STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:2547–2562 (DOI: 10.1002/sim.1123)

Multivariate survival analysis with doubly-censored data: application to the assessment of Accutane treatment for fibrodysplasia ossificans progressiva

Geoffrey Jones[1] and David M. Rocke[2,*,†]

[1]*Department of Statistics, Massey University, Palmerston North, New Zealand*
[2]*Department of Applied Science, University of California, Davis CA 95616, U.S.A.*

SUMMARY

Fibrodysplasia ossificans progressiva is a rare genetic disorder in which the joints of patients become disabled by the formation of heterotopic bone. Data are available on the status of 11 joints of each of 21 patients before, during and after treatment with Accutane. These are compared with data obtained by questionnaire from 40 untreated patients to determine the efficacy of the treatment. Both left- and right-censoring are present in each group, which, together with the multivariate nature of the data and the time-dependent treatment covariate, makes analysis difficult. We consider two alternative parametric models for incorporating within-subject dependence: a marginal model and a frailty model. Both analyses suggest that Accutane treatment is effective. We discuss and illustrate the differences between the two approaches. We also discuss the extent to which the conclusions are compromised by the observational nature of the study. Copyright © 2002 John Wiley & Sons, Ltd.

1. INTRODUCTION

Fibrodysplasia ossificans progressiva (FOP) is an extremely rare and disabling genetic disorder characterized by progressive heterotopic ossification of soft tissues, leading to the immobilization of affected joints. Bone formation can be stimulated by blunt trauma, surgery, intramuscular injection or aggressive physical therapy, but most often occurs spontaneously. At present no effective prevention or treatment is known [1].

In 1992, 44 patient members of the International Fibrodysplasia Ossificans Progressiva Association responded to a postal survey of the age at onset of heterotopic ossification at each of 15 anatomic sites [2]. For each patient in the survey, and for each anatomic site, the patient was asked to record the date at onset of heterotopic ossification. Right-censoring occurred when a particular joint was uninvolved at the time of the survey: left-censoring

*Correspondence to: David M. Rocke, Center for Image Processing and Integrated Computing, Department of Applied Science, One Shields Avenue, University of California, Davis, CA 95616-8553, U.S.A.
†E-mail: dmrocke@ucdavis.edu

Contract/grant sponsor: NSF; contract/grant numbers: AC1 96-19020, DMS 98-70172, DMS 95-10511

Received September 1999
Accepted August 2001

Copyright © 2002 John Wiley & Sons, Ltd.

when the patient replied that a joint was already involved but they could not provide the date of onset. From the 660 onset times in the survey, 231 were right-censored and 41 left-censored.

Retinoids are a plausible family of therapeutic agents for FOP because of their ability to inhibit differentiation of mesenchymal tissue into cartilage and bone. An age-matched internal control study was designed to determine the effectiveness of 13-cis-retinoic acid (Accutane) in the prevention of new heterotopic lesions in patients. After nearly one year of attempted recruitment, it proved impossible to assemble an internal age- and disease-severity-matched control group. Most patients over 21 had severe symptoms and did not wish to receive Accutane, whereas most under 21 had more mild manifestations of FOP and were unwilling to receive a placebo. Because of this, and the extreme rarity of the condition, it was decided to recruit FOP patients into a treatment group only and to use the survey data as an external control.

Twenty-one patients who had FOP were recruited sequentially during an initial or follow-up visit to the FOP clinic at the National Institute of Health. Eleven anatomic regions were assessed in each of the 21 patients by clinical examination, plain roentgenograms and radionuclide bone scans. An anatomic region was considered to be involved if there was clinical, roentgenographic or radionuclide evidence of orthotopic or heterotopic ossification anywhere in that region. The regions thus assessed in the treatment group were the neck, spine, jaw, right shoulder, left shoulder, right elbow, left elbow, right hip, left hip, right knee and left knee; these regions were also included in the survey data from the external control, along with right/left wrist and ankle. The duration of treatment ranged from four months to ten years, with median three years and quartiles one and five years. The total number of uninvolved joints prior to treatment was 88; only two of these became involved during treatment, both belonging to the same patient.

There are a number of problems in assessing the effectiveness of the treatment. The lack of an internal control group may lead to bias, especially since the selection criterion and the criterion for diagnosis of involvement of an anatomic region is different in each group. We postpone discussion of this point until Section 5, where possible group differences are analysed. Our main focus of consideration is the multivariate nature of the data. The total number of patients is too small for the effect of treatment on any single uninvolved joint to be estimated with any precision, so it is necessary to examine all anatomic regions simultaneously and to somehow account for the within-patient dependence. Thus the data are highly multivariate, with 11 observations on each subject corresponding to the status of each of 11 anatomical sites. The presence of both left- and right-censoring, and a time-dependent covariate, increase the technical difficulty of analysis.

