STATISTICS IN MEDICINE, VOL. 14, 1447–1461 (1995)

A COMPARISON OF FRAILTY MODELS FOR MULTIVARIATE SURVIVAL DATA

ANDREW PICKLES

MRC Child Psychiatry Unit and Department of Biostatistics and Computing, Institute of Psychiatry, DeCrespigny Park, London SE5 8AF, U.K.

AND

ROBERT CROUCHLEY

Department of Statistics and Mathematical Sciences, London School of Economics, Houghton Street, London WC2A 2AE, U.K.

SUMMARY

This paper reviews some of the main approaches to the analysis of multivariate censored survival data. Such data typically have correlated failure times. The correlation can be a consequence of the observational design, for example with clustered sampling and matching, or it can be a focus of interest as in genetic studies, longitudinal studies of recurrent events and other studies involving multiple measurements. We assume that the correlation between the failure or survival times can be accounted for by fixed or random frailty effects. We then compare the performance of conditional and mixture likelihood approaches to estimating models with these frailty effects in censored bivariate survival data. We find that the mixture methods are surprisingly robust to misspecification of the frailty distribution. The paper also contains an illustrative example on the times to onset of chest pain brought on by three endurance exercise tests during a drug treatment trial of heart patients.

1. INTRODUCTION

In many medical studies the sampling of response times may be clustered, for example in a sample of related individuals, in matched subjects, or in studies with repeated measurements or a set of different measures for each individual. Hougaard[1] analysed data of the first kind, in which the survival times of each of 50 treated female rats were compared with those of two female sibs drawn from the same litters.[2] Harrington *et al.*[3] examined matched data, comparing the age of onset of depression in adulthood of subjects who had experienced depression in childhood with those matched on other childhood symptoms. In this paper we consider a repeated measures example using the data shown in Table I.[4] Patients with coronary heart disease pedalled an exercise bike until they experienced angina. Having then been administered an oral dose of isosorbide dinitrate, they were persuaded back on the bike to provide exercise times at 1 hour and 3 hours after drug treatment.

In all of these types of study the response times are commonly censored. In the rat example, in only one experimental sibship did all three rats die during the 104 week study; in 11 of them two rats died, in 15 of them only one died, and 23 sibships survived complete. In the heart disease example, although all 21 patients experienced angina when untreated, a third did not 1 hour after

CCC 0277–6715/95/131447–15

Received June 1992
Revised August 1994

Table I. Exercise times to angina pectoris* (seconds) on three occasions after oral isosorbide dinitrate (mg/kg) (Danahy *et al.*[4])

Time			Dose	Time			Dose
0	1	3		0	1	3	
136	(445)	(393)	0·58	147	403	290	0·44
250	306	206	0·34	231	(540)	370	0·49
215	232	258	0·24	224	432	291	0·31
235	248	298	0·37	152	(733)	492	0·20
129	121	110	0·38	417	(743)	566	0·24
425	580	613	0·32	213	250	150	0·38
441	(504)	(519)	0·41	490	(559)	(557)	0·27
208	264	210	0·37	406	651	624	0·51
154	110	123	0·37	229	327	280	0·24
89	145	172	0·53	265	(565)	(505)	0·51
250	230	264	0·24				

* Observations censored by fatigue in brackets.

treatment and four did not 3 hours after treatment. These exercise times were censored through patients becoming too exhausted to continue.

In this paper we review some of the main approaches to the analysis of correlated survival data. We assume that the correlation between responses occurs because they are dependent upon exogenous causal variables. Sometimes conditioning on an observed set of such variables, typically by their inclusion as covariates within some regression function, can achieve approximate conditional independence. Then, the analysis can proceed along familiar lines using standard univariate methods. More commonly, however, the correlation arises from both observed and unobserved covariates, the latter now being commonly referred to as 'frailties'[5] or 'unobserved heterogeneity'.[6]

Univariate survival models use a mixture likelihood to integrate out the frailty effects. Elbers and Ridder,[7] Ridder[8] and Heckman and Singer[6] study the conditions necessary to achieve identifiability of both parameters of the hazard function and the frailty distribution in univariate data. Elbers and Ridder[7] focus on mixing densities with finite mean, and emphasize the importance of having regressor variables with sufficient variation in order to identify uniquely the functions of interest. Heckman and Singer[6] show how identifiability is maintained for a Weibull hazard function even without regressors provided the frailty distribution has finite mean. If it is assumed that the frailties have infinite mean,[9,1] then the parameters of the hazard function and frailty density cannot be separately estimated in single-spell data. Hougaard[9,10] argued that being able to estimate the frailty distribution from the univariate data is scientifically unreasonable, a point that we shall return to.

The importance of identifiability, and of assumptions concerning the form of the frailty, were emphasized by Heckman and Singer,[6] who appeared to demonstrate high sensitivity of results to alternative choices of finite mean frailty distribution. In a study of the factors affecting the duration of single-spell male unemployment, they compared the estimated parameters for a Weibull baseline hazard with different distributions for the frailty. They noted many changes in sign and magnitude of the estimated covariate parameters. This evidence was used to support their argument for the need for 'non-parametric' methods that made weaker assumptions about the form of distribution. By constrast the work of Lancaster and Nickell,[11] Struthers and Kalbfleisch,[12] Schumacher *et al.*[13] and Lancaster[14] noted that ignoring frailty with finite mean

would, in models of single-spell data with time constant covariates, result in a bias towards zero in the parameter estimates. Allowing for such frailty would therefore only cause the covariate parameter estimates to increase in magnitude, the extent of increase depending on the extent to which the assumed frailty distribution approximates the true frailty distribution. There should be no changes of sign. This suggests that the Heckman and Singer model was misspecified in some other way.[15]

