(3, 6)$
A	0.271	0.384	0.090	0.226
B	0.272	0.380	0.091	0.226
C	0.265	0.393	0.085	0.222
D	0.274	0.386	0.092	0.229
E	0.263	0.366	0.087	0.218
	Conditional on $T > \tau_1$			
	$M(3, 4.5)$	$M(3, 6)$	$D(3, 4.5)$	$D(3, 6)$
A	0.289	0.423	0.094	0.241
B	0.297	0.434	0.096	0.247
C	0.289	0.430	0.093	0.241
	Relative values			
	$R_M(3, 4.5)$	$R_M(3, 6)$	$R_D(3, 4.5)$	$R_D(3, 6)$
A	0.061	0.093	0.041	0.061
B	0.085	0.125	0.058	0.084
C	0.082	0.086	0.080	0.081

in a proportional hazards model, gives results which are very similar to the random intercept model A, as expected. In contrast, there can be very large differences between model A and model E, which uses only the most recent prothrombin value as a covariate.

With intermittently observed markers, the results can be sensitive to the choice of the interim analysis time τ_1, as the estimated survival probabilty to an endpoint τ_2 changes in response to new marker data. To illustate this, we now fix τ_2 at 6 years and let τ_1 increase from 0 to 3 years. Figures 4 and 5 show how the estimated conditional probability of surviving to the 6 year endpoint develops as τ_1 increases, for four individuals selected to illustrate the types of patterns which can occur. For reference, estimates in Figure 3 for these four people are marked with circles labelled 1–4. Figure 4 shows values under model C, which is the best-fitting model overall as judged by the likelihood criterion, with and without using prothrombin information, and under model B with the use of prothrombin information. Figure 5 again shows the marginal estimates under model C ignoring prothrombin, which acts as a reference, but includes also estimates under the two proportional hazards models, D and E. Results for model A are very close to those for model C and are not presented. It is clear that there can be high sensitivity to changes in marker values, especially for low τ_1. For most models, this sensitivity decreases as more values are recorded and some stability is achieved. An exception is the Cox model using only the most recent value of Y as covariate, where necessarily the sensitivity is both high and maintained over time no matter how many previous measurements are available.