There is now a large literature on multivariate survival analysis. The marginal model approach of Wei *et al.* [3] fits univariate models to the marginal distributions and uses Huber's sandwich estimator [4] to make joint inference about the regression parameters from each model, with no attempt made to model the within-patient dependence. The frailty model approach, on the other hand, attempts a parametric model of this dependence by postulating a patient-specific parameter, a 'frailty', to account for each patient's particular susceptibility to new events [5–8]. It is usual in both approaches to assume proportional hazards and to model the baseline hazard functions non-parametrically, as in the Cox model, although the interpretation of proportionality varies, applying to the marginals in the first approach and the conditional distributions given the frailty in the second. Parametric models for the hazard

Copyright © 2002 John Wiley & Sons, Ltd.

functions are also possible. Huster *et al.* [9] explicitly compared the two approaches in a parametric analysis of paired survival data, finding that the marginal approach was computationally simpler but less efficient.

The commonly used partial likelihood approach cannot be used for our data because of the double censoring. Several authors [10–12] have examined the estimation of semi-parametric proportional hazard models with doubly-censored or interval-censored data, but time-dependent covariates seem to present a special challenge in this context. Recent work by Goggins *et al.* [13] applies the Cox model for an interval-censored time-dependent covariate to data which is exact or right-censored. We have attempted to solve the combination of difficulties in the FOP data by using parametric versions of both a marginal model and a frailty model, which we then compare. The advantage of a parametric approach here is the relative ease of handling of different censoring patterns. Although our data contained no interval-censoring the methods are easily extended to cover such cases.

In applications the multiple events are either homogeneous (same hazard structure) or heterogeneous. Our data set arguably has both types: some joints, such as the neck and jaw, clearly have different baseline hazards whereas pairs such as right knee/left knee might be expected to be the same. An additional complication is the small number of patient-years on treatment, which together with the few observed failures during treatment makes a straightforward application of the robust variance in the marginal model invalid. We propose a transformation which improves the performance of this method in such circumstances.

The paper is organized as follows. We first discuss the modelling of the marginal distributions, then in Section 3 investigate inference for the treatment effect using this marginal model. In Section 4 we develop a frailty model and discuss its estimation. Possible group differences are investigated in Section 5 using both approaches. Finally, in Section 6 we compare two approaches and discuss their implications for the use of Accutane in treating FOP. Some of the statistical results from this paper were used in Zasloff *et al.* [14].

2. THE MARGINAL MODEL

Here we consider modelling the age T_{ij} of the ith patient at the onset of involvement of joint j. Ignoring for the moment the treatment effect and any other covariates, we observe (t_{ij}, δ_{ij}), $i=1,\ldots,n$ where $\delta=0$ indicates right-censoring (in which case $T_{ij}>t_{ij}$), $\delta=1$ for uncensored observations ($T_{ij}=t_{ij}$) and $\delta=2$ for left-censoring ($T_{ij}<t_{ij}$). Rocke *et al.* [15] examined the marginal distributions for the control group using both Weibull models and Turnbull's non-parametric estimator [16]. Examples of the resulting survival curves, from the control group survey data, are given in Figure 1. The Weibull model gives results similar to the Turnbull estimates, and is more attractive since it can handle the time-dependent treatment covariate relatively easily.

Adopting the Weibull model, we assume that T_{ij} has a survival function $S_j(\cdot)$ given by

$$S_j(t)=P[T_{ij}\geqslant t]=\mathrm{e}^{-(\rho_j t)^{\kappa_j}}$$

for the control group, and

$$S_j^*(t)=P[T_{ij}\geqslant t]=\mathrm{e}^{-\tau(\rho_j t)^{\kappa_j}}$$

Copyright © 2002 John Wiley & Sons, Ltd.

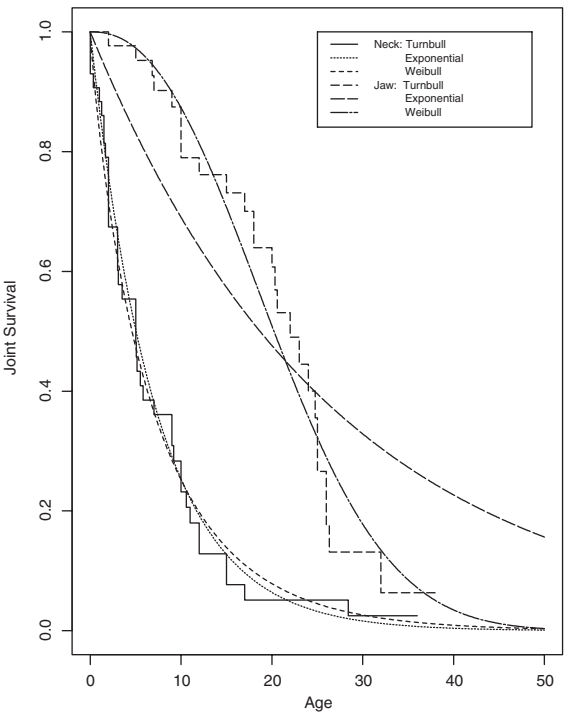

Figure 1. Estimated survival curves for the neck and jaw using the control group data. The methods used are Turnbull's non-parametric estimator, a fitted Weibull model and an exponential model.

for the treatment group, where τ represents the effect of treatment on the hazard rate. We assume here that the treatment effect is the same for each joint, and acts multiplicatively on the marginal hazard rates. We denote the corresponding hazard functions by $h_j(t)$, $h_j^*(t)$, respectively.