Another consequence of ignoring frailty effects with finite mean is a negative bias in the estimated time dependence, often referred to as spurious duration dependency or cumulative inertia.[11,16]

Multivariate survival data allow the hazard function to be estimated with either finite or infinite mean frailty distributions. We examine the empirical importance of such distributional assumptions for multivariate survival analysis, describing various models, their estimation and comparative performance. Section 2 introduces the notation. We use an example of the three endurance exercise tests as a motivating illustration. Section 3 reviews conditional likelihood methods. Section 4 discusses mixture likelihood methods for models with parametric and non-parametric baseline hazards. Section 5 presents results of a small simulation study that assesses the relative merits of these methods on randomly censored bivariate data. Section 6 extends these considerations to examine their robustness to what we considered a potentially quite extreme form of misspecification, namely the presence of a resilient subpopulation unaffected by the causal risk in question. Section 7 considers generalizations of these different approaches to allow for multiple factor forms of frailty, and Section 8 illustrates their application to the angina example introduced in Section 2. Section 9 concludes the paper.

2. NOTATION

Consider a collection of n sample units i, with m_i measures of survival/failure time t_{ij}, $t_{ij} > 0$ ($i = 1, \ldots, n; j = 1, \ldots, m_i$) and their respective indicator variables d_{ij} (0 = censored, 1 = otherwise). In a clustered sampling design the sampling units might be families, school classes or towns with several subjects drawn from within each unit. For the repeated measures data of Table I, $n = 21$ and $m_i = 3$ for all i.

In general, we consider covariates as falling into two groups. The first group of $p + 1$ covariates, of which the first takes the value 1 to enable a constant to be estimated, have values that are common across measurement occasions j. For sample unit i these will be denoted by the vector x_i. The other group of r covariates, denoted by the vector z_{ij}, have values that vary across both sample units and measurement occasions. For simplicity all the covariates are assumed constant within each measurement occasion/type, but this is not a necessary requirement for the great majority of what follows.

In the absence of frailty effects a typical proportional hazards specification for the hazard associated with each survival time is of the form

$$\lambda_{ij}(t_{ij}) = \lambda_j^0(t_{ij})\, \mu_{ij},$$

where

$$\mu_{ij} = \exp(\eta_{ij}) = \exp(\beta_x x_i + \beta_z z_{ij})$$

and $\lambda_j^0(t_{ij})$ is the baseline hazard for the jth response time of sample unit i. For data from exercise time 0, immediately preceding oral administration of isosorbide dinitrate,

$$\mu_{i0} = \exp(\eta_{i0}) = \exp(\beta_0)\,, \tag{1a}$$

while for the exercise times at 1 and 3 hours we might have

$$\mu_{i1} = \exp(\eta_{i1}) = \exp(\beta_0 + \beta_1 \text{Dose}_i) \tag{1b}$$

$$\mu_{i3} = \exp(\eta_{i3}) = \exp(\beta_0 + \beta_3 \text{Dose}_i) \tag{1c}$$

respectively.

There are several reasons why frailty effects may need to be added to a model such as this. The first and most widely quoted reason occurs when the observed sources of variation in the explanatory variables fail to fully account for the true differences in risk, that is in addition to x and z there are other important but omitted variables present. It is often assumed that the total effect of the omitted variables can be captured by individual specific effects α, which can be discrete or continuous. They can also represent observed but not included important explanatory variables as well as unobserved or unmeasured explanatory variables and may also be correlated with the included covariates.[17]

The second justification for including frailty effects occurs when the model is correctly specified for the true covariates but measurement problems have resulted in error in the observed covariates. In this case the individual specific effects will usually be correlated with the covariates. Measurement error in the observed response times can also be modelled as frailty effects, for example Lancaster's Weibull model[14] with multiplicative error in t, and the Gompertz model with additive error in t.

The third justification for including frailty effects occurs when there is variation in the coefficients of the model between individuals, for example when it is expected that individuals respond differently to the same treatment.

If we have a response specific frailty α_{ij}, then

$$\lambda_{ij}(t_{ij};\alpha) = \lambda_j^0(t_{ij})\exp(\eta_{ij} + \alpha_{ij}) .$$

This gives for individual i, conditional upon the value of the frailty term, a likelihood

$$L_i = \prod_j [\lambda_{ij}(t_{ij};\alpha)]^{d_{ij}}\exp[-\Lambda_{ij}(t_{ij};\alpha)] ,$$

where

$$\Lambda_{ij}(t_{ij};\alpha) = \int_0^{t_{ij}} \lambda_{ij}(s;\alpha)\mathrm{d}s .$$

If $\beta = (\beta_x, \beta_z)$, and the vector of parameters in the baseline hazard is γ, then attempts to directly estimate (β, γ) with the α_{ij} as fixed effects dummy variables fall foul of the 'incidental parameter' problem,[18] and give rise to inconsistent estimates of the hazard function and regression coefficients. The mixture and conditional likelihood methods developed to overcome this problem have for the most part postulated a 'one factor' model of frailty, in which the α_{ij} are constant over j and independent over i (see for example Aalen and Husebye[19]). The correlation between the α_{ij} and the x_i and z_{ij} cannot be recovered for covariates which do not vary within measures in single-measure data.