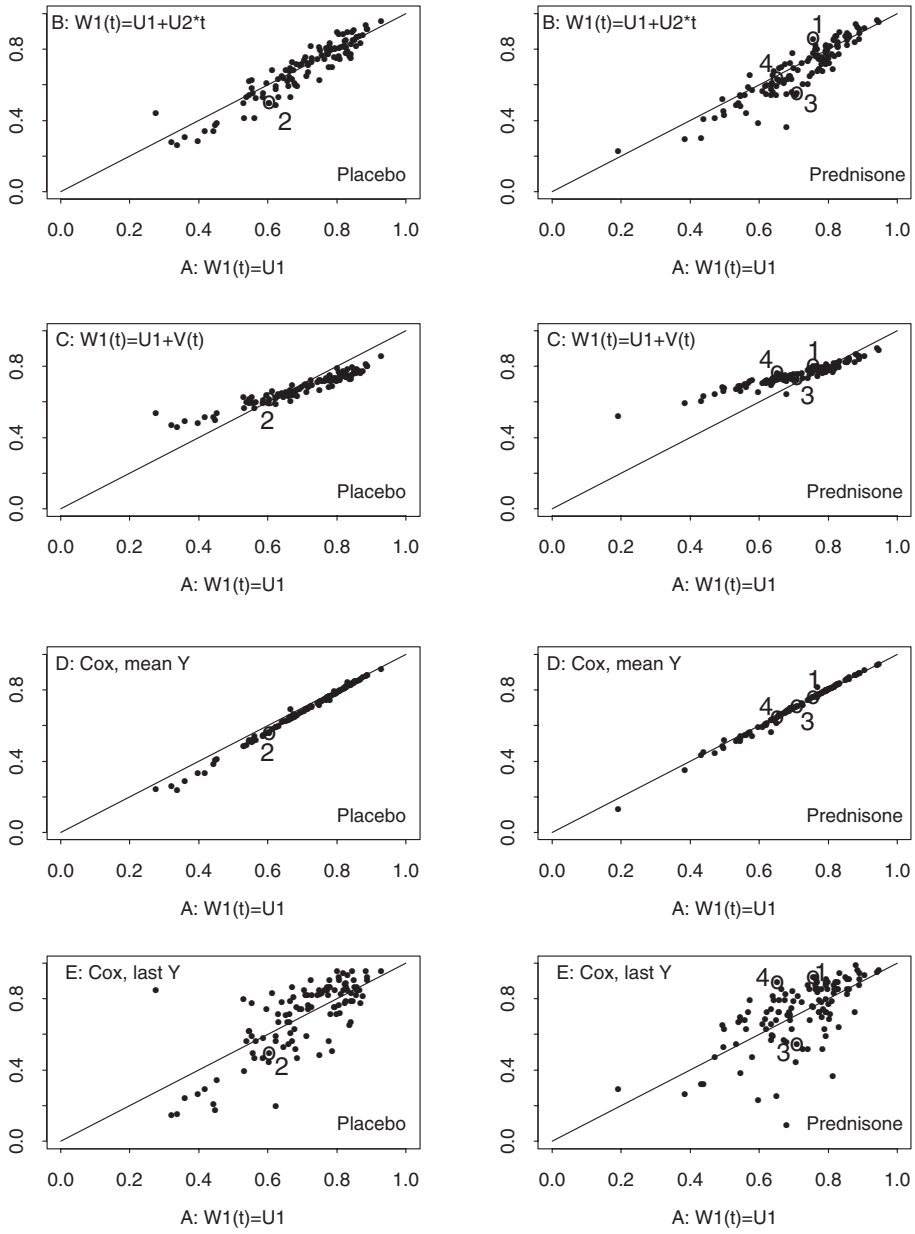

Fig. 3. Comparison of survival probabilities to $\tau_2 = 6$ years made at $\tau_1 = 3$ years, for five different models. In each plot the horizontal axis shows the subject-specific estimate of surviving to six years made under model A and given all information available at three years. The vertical axes show the corresponding estimates under each of the other models.

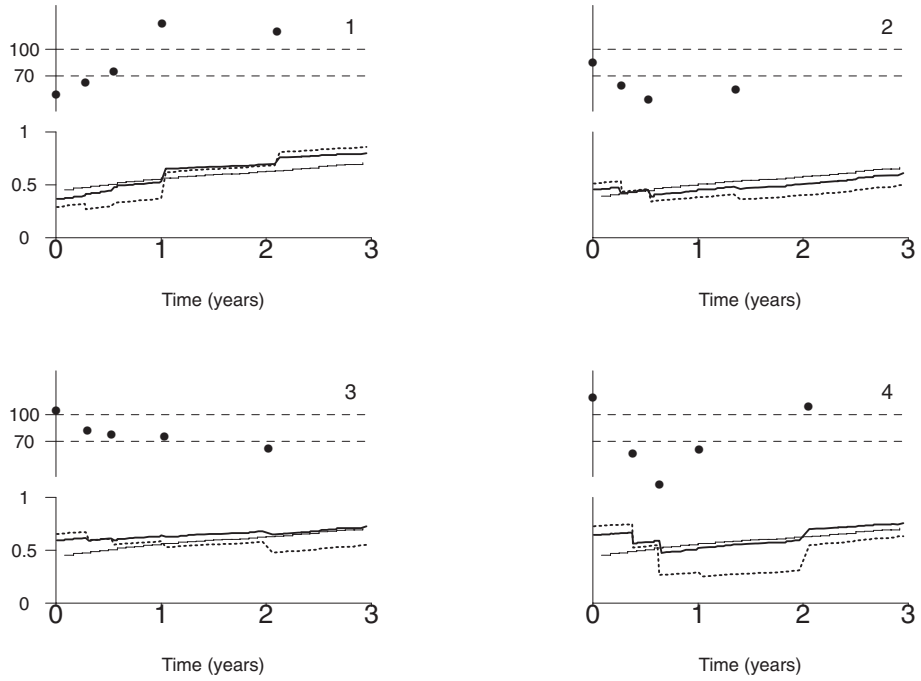

Fig. 4. Four case illustrations of the development of $S(\tau_2|\tau_1)$. The upper part of each plot shows prothrombin value against time recorded. The lower part shows how the estimated conditional probability of surviving to time $\tau_2 = 6$ years changes as information is accrued over the first three years: model C, ignoring Y_{01} (fine solid curve); model C using Y_{01} (bold solid curve); model B using Y_{01} (broken curve).