Assuming independent censoring, the contribution to the overall log-likelihood of a patient in the control group is

$$\ell_{ij}(\rho_j, \kappa_j, \tau) = \begin{cases} \log S_j(t_{ij}) & \text{if } \delta_{ij} = 0 \\ \log h_j(t_{ij}) + \log S_j(t_{ij}) & \text{if } \delta_{ij} = 1 \\ \log(1 - S_j(t_{ij})) & \text{if } \delta_{ij} = 2 \end{cases}$$

and the first and second derivatives are easily calculated, so that it is fairly straightforward to estimate the Weibull parameters for each joint by maximum likelihood using only the control group data. This was done by Rocke *et al.* [15], who found that some joints experienced an apparently constant hazard rate ($\kappa = 1$) whereas others had a hazard rate which increased with age ($\kappa > 1$). The parameter estimates and their asymptotic standard errors are given in Table I, together with the expected total number of joint failures in the treatment group during the study, assuming no treatment effect, and the actual number of joint failures observed. The expected values N_j are calculated by summing the reduction in $S_j(t)$ over the treatment period

Copyright © 2002 John Wiley & Sons, Ltd.

Table I. Estimated Weibull parameters (standard errors in parenthesis) from the control group (Rocke *et al.* [15]), with expected and observed joint failures for the treatment group.

Joint	ρ	κ	Expected	Observed
Neck	0.144 (0.026)	0.88 (0.11)	0	0
Spine	0.116 (0.018)	1.07 (0.14)	0	0
Jaw	0.042 (0.004)	2.32 (0.37)	1.41	0
Right shoulder	0.113 (0.016)	1.21 (0.16)	0.36	0
Right elbow	0.047 (0.007)	1.21 (0.18)	1.36	0
Right hip	0.052 (0.006)	1.62 (0.23)	1.17	0
Right knee	0.041 (0.003)	2.55 (0.37)	1.14	0
Left shoulder	0.123 (0.018)	1.14 (0.14)	0.91	0
Left elbow	0.038 (0.007)	1.14 (0.19)	1.34	1
Left hip	0.044 (0.007)	1.26 (0.20)	1.14	0
Left knee	0.037 (0.005)	1.59 (0.26)	1.28	1

for all joints uninvolved at the start of treatment, that is

$$N_j = \sum_{i:t_{ij} > s_i} [S_j(s_i) - S_j(t_{ij})]$$

where s_i denotes the age of the ith patient at the start of treatment and $S_j(\cdot)$ is evaluated using the estimated parameter values. Because the hazard rates are in general not constant, these expected values incorporate information on the ages of the treatment group subjects as well as the duration of their participation in the study and the status of each joint at recruitment. The expected values are zero for the neck and spine because both of these joints had already failed for all treatment group patients.

Where joints are in right–left pairs, it seems reasonable to assume that the parameters are the same for each side; we can estimate these parameters by pooling the data for right and left sides, thus allowing two observations per patient. Although some of the right and left parameters in Table I seem to be different (in particular κ for the knee joints), the pooling results in a drop in the total log-likelihood of 4.17 on 8 d.f., which is not significant. This test is not really valid because it is based not on the true likelihood but the working likelihood from the marginal model, which ignores the dependence between joints. However, assuming that correlation is positive between right–left pairs, the above test will result in underestimated p-values, so a valid test would also fail to reject. We assume in the subsequent analysis that right–left pairs have the same parameters, and return to discuss this point further in the final section.

For the treatment group patients three times are relevant: the age s_i at the start of treatment; the age e_i at the end of the study, and the age t_{ij} at which the joint became involved. Here we take the censoring variable δ_{ij} to be 0 if the joint was still uninvolved after treatment, 1 if involved during treatment and 2 if involved before treatment. The contribution to the overall log-likelihood of a patient in the treatment group is

$$\ell_{ij}(\rho_j, \kappa_j, \tau) = \begin{cases} \log S_j^*(e_i) + \log S_j(s_i) - \log S_j^*(s_i) & \text{if } \delta_{ij} = 0 \\ \log h_j^*(t_{ij}) + \log S_j^*(t_{ij}) + \log S_j(s_i) - \log S_j^*(s_i) & \text{if } \delta_{ij} = 1 \\ \log(1 - S_j(s_i)) & \text{if } \delta_{ij} = 2 \end{cases}$$

Copyright © 2002 John Wiley & Sons, Ltd.