3. CONDITIONAL LIKELIHOOD ESTIMATION

The simplest conditional approach can be used when the baseline hazard is $\lambda_{ij}^0(t_{ij}) = \lambda^0(t_{ij})$, being constant across sample units i and measures j and when $\alpha_{ij} = \alpha_i$. In this case

$$\lambda_{ij}(t_{ij};\alpha) = \lambda^0(t_{ij})\exp(\eta_{ij} + \alpha_i) .$$

In the absence of censoring, simple conditional estimation is possible for some baseline hazards $\lambda^0(t_{ij})$. But in general, in the presence of censoring the appropriate sufficient statistics for the $\{\alpha_i\}$ make estimation intractable.[20] Rather curiously, censoring does not present any such problem if a sample unit specific non-parametric hazard function is assumed. Forming the standard partial likelihood,[21] but one stratified by sample unit, removes both the baseline hazard and sample unit specific constant to give

$$L_i(\beta_z) = \prod_j \left[\exp(\beta_z z_{ij}) \Big/ \left\{ \sum_{k \in R_i(t_{ij})} \exp(\beta_z z_{ik}) \right\} \right], \tag{2}$$

where $R_i(t_{ij})$ is the usual proportional hazards risk set but defined for the ith sample unit alone. Maximization of equation (2) for β_z is easily undertaken using any proportional hazards program that allows stratification, but of course only β_z and not β_x can be estimated.

For univariate data this partial likelihood is known to possess good efficiency even in the presence of substantial censoring. However, with censored multivariate data the loss of efficiency in using the above conditional estimator may be more severe since the only sample units that contribute to the likelihood are those that can provide at least one survival time that is neither censored nor the longest of the observed durations for that sample unit.

4. MIXTURE LIKELIHOOD ESTIMATION WITH PARAMETRIC AND NON-PARAMETRIC HAZARDS

We can also form a likelihood for t_i, the $1 \times m_i$ vector of responses, that is marginal with respect to α under the assumption that $\alpha_{ij} = \alpha_i$ and the less restrictive assumption that $\lambda^0_{ij}(t_{ij}) = \lambda^0_j(t_{ij})$, under which the baseline hazard varies by response. If the α_i are independent of the included covariates and have probability density $g(\alpha)$ with parameters κ, then estimates of $\theta = (\gamma, \beta, \kappa)$ can be obtained by maximizing

$$L_i = \int \prod_j \lambda_{ij}(t_{ij}; \alpha)^{d_{ij}} \exp\{ - \Lambda_{ij}(t_{ij}; \alpha)\} \, dG(\alpha) . \tag{3}$$

Maximization of equation (3) for θ is made easier if the integral gives a closed form. Some tractability is obtained if we make the substitution $\tau_i = \exp(\alpha_i)$, where τ has probability density $h(\tau)$. If the model contains a constant β_0 there is no loss of generality in assuming that $E(\tau) = 1$. Many researchers have assumed $h(\tau)$ to be the gamma distribution,[11,14,19,22-24] with,

$$h(\tau) = \kappa^\kappa \tau^{\kappa - 1} \exp(- \tau\kappa)/\Gamma(\kappa),$$

which has $\text{var}(\tau) = 1/\kappa$.

If $\Lambda_{ij} = \Lambda^0_j(t_{ij})\mu_{ij}$ and $\lambda_{ij} = \lambda^0_j(t_{ij})\mu_{ij}$, the gamma distribution for τ_i gives a mixture likelihood of the form

$$L_i = \left\{ \prod_{j=1}^{d_i} (\kappa + j - 1) \right\} \kappa^{-d_i.} \left\{ \prod_{j=1}^{m_i} \lambda_{ij}^{d_{ij}} \right\} (1 + \Lambda_{i.}/\kappa)^{-(\kappa + d_i.)} , \tag{4}$$

where $d_i. = \Sigma_j d_{ij}$ and $\Lambda_{i.} = \Sigma_j \Lambda_{ij}$.

The gamma distribution is a member of a family of distributions suggested by Hougaard,[1] which also includes the inverse Gaussian and models in which $E(\tau) = \infty$, such as the positive stable law.[1,10] A normal distribution for $G(\alpha)$ in equation (3) does not lead to an analytically

tractable integral, but numerical integration by Gaussian quadrature, in which the α_i are replaced by a set of masses of known weight and location,[25] has been used. Increasing use of estimation via Gibbs sampling[26] or other stochastic integration method is likely to make the normal model more popular.

Bock and Aitkin,[26] in the context of binary item response modelling, showed how the Gaussian quadrature approach could be extended to estimate the mass of each quadrature point, rather than treating each as known. Heckman and Singer[6] extended this further for application to parametric survival data by showing how both the weights and the locations could be estimated. Theoretical results[27-30] showed how under apparently quite weak conditions the discrete distribution formed by such a finite set of free points of mass corresponded to the non-parametric maximum likelihood (NPML) representation of any finite mean continuous frailty distribution. Numerous empirical applications have subsequently shown how the number of mass points required to obtain the non-parametric representation is usually quite small, rarely more than ten and often as few as three or four. Estimation can follow a generalization of the procedures described above or can be undertaken in more specialized programs, for example MIXTURE.[31]

If the τ_i were known fixed effects, inference about θ could be based on the log-likelihood constructed from the joint distribution of t_{ij} and τ, namely

$$l_n = \sum_i \sum_j [d_{ij} \log\{\tau_i \mu_{ij} \lambda_j^0(t_{ij})\} - \tau_i \mu_{ij} \Lambda_j^0(t_{ij})] + \sum_i \log h(\tau_i) ,$$

and this could be maximized more straightforwardly using standard survival analysis software. Of course, the τ_i are not known but an initial guess at θ, θ_m, can be used to calculate the expected value of l_n with respect to τ given t. The expected log-likelihood, $Q(\theta, \theta_m) = E[l_n; \tau \mid t, \theta_m]$, can be maximized with respect to θ, and the solution used to replace the initial guess θ_m in a recalculation of $Q(\theta, \theta_m)$. This sequence of expectation and maximization steps is an example of an EM algorithm, and forms the basis of the GLIM[32] macro for the gamma-Weibull model of equation (4) in Clayton.[33]