6. DISCUSSION

In this work we have concentrated on the initial screening of biomarkers for survival and the subsequent use of marker information in attempting to improve survival probability estimates for individual patients. In the terminology of Buyse *et al.* (2000), this implies a concern for individual-level surrogacy, rather than for trial-level surrogacy. Within this context, the score test for association could be used for interim analysis, to help decide whether continued monitoring of large numbers of potential surrogate markers is worthwhile. The predictive calculations of survival probabilities conditional on observed values to date of a selected marker are a potential contribution to prognosis for individual patients. In this respect, one conclusion is that prognosis should somehow integrate the information available over time for the patient in question, to avoid over-sensitivity to single atypical measurements. A second conclusion is that different models which fit average characteristics equally well may differ substantially in their predictions for individual patients. This is perhaps not surprising.

The area of joint modelling of longitudinal measurements and survival outcomes is closely related to the problem of dealing with potentially informative dropout in longitudinal studies, where the consensus is that conclusions can be sensitive to modelling assumptions which are difficult, or even impossible, to validate from the available data. See, for example, Scharfstein *et al.* (1999) and the associated discussion. In this respect, our experience has been that the widely used random intercept and slope model for longitudinal data introduced by Laird and Ware (1982) is particularly fragile when applied to data consisting of relatively long sequences but with a relatively high dropout rate.

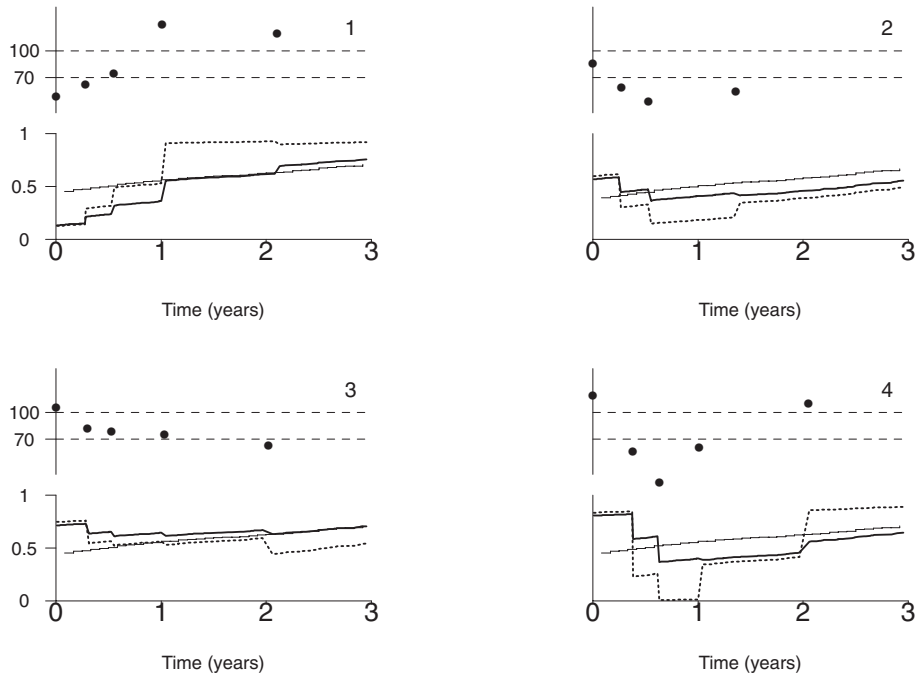

Fig. 5. As previous figure. Lower plots from: model C ignoring Y_{01} (fine solid curve); Cox with mean Y_{01} as covariate (bold solid curve); Cox with most recent Y as covariate (broken curve).

ACKNOWLEDGEMENTS

We thank Per Kragh Andersen for providing the liver cirrhosis data considered in Section 5 and the editor and referees for useful comments. This research has been supported by National Institute of Mental Health grant number R01 MH56639, NATO grant CRG960737, and a UK Medical Research Council studentship to AD.