For example, if $\delta_{ij}=1$ then joint j survived without treatment until time s_i but failed during treatment at time t_{ij}, the probability of this being

$$S_j(s_i)\times\frac{h_j^*(t_{ij})S_j^*(t_{ij})}{S_j^*(s_i)}$$

For any given value of the treatment effect (for example, $\tau=1$, corresponding to no effect) the Weibull parameters for a particular joint can now be estimated by maximum likelihood using the combined data from both treatment and control groups. If we assumed a joint-specific treatment effect τ_j it could also be estimated by maximum likelihood using the marginal model for joint j. Since only two new joints became involved during treatment (see Table I), the point estimate would in most cases be zero (that is, no hazard during treatment). The small number of patients in the treatment group, together with the fact that not all joints were initially uninvolved, means that no reliable inference about the treatment effect can be made from considering a single joint, so we are forced to assume a common τ and therefore to perform a multivariate analysis. Since all patients in the treatment group had both the neck and spine involved before the start of treatment, there is no information about τ in the data on these two joints so they may be omitted from the following analysis.

3. INFERENCE FOR THE MARGINAL MODEL

Using the approach of Wei *et al.* [3] with our fully parametric model, we consider estimation of the full parameter set $\beta=(\tau,\rho_1,\kappa_1,\ldots,\rho_J,\kappa_J)^{\mathrm{T}}$ by maximizing the working likelihood

$$\mathscr{L}(\beta)=\sum_i\sum_j\ell_{ij}(\rho_j,\kappa_j,\tau)$$

so that the estimator $\hat{\beta}$ solves the estimating equation

$$\sum_{i,j}\frac{\partial}{\partial\beta^{\mathrm{T}}}\ell_{ij}(\rho_j,\kappa_j,\tau)=0$$

We note that here β has 11 components, corresponding to τ and the Weibull parameters for each joint type. The neck and spine data have been dropped, and the data on left/right pairs pooled, so the number of joint types $J=5$. Assuming that the marginal models are correct, the consistency of $\hat{\beta}$ follows from a simple extension of the consistency argument of Huster *et al.* [9]. Asymptotic normality follows from Huber [4] under mild regularity assumptions. The sandwich estimator of the covariance of $\hat{\beta}$ is then given by

$$\hat{V}=A^{-1}\tilde{U}^{\mathrm{T}}\tilde{U}A^{-1}$$

where A is the observed information matrix

$$A=-\frac{\partial^2}{\partial\beta\partial\beta^{\mathrm{T}}}\mathscr{L}(\hat{\beta})$$

Copyright © 2002 John Wiley & Sons, Ltd.

and $\tilde{U}$ is the 'collapsed score matrix' with 61 rows, one for each patient, and 11 columns. The ith row of $\tilde{U}$ is

$$\sum_j \frac{\partial}{\partial \tau} \ell_{ij}, \frac{\partial}{\partial \rho_1} \ell_{i1}, \frac{\partial}{\partial \kappa_1} \ell_{i1}, \ldots, \frac{\partial}{\partial \rho_5} \ell_{i5}, \frac{\partial}{\partial \kappa_5} \ell_{i5}$$

the first element being zero for the 40 control group patients. A is block-diagonal apart from the first row/column. Specifically the first row is

$$\sum_i \left(\sum_j \frac{\partial^2}{\partial \tau^2} \ell_{ij}(\hat{\rho}_j, \hat{\kappa}_j, \hat{\tau}), \quad \frac{\partial^2}{\partial \rho_1 \partial \tau} \ell_{i1}(\hat{\rho}_1, \hat{\kappa}_1, \hat{\tau}), \ldots, \frac{\partial^2}{\partial \kappa_5 \partial \tau} \ell_{i5}(\hat{\rho}_5, \hat{\kappa}_5, \hat{\tau}) \right)$$

and the diagonal elements are the 2×2 joint-specific information matrices $\mathscr{I}_1, \ldots, \mathscr{I}_5$ where

$$\mathscr{I}_j = -\sum_i \begin{pmatrix} \dfrac{\partial^2}{\partial \rho_j^2} & \dfrac{\partial^2}{\partial \rho_j \partial \kappa_j} \\ \dfrac{\partial^2}{\partial \rho_j \partial \kappa_j} & \dfrac{\partial^2}{\partial \kappa_j^2} \end{pmatrix} \ell_{ij}(\hat{\rho}_j, \hat{\kappa}_j, \hat{\tau})$$

These are calculated routinely during the estimation of the Weibull parameters. The remaining elements of A are zero. This structure occurs because $\ell_{ij}(\rho_j, \kappa_j, \tau)$ does not depend on $\rho_{j'}$, $\kappa_{j'}$ for $j' \neq j$.

We are interested only in the first element $\hat{v}_{11}$ which estimates the standard error of $\hat{\tau}$. An approximate confidence interval for τ can now be constructed by assuming approximate normality for $\hat{\tau}$. However this leads to a 95 per cent confidence interval which includes negative values, which is particularly unsatisfactory because we are most concerned with the upper confidence limit. This is caused by the fact that the (pseudo-) likelihood surface is unsymmetrical with respect to τ. We can try to remedy this by reparameterizing the treatment effect to give a more symmetrical likelihood surface. Transformation of τ is easily incorporated into the above analysis since the τ derivatives can simply be adjusted using the chain rule.