The EM approach can also be applied to the multivariate form of Cox's partial likelihood. Gill[34] presents some results for a gamma frailty piece-wise exponential model, illustrating some of the links between the work of Clayton and Cuzick[23] and that of Self and Prentice.[35] Self and Prentice[35] derive a model with hazards $\lambda^0(t_{ij})\mu_{ij} E[\tau \mid t_{hi}, \theta]$, in which the expected values of the frailties are conditional on the subset (t_{hi}) of t_i which have failed by time t_{ij}. (Unlike EM the expectations are not based on the final posterior density of τ.) The density of τ changes with time and as a result the estimate of $E[\tau \mid t_{hi}, \theta]$, given by $[\kappa + \Sigma_j d_{ij}(t_{hi})]/[\kappa + \Sigma_j \Lambda_j(t_{hi})\mu_{ij}]$ for gamma frailty, varies as information about t_i accumulates, even though the true value is assumed constant over time. In addition Self and Prentice[35] suggest an approximation to $E[\tau \mid t_i, \theta]$ which allows the maximization step to be performed by standard partial likelihood software by the inclusion of the term $\Sigma_j[d_{ij}(t_{hi}) - \Lambda_j(t_{hi})\mu_{ij}]$ as a time varying covariate with parameter κ. The estimation process is cycled until the values of θ remain unchanged. Models and EM algorithms based on counting process derivations have also been proposed.[36,37]

The next section presents the main results from a small simulation study that assesses the relative merits of the alternative models on a particular set of randomly censored bivariate data.

5. MONTE CARLO COMPARISON OF CONDITIONAL, GAMMA, SELF AND PRENTICE, NPML AND STABLE LAW MODELS

Before proceeding to consider more complex models and empirical examples an exploratory simulation study of the performance of the basic forms of these models was undertaken. While the

Table II. Monte Carlo for samples of 100 bivariate survival times with log-gamma frailty, 60 per cent random censoring and pairs 50 per cent discordant for a dummy variable: parameter sample means and standard deviations ($n = 100$, $m = 2$, 100 replicate samples)

Model	Beta	Constant	Weibull
True values	$\beta_1 = 1$	$\beta_0 = 0$	$\gamma = 2$
Conditional	1·15 (0·27)		
Gamma	1·02 (0·11)	0·06 (0·05)	2·04 (0·05)
Self and Prentice	1·02 (0·07)	− 0·06 (0·06)	2·03 (0·06)
Normal compound	1·04 (0·12)	− 0·48 (0·06)	2·09 (0·07)
Mass point	1·10 (0·18)	− 1·15 (0·92)	2·18 (0·08)
Stable law	0·96 (0·10)	− 0·62 (0·07)	1·95 (0·07)

simulations may not be informative about the properties of the models in contexts different from those assumed, they may help to shed some light on the differences between the models in a particular context. Samples of 100 pairs, subscripted i, of bivariate Weibull distributed survival times were generated with rate parameters $\tau_i \exp(\beta x_{ij})$ and a common shape parameter. The values of τ_i were drawn from a gamma distribution with mean and variance 1, the β coefficient was set equal to 1 and the shape parameter equal to 2. The exogenous covariates $\{x_{i1}, x_{i2}\}$ were time constant dummy variables with values $\{0, 0\}$ for 50 pairs and $\{0, 1\}$ for the remaining 50. The response times were then subject to approximately 60 per cent random censoring with censoring times generated from a unit exponential distribution.

Although, as explained above, several of the models examined could have been estimated using modifications of existing software, for this exercise programs were written in FORTRAN. The conditional, gamma, normal (using ten-point Gaussian quadrature) and stable law models were all fitted by direct maximization of the likelihood function. The model of Self and Prentice[35] was estimated by maximizing the partial likelihood that included a time varying multiplicative term $[\kappa + \Sigma_j d_{ij}(t_{hi})]/[\kappa + \Sigma_j \Lambda_j(t_{hi})\mu_{ij}]$. The integrated hazard was estimated using Breslow's method.[38,39] The mass point model was estimated using direct likelihood maximization, with an additional routine to examine solutions to determine if and where an additional mass was required. Results on parameter estimation are shown in Table II. We concern ourselves primarily with the estimation of the regression coefficient β.

The poor performance of the conditional estimator is due to the large amount of censoring in these samples. The normal frailty model gave results close to that of the fully parametric and correct gamma model, unsurprising in view of the similarity of the log-normal distribution to the gamma distribution. As indicated earlier, the number of masses required to achieve the ML non-parametric representation for the frailty distribution was not large, on average only four masses being required. However, the mass point model showed substantially larger sampling variance and quite marked bias in the estimate of the constant, calculated in this model as the mean location of the points weighted by their mass. Closer inspection of the fitted models

Table III. Monte Carlo results for LR test statistic ($n = 100$, $m = 2$, 100 replicate samples, 50 per cent discordant, 60 per cent random censoring, log-gamma frailty)

	Quantiles of distribution						
χ^2_1	0·065	0·46	1·64	2·71	3·84	5·41	6·63
p	0·20	0·50	0·80	0·90	0·95	0·98	0·99
Conditional	0·15	0·57	0·83	0·91	0·97	0·99	1·00
Gamma	0·21	0·48	0·79	0·91	0·97	0·99	1·00
Self and Prentice	0·24	0·50	0·85	0·93	0·95	0·97	0·99
Stable law	0·24	0·46	0·78	0·95	0·97	0·98	0·99
Normal	0·23	0·48	0·83	0·92	0·95	0·98	1·00
Mass point	0·16	0·38	0·74	0·84	0·91	0·93	0·96

suggested that this arose from a sporadic tendency to represent censored observations by individuals with very low estimated frailty, these being represented by a mass point with a very large negative location. The stable law model, in spite of departing from the true model in important theoretical respects, performs surprisingly well for the two parameters of main interest (β_1 and γ), although some evidence for compensating effects for its known distributional differences is shown in the negative bias in the constant.