APPENDIX: SCORE TEST FOR ASSOCIATION

Let the combined vector of unknown parameters be $(\theta, \gamma, \beta_2, A_0)$, where θ includes all parameters determining the distribution of Y. Let the maximum follow-up time be τ. Assume temporarily that β_2 and $A_0(t)$ are known and that A_0 is continuous. If W is known we can write the conditional likelihood of the event history data as

$$L_\gamma = \left(\prod_t \prod_i (\mathrm{e}^{x_{2i}(t)'\beta_2 + \gamma W_i(t)} \, \mathrm{d}A_0(t))^{\Delta N_i(t)} \right) \exp\left\{ - \int_0^\tau S_\gamma^{(0)}(t, W, \beta_2) \, \mathrm{d}A_0(t) \right\},$$

where

$$S_\gamma^{(0)}(t, W, \beta_2) = \sum_{i=1}^m H_i(t) \, \mathrm{e}^{x_{2i}(t)'\beta_2 + \gamma W_i(t)}.$$

Let

$$U_\gamma(\tau) = \sum_{i=1}^{m} \left\{ \int_0^\tau W_i(t)\, \mathrm{d}N_i(t) - \int_0^\tau W_i(t) H_i(t)\, \mathrm{e}^{x_{2i}(t)'\beta_2 + \gamma W_i(t)}\, \mathrm{d}A_0(t) \right\}$$

and note that

$$\frac{\partial L_\gamma}{\partial \gamma} = U_\gamma(\tau) L_\gamma.$$

With $l_1(\theta, Y)$ denoting the marginal likelihood of the longitudinal measurements, the overall log likelihood is

$$l = l_1(\theta, Y) + \log E_{W|Y}[L_\gamma]$$

which has derivative with respect to γ

$$\frac{\partial l}{\partial \gamma} = \frac{E_{W|Y}[U_\gamma(\tau) L_\gamma]}{E_{W|Y}[L_\gamma]}.$$

Hence the score statistic is

$$\begin{aligned} U(\tau) &= E_{W|Y}[U_0(\tau)] \\ &= E_{W|Y}\left[\sum_{i=1}^{m} \left\{ \int_0^\tau W_i(t)\, \mathrm{d}N_i(t) - \int_0^\tau W_i(t) H_i(t)\, \mathrm{e}^{x_{2i}(t)'\beta_2}\, \mathrm{d}A_0(t) \right\} \right] \\ &= \sum_{i=1}^{m} \int_0^\tau E_{W|Y}[W_i(t)]\, \mathrm{d}M_i(t) \end{aligned}$$

where

$$M_i(t) = N_i(t) - \Lambda_i(t) = N_i(t) - \int_0^t H_i(u)\, \mathrm{e}^{x_{2i}(u)'\beta_2}\, \mathrm{d}A_0(u)$$

is the usual counting process martingale.

Now consider $U(\tau)$ to be a particular value of a process $\{U(s) : s > 0\}$. Since W is predictable it follows that $U(s)$ is itself a martingale process with predictable variation

$$V_1(s) = \sum_{i=1}^{m} \int_0^s E_{W|Y}[W_i(t)]^2\, \mathrm{d}\Lambda_i(t).$$

With independence between subjects and under mild conditions, the martingale central limit theorem implies $U(\tau)/\sqrt{V_1(\tau)}$ is asymptotically $N(0, 1)$ under H_0 as $m \to \infty$ (Andersen *et al.*, 1993, p. 83).

Alternatively, regular likelihood theory can be used to provide an information-based variance for the score statistic:

$$V_2(\tau) = \sum_{i=1}^{m} \left\{ \int_0^\tau E_{W|Y}[W_i^2(t)]\, \mathrm{d}\Lambda_i(t) - \int_0^\tau \int_0^\tau \mathrm{Cov}_{W|Y}(W_i(t), W_i(s))\, \mathrm{d}M_i(t)\, \mathrm{d}M_i(s) \right\}.$$

Again the limiting distribution of $U(\tau)/\sqrt{V_2(\tau)}$ is $N(0, 1)$ under mild conditions.

In practice, we replace the unknown β_2 by the usual maximum partial likelihood estimator $\hat{\beta}_2$ and $A_0(t)$ by the non-parametric maximum likelihood estimator (under H_0)

$$\hat{A}_0(t) = \int_0^t \frac{J(u)}{\sum_{i=1}^m H_i(u)\, \mathrm{e}^{x_{2i}(u)'\hat{\beta}_2}}\, \mathrm{d}N(u)$$

where $N(u) = \sum N_i(u)$ and $J(u) = I\big(\sum H_i(u) > 0\big)$. Thus the expectations need only be determined at event times.