We can investigate the effect of transformation on the likelihood surface by drawing the profile log-likelihood

$$\mathscr{L}(\tau) = \sum_j \max_{\rho_j, \kappa_j} \sum_i \ell_{ij}(\rho_j, \kappa_j, \tau)$$

for a range of values of τ. For τ untransformed this profile log-likelihood graph is far from symmetric, as can be seen in Figure 2(a). The natural transformation here would be to $\log \tau$ since this removes the boundary $\tau = 0$; it is also the usual parameterization in the Cox model. However the profile likelihood for $\log \tau$ is still far from symmetrical; in fact it errs in the opposite direction to that of the untransformed τ (Figure 2(b)). The problem here seems to be that, as noted earlier, very few uninvolved joints became involved during treatment, and it is only these which prevent $\hat{\tau}$ from being zero. To apply the robust variance estimator in constructing a confidence interval, we seek a transformation which will approximately symmetrize the profile log-likelihood. In Figure 2(c) we show the profile log-likelihood for the 0.5 power, and in Figure 2(d) we show the plot for 0.3, which appears to be a suitable choice.

Copyright © 2002 John Wiley & Sons, Ltd.

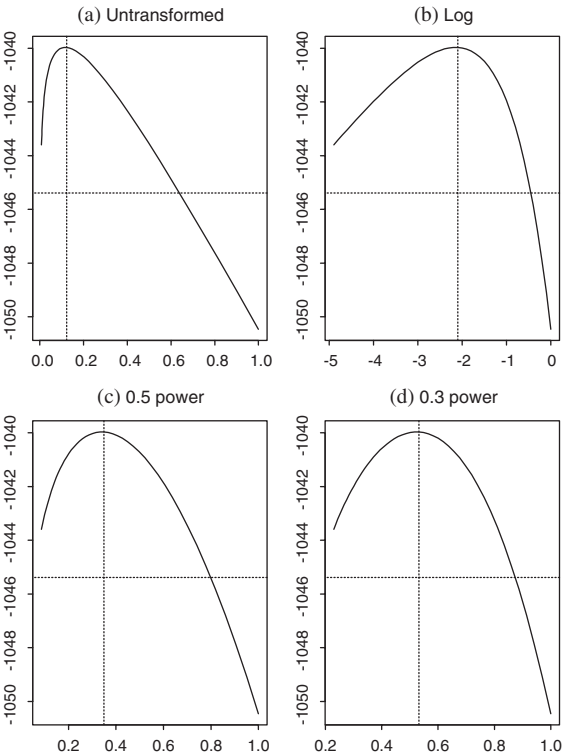

Figure 2. Effect of transformation of the treatment effect parameter on the shape of the profile likelihood: (*a*) untransformed profile likelihood; (*b*) logarithmic transformation; (*c*) 0.5 power; (*d*) 0.3 power.

An alternative consideration of the shape of the likelihood surface is obtained by allowing τ to vary while keeping the other parameters fixed. Because there are no left-censored observations during treatment, the second partial derivative with respect to τ of the log-likelihood is $-n_T/\tau^2$ where n_T is the number of observed events during treatment. It follows by application of the chain rule that the third derivative vanishes at the maximum when a cube-root transformation is used. (This result holds for any proportional hazards model with only right-censoring.) This is in broad agreement with the graphical approach using the profile likelihood. Using the cube-root transformation and applying the chain rule to the above robust variance methodology, we arrive at the 95 per cent confidence interval for τ of $(0.008, 0.497)$, or a reduction in the hazard of new joint involvement of between 50 and 99 per cent.

The point estimate is $\hat{\tau} = 0.117$. Table II gives the point estimates and standard errors for the joint-specific Weibull parameters, together with the expected and observed numbers of joint failures for both control and treatment groups. For the treatment group we give expected and observed counts before and during the treatment period, based on the estimated Weibull parameters and treatment effect. It is noticeable that the number of involved joints in this group prior to treatment is in all cases greater than that predicted by the model, suggesting that there may be a systematic difference between the groups. We postpone discussion of

Copyright © 2002 John Wiley & Sons, Ltd.

Table II. Estimated Weibull parameters (standard errors in parenthesis) from the marginal model, with expected (observed) joint failures for the control group and the treatment group before and during treatment.

Joint	ρ	κ	Control	Before	During
Neck	0.181 (0.033)	0.85 (0.13)	36.5 (37)	16.5 (21)	0.32 (0)
Spine	0.149 (0.022)	0.99 (0.15)	36.0 (35)	15.9 (21)	0.37 (0)
Jaw	0.043 (0.003)	2.05 (0.36)	25.4 (26)	5.6 (6)	0.26 (0)
Shoulders	0.136 (0.018)	1.09 (0.15)	71.4 (74)	31.0 (36)	0.82 (0)
Elbows	0.044 (0.006)	1.11 (0.13)	50.5 (52)	15.8 (16)	0.49 (1)
Hips	0.060 (0.006)	1.19 (0.21)	58.3 (59)	19.6 (29)	0.62 (0)
Knees	0.041 (0.003)	1.81 (0.23)	49.0 (50)	11.2 (14)	0.49 (1)

this until Section 5. We also note that if the expected number of joint involvements during treatment is calculated conditional on the observed number of failures prior to treatment, these become 0.19, 0.22, 0.36, 0.29, 0.35.