Table III shows the distribution of the LR test statistic for β_1, the regression coefficient for the dummy variable, equal to its true value. All except the mass point model performed reasonably well.

6. THE PRESENCE OF A NOT-AT-RISK SUBPOPULATION

The form of frailty considered in the previous sections has excluded the possible occurrence of a subpopulation who are entirely robust, immune or otherwise simply not at risk from the cause in question. Farewell[40] suggests that such individuals might arise in some treatment trials in which complete 'cure' is possible, but otherwise uses the more neutral terminology of 'long-term survivors'. Heckman and Walker[41] consider a sterile subgroup within an analysis of the time to first birth. In state transition processes such individuals have been commonly referred to as 'stayers'.[42]

In the presence of such a subpopulation the conditional approach can be used without modification. In practice this should also be true of the mass point approach, since such a subpopulation can be represented by the occurrence of a mass located on the far left of the distribution corresponding to a negligible transition intensity, though computational convergence is quicker if explicit allowance is made for a subgroup with zero intensity.[43] Formally, none of the other estimated models account for such stayers. They can be extended in finite mixture fashion to allow a zero-intensity spike to the distribution or by using the more general Hougaard family[1] along the lines of Aalen.[44]

The upper half of Table IV shows the results of some further simulations similar to those of Table II, but with x_{i1} and x_{i2} discordant in value for all pairs. The performance of the conditional estimator, in particular, should increase as the proportion of discordant pairs is increased. This is borne out by a comparison of the top of Table IV with Table II. In the lower half of Table IV a random 33 per cent of the pairs of observations were 'stayers' (or 'cured'), and were assigned

Table IV. Monte Carlo for samples of 100 bivariate survival times with log-gamma frailty, 60 per cent random censoring pairs 100 per cent discordant for a dummy variable and 33 per cent 'stayers': parameter sample means and standard deviations ($n = 100$, $m = 2$, 100 replicate samples)

Model	Beta	Constant	Weibull
True values	$\beta_1 = 1$	$\beta_0 = 0$	$\gamma = 2$
Conditional	1·05 (0·17)		
Gamma	1·02 (0·08)	0·06 (0·06)	2·05 (0·05)
Stable law	0·96 (0·09)	− 0·63 (0·07)	1·96 (0·07)
Mass point	1·10 (0·12)	− 1·05 (0·86)	2·19 (0·09)
With 33 per cent stayers			
Conditional	1·04 (0·29)		
Gamma	1·09 (0·19)	0·20 (0·16)	2·26 (0·12)
Stable law	0·88 (0·11)	− 4·58 (0·20)	1·92 (0·07)
Mass point	1·08 (0·19)	—	2·21 (0·16)

a zero hazard rate. The models formally more capable of incorporating stayers show lower bias in the estimates of β, with both the conditional and the mass point models actually showing, in these samples, lower bias in the presence of stayers than in their absence. Also as expected, the bias shown by the gamma model, the true model in the upper but not the lower half of the table, is not as great as that shown by the stable law. Even so, in all four models the bias is not large in comparison with the sampling variance of the estimates.

Table V shows the distribution of the LR test statistic for the regression coefficient β. Again the mass point model was the worst of the four models tested, even in the presence of 'stayers' for which it might have been expected to have some relative advantage. By contrast, the stable law model that would have been expected to have some difficulty with 'stayers' gave a test statistic closest in distribution to the nominal chi-square.

7. GENERALIZATIONS TO MORE COMPLEX COVARIANCE STRUCTURES

The conditional maximum likelihood estimator is not easily extended to more complex covariance structures such as might be present in multilevel data. This is because of the difficulty in constructing an appropriate sufficient statistic for the sample unit frailty.

A marginal likelihood model with a normal frailty term in the linear predictor could be generalized along the same lines as normal theory linear variance components models, to give

Table V. Monte Carlo results for LR test statistic ($n = 100$, $m = 2$, 100 replicate samples, 100 per cent discordant, log-gamma frailty, 60 per cent random censoring, 33 per cent 'stayers')

	Quantiles of distribution						
χ^2_1	0·0647	0·455	1·642	2·706	3·841	5·412	6·633
p	0·20	0·50	0·80	0·90	0·95	0·98	0·99
No stayers							
Conditional	0·29	0·54	0·84	0·91	0·94	0·94	0·95
Gamma	0·23	0·45	0·83	0·92	0·93	0·96	0·99
Stable law	0·26	0·54	0·80	0·89	0·93	0·96	0·98
Mass point	0·19	0·51	0·74	0·88	0·92	0·92	0·96
Approximately 33 per cent stayers							
Conditional	0·21	0·48	0·75	0·87	0·93	0·97	0·99
Gamma	0·08	0·38	0·74	0·86	0·91	0·95	0·98
Stable law	0·20	0·48	0·75	0·87	0·94	0·98	0·99
Mass point	0·12	0·38	0·72	0·87	0·90	0·93	0·96

multilevel models.[45] Random coefficients as well as random constants would be possible, and several components organized in hierarchical or non-hierarchical fashion and with possible correlations between them. A simple extension that maintains the univariate error structure of the basic model is to allow different random error variances from measure to measure, equivalent to allowing different scale factors. This would appear essential where the observations represent quite different measures rather than being replications of the same measurement.