Some compensation may be required when sample size is small to allow for the effect of using estimated values. In practice, the effect of estimating θ is likely to be small (as it is obtained from the measurement data only) but estimation of β_2 may inflate the variance estimate and lead to a conservative test. For example, Crowder and Kimber (1997) show this to be a severe problem in a parametric test for frailty. In the semiparametric case a variance estimate corrected for uncertainty in β_2 is of the form

$$V_c = V - J' I(\beta_2)^{-1} J$$

where $I(\beta_2)$ is the information matrix obtained from the partial likelihood and $J = E[\partial U/\partial \beta_2]$ (Commenges and Andersen, 1995). In practice the observed value of $\partial U/\partial \beta_2$ can be used in place of J.

References

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

Begg, C. B. and Leung, D. H.Y. (2000). On the use of surrogate end points in randomized trials (with comments by Day, N. E., Duffy, S. W. and Prentice, R. L.). *Journal of Royal Statistical Society* A **163**, 15–28.

Boscardin, W. J., Taylor, J. M.G. and Law, N. (1998). Longitudinal models for AIDS marker data. *Statistics in Medical Research* **7**, 13–27.

Buyse, M. and Molenberghs, G. (1998). The validation of surrogate endpoints in randomized experiments. *Biometrics* **54**, 1014–1029.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* **1**, 49–67.

Bycott, P. and Taylor, J. (1998). A comparison of smoothing techniques for CD4 data measured with error in a time-dependent Cox proportional hazards model. *Statistics in Medicine* **17**, 2061–2077.

Christensen, E. *et al.* for the Copenhagen Study Group for Liver Diseases (1986). Updating prognosis and therapeutic effect evaluation in cirrhosis with Cox's mulitiple regression model for time-dependent variables. *Journal of Gastroenterology* **21**, 163–174.

Commenges, D. and Andersen, P. K. (1995). Score test of homogeneity for survival data. *Lifetime Data Analysis* **1**, 145–159.

Crowder, M. and Kimber, A. (1997). A score test for the multivariate Burr and other Weibull mixture distributions. *Scandinavian Journal of Statistics* **24**, 419–432.

De Stavola, B. L. and Christensen, E. (1996). Multilevel models for longitudinal variables prognostic for survival. *Lifetime Data Analysis* **2**, 329–347.

Einspahr, J. G., Alberts, D. S., Gapstur, S. M., Bostick, R. M., Emerson, S. S. and Gerner, E. W. (1997). Surrogate end-point biomarkers as measures of colon cancer risk and their use in cancer chemoprevention trials. *Cancer Epidemiology, Biomarkers and Prevention* **6**, 37–48.

Faucett, C. L. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine* **15**, 1663–1686.

FREEDMAN, L. S., GRAUBARD, B. I. AND SCHATZKIN, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine* **11**, 167–178.

HENDERSON, R., DIGGLE, P. AND DOBSON, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–480.

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

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

PRENTICE, R. L. (1989). Surrogate endpoints in clinical trials: are we being misled? *Statistics in Medicine* **8**, 431–440.

SATTEN, G. A. AND LOGINI, I. M. (1996). Markov chains with measurement error: estimating the true course of a marker of the progression of human immunodeficiency virus diesease. *Applied Statistics* **45**, 275–309.

SCHARFSTEIN, D. O., ROTNITZKY, A. AND ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94**, 1096–1146.

SCHEMPER, M. AND HENDERSON, R. (2000). Predictive accuracy and explained variation in Cox regression. *Biometrics* **56**, 249–255.

TSIATIS, A. A., DAFNI, U., DE GRUTTOLA, V., PROPERT, K. J., STRAWDERMAN, R. L. AND WULFSOHN, M. (1991). The relationship of CD4 counts over time to survival in patients with AIDS: is CD4 a good surrogate marker? In Jewell, N., Dietz, K. and Farewell, V. (eds), *AIDS Epidemiology: Methodological Issues*, Boston: Birkhauser, pp. 256–274.

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

[*Received 7 September, 2000; revised 22 January, 2001; accepted for publication 22 January, 2001*]