To investigate the validity of the robust variance approach, which is based on an asymptotic approximation, we simulated new data sets using the estimated parameter values of our model, keeping the same censoring times (that is, the ages of patients at interview, or start and end of treatment). To simulate correlated Weibulls with the given marginals we started with 11-dimensional standard normals, transformed these to equicorrelated normals, squared and added pairs to give correlated 11-dimensional exponentials, and then used the joint-specific parameters to convert the components to Weibulls. The degree of correlation is controlled by the off-diagonal element r in the matrix used to transform the standard normals, the correlation in the transformed normals being $(2r+9r^2)/(1+10r^2)$. Values of 0.1, 0.2, 0.3, 0.4 give correlations of 0.26, 0.54, 0.74, 0.86, respectively, thus ranging from fairly weak to quite strong correlation between the within-patient joint failure times. Based on 5000 simulations each, we found that the rates of inclusion of the true value of τ of nominal 95 per cent confidence intervals, using robust variance with the cube-root transform, were 95.7, 93.5, 91.4 and 90.4 per cent, respectively. Thus there appears to be a deterioration in performance as the strength of the correlation increases. To match the actual data to the simulation results we estimated within-patient correlations using the observed failure times and matched these informally with the correlations in the simulated data, which suggested that our data are closest to $r=0.3$. Thus our 95 per cent interval is perhaps closer to 90 per cent in coverage. The averages of the robust standard errors of the cube-root transformed treatment effect for each set of simulations were, respectively, 0.0847, 0.0845, 0.0849, 0.0862; the corresponding standard deviations of the point estimates were 0.0852, 0.0883, 0.0923, 0.0968. This seems to suggest that some form of parametric bootstrap might be a better procedure, but this would require a parametric model of the dependence. Moreover, neither the parametric nor the non-parametric bootstrap will work with our data because many of the bootstrap samples have no observed failures on treatment, resulting in an infinite estimated treatment effect ($\hat{\tau}=0$). Thus the use of the asymptotic robust variance, while not ideal, would seem to be the only tractable method, and is greatly improved by the cube-root transform.

4. FRAILTY MODEL

Here we take the hazard rate for patient i to be of the form $h_i(t)=Z_i h_0(t)$ where $h_0(t)$ is a baseline hazard which might depend on covariates such as, in our case, treatment with

Copyright © 2002 John Wiley & Sons, Ltd.

Accutane. The frailties $\{Z_i\}$ are usually assumed to come from a parametric family of distributions, for example the gamma distribution with mean 1. This gamma family is often convenient for estimation via the EM algorithm, the frailties being regarded as unobserved data. If the Cox partial likelihood is used, the frailties Z_i enter linearly into that part of the log-likelihood which depends on the regression and gamma parameters, and the conditional expectation of each Z_i is also easy to calculate.

In the presence of double-censoring, however, this convenience is lost. There is no partial likelihood, and the left-censoring introduces terms of the form $(1 - e^{-Z_i H_{ij}(\beta,t)})$ into the full likelihood so that the EM algorithm becomes intractable. When the terms for the ith person are multiplied out we get the sum of terms of the form $Z_i^{n_i} e^{-Z_i K_{ij}(\beta,t)}$, so the assumption of a gamma distribution still allows the frailties to be integrated out of the likelihood reasonably simply. However the resulting function is extremely complex in its dependence on the regression parameters β and the baseline survivor functions, so for highly multivariate problems such as ours a direct maximization of the likelihood is again intractable. Our approach instead is to use the conditional independences in the model as much as possible, iteratively estimating the frailties Z_i and the survivor function parameters ρ_j, κ_j to maximize the complete likelihood for each value of the parameters of interest, and to base inferences on these parameters on the resulting profile likelihood.

Frailties are easily incorporated into our model, the survival function for an individual with frailty Z_i becoming

$$S_j(t|Z_i) = \begin{cases} e^{-Z_i(\rho_j t)^{\kappa_j}} & \text{in the control group} \\ e^{-\tau Z_i(\rho_j t)^{\kappa_j}} & \text{in the treatment group} \end{cases} \tag{1}$$

The treatment effect is here assumed to have a multiplicative effect on the conditional hazards $h_j(t|Z_i)$.

Our approach to the treatment of the frailties is similar to that used by McGilchrist and Aisbett [17] to fit a frailty model to bivariate catheter infection data. They first consider the frailties to be fixed effects, which they estimate together with a regression parameter β by maximizing the log-likelihood $\mathscr{L}(z,\beta)$ subject to the constraint $\sum \log(Z_i) = 0$. They use the Cox model so their log-likelihood comes from the Cox partial likelihood. They then consider an alternative random effects model in which the log frailties are Gaussian with zero mean and unknown variance. They construct a 'penalized likelihood' by augmenting the previous log-likelihood of the observations with z conditionally fixed by the likelihood of z as given by the log-normal frailty distribution. This is again maximized subject to the same constraint on the vector z of frailties. This second alternative can be thought of as a hierarchical model in which the $\log Z_i$ for each patient is first chosen from $N(0, \sigma^2)$ and then the observations arise from the conditional distributions given Z_i. This approach is generally preferred because direct maximization over a large number of nuisance parameters as in the first method is likely to produce inconsistent estimates.