Interest in models using the multivariate extreme value (MEV) distribution has recently been renewed in the work of Crowder,[24] Hougaard,[9,1] McFadden[46] and Tawn.[47] McFadden[46] applied it to consumer choice over categorical alternatives, a distribution equivalent, in our context, to the following two level nested form:

$$\Pr[t > y] = \exp\left\{ -\left(\left[\sum_{j \in S_1} \Lambda_{ij} \right]^{\kappa_2} + \left[\sum_{j \in S_2} \Lambda_{ij} \right]^{\kappa_3} \right)^{\kappa_1} \right\},$$

where the m measurements are exhaustively allocated to the mutually exclusive sets S_1 and S_2, and $0 < \kappa_j \leqslant 1$. Hougaard[9] showed that this structure arises from a model of hierarchically organized independent stable law random components, each acting multiplicatively. The correlation of measurements between sets arises through one shared component and is determined by κ_1. Measures within set 1 share this component and a second component, giving a correlation determined by both κ_1 and κ_2. The correlation within set 2 is correspondingly determined by κ_1 and κ_3.

A rather different generalization to the notion of frailty can be obtained by the use of a class of MEV distribution introduced by Tawn.[47] This structure arises from a model of hierarchically organized additive stable law components. Alternatively, the sample units may be considered as made up of a mixture of those that exhibit a dependency among measurements and those for whom they are independent. In the simple case of one independent component and one

dependent component this is given by

$$\Pr[t > y|\tau] = \exp\left[- \sum_{j=1}^{m} (1 - \theta_j)\Lambda_{ij} - \tau \sum_{j=1}^{m} (\theta_j \Lambda_{ij})^{1/\kappa} \right],$$

with the joint marginal distribution

$$\Pr[t > y] = \exp\left[- \sum_{j=1}^{m} (1 - \theta_j)\Lambda_{ij} - \left\{ \sum_{j=1}^{m} (\theta_j \Lambda_{ij})^{1/\kappa} \right\}^{\kappa} \right].$$

Such models would be valuable where co-morbidity in the form of the onsets of two conditions, such as childhood depression and anti-social behaviour, may represent a disorder distinct from either one alone.[48]

The inclusion of random coefficients within either of the hierarchical or Tawn classes of MEV is possible by making κ a function of covariates.

8. ILLUSTRATIVE EXAMPLE: EXERCISE TIMES

Table VI presents the results from model fitting to the data of Table I with β_1 and β_3, as shown in equations (1a)–(1c), measuring the effects of the administered dose after 1 and 3 hours respectively. The dose administered to each patient had been previously determined as the level above which side-effects occurred. All models fitted by marginal likelihood assumed Weibull baseline hazards with a single shape parameter γ common across exercise occasions.

The basic models allowed for a single component random effect or one-factor frailty. The mixture likelihood models gave estimates of the β coefficients that were close to those from the conditional model. In this example differences might have arisen either through misspecification of the hazard or through the frailty being correlated with the included covariate, dose. As one might have expected, these data possess a slight negative correlation between the initial exercise times and dose, those subjects showing more severe initial incapacitation being administered larger doses. The conditional model makes no assumption as to the correlation of covariates with frailty. The mixture likelihood models would have to be extended to allow for such a correlation.[49,50]

As well as variation in exercise times as the result of an overall 'frailty' component of variance τ_0, which is common to all the responses, it might be reasonable to expect variation in response to the drug. This would require a second frailty effect τ_2 or component of variance which is only present in the last two responses. In this case the hazard for time 0 takes the form

$$\lambda_{i0}(t_{i0}; \tau_0) = \lambda^0(t_{i0})\tau_0 \exp(\eta_{i0}),$$

and

$$\lambda_{i1}(t_{i1}; \tau_0, \tau_2) = \lambda^0(t_{i1})\tau_0^{1/\kappa_2}\tau_2 \exp(\eta_{i1}),$$

$$\lambda_{i3}(t_{i3}; \tau_0, \tau_2) = \lambda^0(t_{i3})\tau_0^{1/\kappa_2}\tau_2 \exp(\eta_{i3}),$$

for times 1 and 3 respectively. If we let Λ_{i0}, Λ_{i1} and Λ_{i3} be the integrated hazards for time 0, 1 and 3, then the two-level MEV model of Hougaard[9] that is appropriate in this context is

$$\Pr[t > y] = \exp\{ - [\Lambda_{i0} - (\Lambda_{i1} + \Lambda_{i3})^{\kappa_2}]^{\kappa_0} \},$$

Table VI. Analysis of angina pectoris data: parameter estimates and LR tests (β_1, β_3 and κ_2 against 0, γ against 1)

Parameter	Conditional model	Normal compound 1-level	Normal compound 2-level	Hierarchical MEV 1-level	Hierarchical MEV 2-level
β_0		− 5·41	− 7·23	− 5·21	− 7·36
β_1	− 5·10 (15.40)	− 6·05 (24.16)	− 11.12 (15.09)	− 5·61 (23.87)	− 9·22 (20.81)
β_3	− 3.56 (8.54)	− 3.35 (17.22)	− 8.49 (6.98)	− 3.49 (11.46)	− 6.30 (12.81)
γ		4.60 (74.37)	6.64 (79.82)	4.59 (12.82)	6.73 (13.69)
κ_0		3.40	3.22	2.36	1.99
κ_2			3.07 (5.85)		3.41 (6.33)
Deviance	53.8	168.2	162.3	− 175.7	169.4

which can also be written in terms of the marginal survivor functions, that is $\Pr[t_{ij} > y_{ij}] = [\exp(-\Lambda^*_{ij})]$. In this case

$$\Pr[t > y] = \exp\{-\Lambda_{i0}^{*1/\kappa_0} - (\Lambda_{i1}^{*1/\kappa_2\kappa_0} + \Lambda_{i3}^{*1/\kappa_2\kappa_0})^{\kappa_2}]^{\kappa_0}\} .$$

An equivalent two-factor frailty model with normal components is also easily constructed. Of course, the marginal survivor function for this normal frailty model does not have a simple form.

The three one-factor models all gave similar parameter estimates, with the LR chi-squares being larger for the normal frailty model than the MEV model, and for both marginal likelihood models being larger than those from the conditional model. The two two-factor mixture likelihood models both gave a significant improvement in fit over their one-factor versions. The parameter estimates from these two models were again very similar but were roughly twice the size of those obtained from the one-factor models. Allowing for the correlation in drug response at one and three hours after administration reduced the significance of the parameters estimating drug effects, the MEV model now giving more significant effects than the normal frailty model.

9. DISCUSSION

Our simulation study of models for multivariate survival data with univariate or one-factor frailty suggests that with multiple outcome measurements the choice of particular parametric frailty distribution is not critical for the estimation and testing of regression type coefficients. The tractable models that assume log-gamma or normal frailty performed well even in the presence of a substantial not-at-risk subgroup or stayers. Somewhat surprisingly, the mass point method, which had been considered to offer particular robustness, in fact did not perform well with the heavily censored data examined here. For data where univariate frailty is thought to be the dominant source of correlation in response times, computational convenience and the generality of the baseline hazard would seem to be more important criteria in a choice of model than the generality of the frailty distribution.

Although we have not undertaken any formal study, the example suggested that a more important choice that the parametric form of frailty is its dimensionality. This choice will also be fundamental where the frailty is not just a nuisance but of prime interest, as for example in genetic studies. As in the example, increasing the dimensionality allows different outcomes to be correlated with each other to differing degrees. It can also allow the conditions for non-identifiability, where scientifically necessary, to be met while still using finite mean frailty distributions. As discussed, finite mean frailty distributions can be identified from single-response data. Thus, to estimate just the frailty common across a set of responses, mixture likelihood models based on finite mean distributions must include at least bivariate frailty, allowing the estimation of response specific frailty as well as that common across responses. The addition of such complexity to finite mean frailty models makes the unfamiliar stable-law/MEV models look relatively more attractive.

This review has not included approaches in which the frailty effects are of secondary importance to the analysis of multivariate survival data, for example the marginal model approaches of Huster *et al.*[51] and Liang *et al.*[52] Clayton[53] contrasts random effects and marginal model approaches.

REFERENCES

1. Hougaard, P. 'A class of multivariate failure time distributions', *Biometrika*, **73**, 671–678 (1986).
2. Mantel, N., Bohidar, N. R. and Ciminera, J. L. 'Mantel–Haenszel analyses of litter-matched time-to-response data, with modifications for recovery of interlitter information', *Cancer Research*, **37**, 3863–3868
3. Harrington, R. C., Fudge, H., Rutter, M., Pickles, A., and Hill, J. 'Adult outcomes of childhood and adolescent depression. I: Psychiatric status', *Archives of General Psychiatry*, **47**, 465–473 (1990).
4. Danahy, D. J., Burwell, D. T., Aranow, W. S. and Prakash, R. 'Sustained henodynamic and anti-anginal effect of high dose oral isosorbide dinitrate', *Circulation*, **55**, 381–387 (1977).
5. Vaupel, J. W., Manton, K. G. and Stallard, E. 'The impact of heterogeneity in individual frailty on the dynamics of mortality', *Demography*, **16**, 439–454 (1979).
6. Heckman, J. J. and Singer, B. 'A method for minimizing the impact of distributional assumptions in econometric models of duration data', *Econometrica*, **52**, 271–320 (1984).
7. Elbers, C. and Ridder, G. 'True and spurious duration dependence: the identifiability of the proportional hazards model', *Review of Economic Studies*, **49**, 403–409 (1982).
8. Ridder, G. 'The non-parametric identification of generalized accelerated failure time models', *Review of Economic Studies*, **57**, 167–182 (1990).
9. Hougaard, P. 'Survival models for heterogeneous populations derived from stable distributions', *Biometrika*, **73**, 387–396 (1986).
10. Hougaard, P. 'Modelling multivariate survival', *Scandinavian Journal of Statistics*, **14**, 291–304 (1987).
11. Lancaster, T. and Nickell, S. 'The analysis of re-employment probabilities for the unemployed', *Journal of the Royal Statistical Society, Series A*, **143**, 141–165 (1980).
12. Struthers, C. A. and Kalbfleisch, J. D. 'Misspecified proportional hazards models', *Biometrika*, **73**, 363–369 (1986).
13. Schumacher, M., Olschewski, M. and Schmoor, C. 'The impact of heterogeneity on the comparison of survival times', *Statistics in Medicine*, **6**, 773–784 (1987).
14. Lancaster, T. *The Econometric Analysis of Transition Data*, CUP, Cambridge, 1990.
15. Manton, G. K. 'Contribution to discussion of "Natural variability of vital rates" by D. R. Brillinger', *Biometrics*, **42**, 693–734 (1986).
16. Yashin, A. I., Manton, K. G. and Vaupel, J. W. 'Mortality and aging in a heterogeneous population: a stochastic process model with observed and unobserved variables', *Theoretical Population Biology*, **27**, 154–175 (1985).
17. Crouchley, R. and Pickles, A. R. 'An empirical comparison of conditional and marginal likelihood methods in a longitudinal study', in Clogg, C. C. (ed.), *Sociological Methodology 1989*, Leinhardt, 1989.
18. Neyman, J. and Scott, E. 'Consistent estimates based on partially consistent observations', *Econometrica*, **16**, 1–32 (1948).