Taking the first approach, in which the Z_i are regarded as fixed effects, our model gives the log-likelihood as

$$\mathscr{L}(\tau, z, \theta) = \sum_i \sum_j \ell_{ij}(\tau, Z_i, \rho_j, \kappa_j)$$

where z is the vector of frailties and θ the vector of Weibull parameters. Inference for τ can be based on the profile likelihood [18]; for each fixed τ we maximize the above expression

Copyright © 2002 John Wiley & Sons, Ltd.

over z, θ. There is an identifiability problem since we can replace z by kz and ρ_j by $(\rho_j/k)^{1/\kappa_j}$ for any positive scalar k without changing the likelihood, so we impose $\sum \log(Z_i) = 0$. To achieve the optimization we use the following iterative scheme:

1. Initialize $Z_i \equiv 1$.
2. For each joint type j calculate the conditional maximum likelihood estimates for the Weibull parameters ρ_j, κ_j given z by maximizing $\sum_i \ell_{ij}(\tau, Z_i, \rho_j, \kappa_j)$.
3. For each subject i calculate the conditional maximum likelihood estimate for Z_i given θ by maximizing $\sum_j \ell_{ij}(\tau, Z_i, \rho_j, \kappa_j)$.
4. Adjust z to satisfy the constraint, multiplying each Z_i by $\mathrm{e}^{-\sum_i \log(Z_i)/n}$.
5. Repeat steps 2–4 until convergence.

We note that step 4 is not necessary for convergence but is necessary for identifiability of the model. By using alternating conditional maximum likelihood estimators we are able to replace a complex multiparameter optimization by a number of simpler univariate and bivariate ones. Repetition over a range of values of τ can be expedited by using the final z and θ from the previous τ as starting values for the next. The estimate is $\hat{\tau} = 0.031$, and using the usual asymptotic χ^2 approximation to the logged likelihood ratio gives the 95 per cent confidence interval $(0.005, 0.098)$. Each profile likelihood calculation also yields estimates of the frailties z of the individuals in the study. The obvious approach here would be to use the estimates of z at the maximizing values of τ and θ. These might be of some clinical interest, but the estimates are likely to be unreliable because of the large number of parameters involved. Figure 3 shows kernel-smoothed density estimates for these estimated log frailties, separately for each group. There is again some suggestion of a difference between the control and treatment groups, to be discussed in Section 5.

Turning now to the second approach, we consider a hierarchical model in which the Z_i are random drawings from some 'frailty distribution'. Figure 3 suggests that the log frailties are approximately normal, so we follow McGilchrist and Aisbett [17] in using penalized likelihood with a log-normal distribution, but differ slightly in that we consider the Z_i to be independent of each other, whereas they specify a multivariate distribution for z which incorporates the constraint $\sum \log(Z_i) = 0$. This constraint is no longer necessary for identifiability because of the penalty term, but it will be satisfied by the maximum likelihood estimates. We therefore continue to impose the constraint, not as part of the model but as part of the estimation procedure.

Our main focus of interest is again τ, but it may also be relevant to examine the 'frailty parameter' σ. For fixed (τ, σ) we maximize the likelihood over the Weibull and frailty parameters following steps 1–5 to give the 'penalized profile likelihood'

$$\mathscr{L}_{\mathrm{P}}(\tau, \sigma) = \max_{\rho, \kappa, z} \left\{ \sum_j \ell_{ij}(\tau, Z_i, \rho_j, \kappa_j) - 61 \log \sigma - \frac{(\log Z_i)^2}{2\sigma^2} \right\}$$

Evaluating this over a grid of values allows examination of the likelihood surface, as in Figure 4, which shows $\max_{\tau,\sigma} \mathscr{L}_{\mathrm{P}}(\tau, \sigma) - \mathscr{L}_{\mathrm{P}}(\tau, \sigma)$; the contour corresponding to level 3 is an approximate asymptotic 95 per cent confidence region. This suggests that $\log \tau$ is significantly less than zero, pointing again to a significant reduction in hazard with Accutane use. The point estimate for (τ, σ) is $(0.059, 0.94)$.

Copyright © 2002 John Wiley & Sons, Ltd.