19. Aalen, O. O. and Husebye, E. 'Statistical analysis of repeated events forming renewal processes', *Statistics in Medicine*, **10**, 1127–1240 (1991).
20. Holt, T. D. and Prentice, R. L. 'Survival analysis in twin studies and matched pairs experiments', *Biometrika*, **61** 17–30 (1974).
21. Cox, D. R. 'Regression models and life-tables (with discussion)', *Journal of the Royal Statistical Society, Series B*, **34**, 187–220 (1972).
22. Clayton, D. G. 'A model for association in bivariate life-tables and its application in epidemiological studies of chronic disease incidence', *Biometrika*, **65**, 141–151 (1978).
23. Clayton, D. G. and Cuzick, J. 'Multivariate generalizations of the proportional hazards model (with discussion)', *Journal of the Royal Statistical Society, Series A*, **148**, 82–117 (1985).
24. Crowder, M. 'A distributional model for repeated failure time measurements', *Journal of the Royal Statistical Society, Series B*, **47**, 447–452 (1985).
25. Clayton, D. G. 'A Monte-Carlo method for Bayesian inference in frailty models', *Biometrics*, **47**, 467–485 (1991).
26. Bock, R. D. and Aitkin, M. 'Marginal maximum likelihood estimation of item parameters: application of an EM algorithm', *Psychometrika*, **46**, 443–459 (1981).
27. Simar, L. 'Maximum likelihood estimation of a compound Poisson process', *Annals of Statistics*, **4**, 1200–1209 (1976).
28. Laird, N. 'Non-parametric maximum likelihood estimation of a mixing distribution', *Journal of the American Statistical Association*, **73**, 805–811 (1978).
29. Lindsay, B. G. 'The geometry of mixture likelihoods: a general theory', *Annals of Statistics*, **11**, 86–94 (1983).
30. Lindsay, B. G. 'The geometry of mixture likelihoods. Part II: The exponential family', *Annals of Statistics*, **11**, 783–792 (1983).
31. Ezzet, F. L. and Davies, R. B. *A Manual for MIXTURE*, Centre for Applied Statistics, University of Lancaster, 1987.
32. Baker, R. J. and Nelder, J. A. *The GLIM System: Release 3. Generalized Linear Interactive Modelling*, Numerical Algorithms Group, Oxford, 1978.
33. Clayton, D. G. 'The analysis of event history data: a review of progress and outstanding problems', *Statistics in Medicine*, **7**, 819–841 (1988).
34. Gill, R. D. 'Discussion of paper presented by Mr. Clayton and Dr. Cuzick', *Journal of the Royal Statistical Society, Series A*, **148**, 108–109 (1985).
35. Self, S. G. and Prentice, R. L. 'Incorporating random effects into multivariate relative risk regression models', in Moolgavkar, S. H. and Prentice, R. L. (eds.), *Modern Statistical Methods in Chronic Disease Epidemiology*, Wiley, New York, 1986, pp. 167–178.
36. Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. *Statistical Models Based on Counting Processes*, Springer, New York, 1992.
37. Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sorenson, T. I. A. 'A counting process approach to maximum likelihood estimation in frailty models', *Scandinavian Journal of Statistics*, **19**, 25–43 (1992).
38. Breslow, N. 'Covariance analysis of censored survival data', *Biometrics*, **30**, 880–899 (1974).
39. Breslow, N. 'Contribution to discussion of paper by D. R. Cox', *Journal of the Royal Statistical Society, Series B*, **34**, 216–217 (1972).
40. Farewell, V. T. 'The use of mixtures for the analysis of survival data with long-term survivors', *Biometrics*, **38**, 1041–1046 (1982).
41. Heckman, J. J. and Walker, J. R. 'Estimating fecundability from data on waiting time to first conception', *Journal of the American Statistical Association*, **85**, 283–294 (1990).
42. Blumen, I., Kogan, M. and McCarthy, P. J. *The Industrial Mobility of Labour as a Probability Process*, Cornell Studies in Industrial and Labor Relations, Vol. 6, 1955.
43. Davies, R. B. and Ezzet, F. L. 'Software for the statistical modelling of recurrent behaviour', in Uncles, M. D. (ed.), *Longitudinal Data Analysis: Models and Applications*, Pion, London, 1987, pp. 103–115.
44. Aalen, O. O. 'Heterogeneity in survival analysis', *Statistics in Medicine*, **7**, 1121–1137 (1988).
45. Goldstein, H. *Multilevel Models in Education and Social Research*, Griffin, London and OUP, New York, 1987.
46. McFadden, D. 'Econometric models of probabilistic choice', in Manski, C. F. and McFadden, D. (eds.), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, MA, 1978.
47. Tawn, J. A. 'Bivariate extreme value theory: models and estimation', *Biometrika*, **75**, 397–416 (1988).

48. Kovacs, M., Paulauskas, S., Gatsonis, C. and Richards, C. 'Depressive disorders in childhood. III: A longitudinal study of comorbidity with and risk for conduct disorders', *Journal of Affective Disorders*, **15**, 205–217 (1988).
49. Chamberlain, G. 'Heterogeneity, omitted variable bias and duration dependence', in Heckman, J. J. and Singer, B. (eds.), *Longitudinal Analysis of Labor Market Data*, CUP, Cambridge, 1985, pp. 3–38.
50. Brillinger, D. B. and Preisler, H. J. 'Two examples of quantal data analysis: (a) multivariate point process, (b) pure death process in an experimental design', *Proceedings of the 13th International Biometrics Conference*, Seattle, WA, 1986.
51. Huster, W. J., Brookmeyer, R. and Self, S. G. 'Modelling paired survival data with covariates', *Biometrics*, **45**, 145–146 (1989).
52. Liang, K.-Y., Self, S. G. and Chang, Y.-C. 'Modelling marginal hazards in multivariate failure time data', *Journal of the Royal Statistical Society, Series B*, **55**, 441–454 (1993).
53. Clayton, D. G. 'Some approaches to the analysis of recurrent event data', *Statistical Methods in Medical Research* **3**, 244–262 (1994).