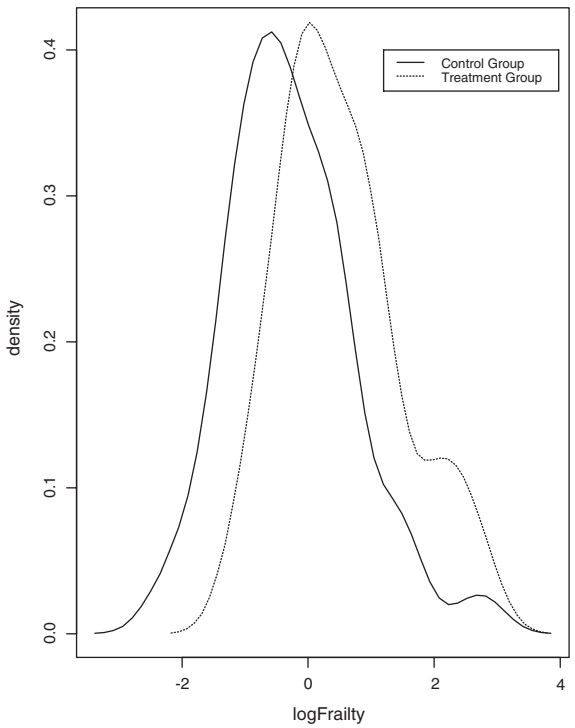

Figure 3. Kernel density estimates of the frailty distribution for the control and treatment groups.

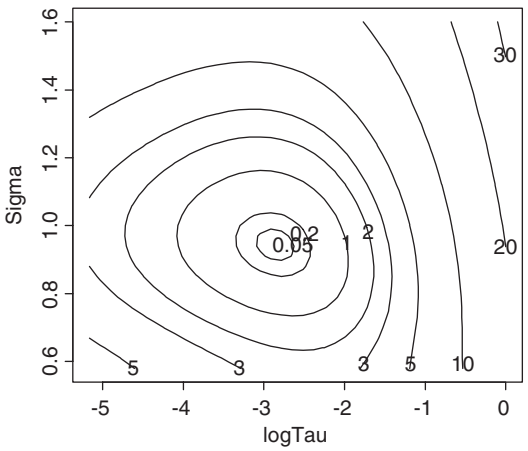

Figure 4. Penalized profile likelihood for the treatment and frailty parameters in the frailty model.

Copyright © 2002 John Wiley & Sons, Ltd.

5. TESTING FOR GROUP DIFFERENCES

The above analysis assumes that the treatment and control groups are comparable, in that they have the same baseline survival functions in the marginal model or the same frailty distribution in the frailty model. Since the allocation was not done on the basis of a randomized trial, there may be systematic differences between the groups which would affect and perhaps invalidate our inference about the treatment effect. We have noted two possible sources of bias: different recruitment criteria and different diagnosis of joint involvement. The second of these could perhaps be accommodated by arguing that the more sensitive radiographical diagnosis applied to the treatment group would tend, if anything, to diminish the apparent effectiveness of the treatment, but little can be said about the first. Furthermore there may be omitted covariates causing a difference between groups.

We saw in the fitted marginal model, from Table II, that the treatment group had a higher than expected number of joint involvements prior to the commencement of the treatment period, particularly in the neck and spine regions. Since these were determined by radiological scans it is possible that in many cases immobilization of the joint had not yet occurred, so that some of these would not be classed as involved using the control group criterion.

It was noted in Section 4 that the frailty analysis suggests, from Figure 3, that there is a systematic difference between the groups, with the treatment group having higher frailties than the control group: this effect, though reduced, still exists when a parametric family is used to model the frailties. This again is consistent with the more sensitive diagnostic techniques used in the clinical examinations. These differences are incorporated into the estimation of the treatment effect; if we eliminated the difference by forcing the log frailties to sum to zero separately for each group, this would then reduce the estimated frailties for the treatment group and hence reduce the apparent effectiveness of the treatment. It might be better in such situations to assume different distributions for the frailties in the control and treatment groups.

It should be noted that since the treatment is time-dependent, with some information available on the treatment patients before their treatment began, it would be possible to estimate τ using only this group. This does not work well in practice because there are only 21 subjects and nearly all the observations are either right- or left-censored. We can however use this pre-treatment information to examine possible differences between the groups. Using the marginal model approach we can subsume any recruitment, diagnostic and covariate differences into a single group difference parameter γ, assumed to have a multiplicative effect on the marginal hazard rates for the treatment group. Proceeding as before, we use a working independence model to estimate γ. The contributions to the working log-likelihood for the ith treatment group patient are now

$$\ell_{ij}(\rho_j, \kappa_j) = \begin{cases} \gamma \log S_j(s_i) & \text{if } \delta_{ij} = 0 \\ \log(1 - [S(s_i)]^{\gamma}) & \text{if } \delta_{ij} = 2 \end{cases}$$

with all observations being either left- or right-censored according to their status at the time of commencement of the treatment period. This gives $\hat{\gamma} = 1.84$ so the pre-treatment information suggests that if anything the treatment group are more prone to joint involvement than the controls at baseline, perhaps because of the more sensitive radiological diagnosis. This would tend to strengthen our conclusions about the effectiveness of the Accutane treatment. However the robust variance estimate gives the standard error of $\hat{\gamma}$ as 0.81, and the profile likelihood

Copyright © 2002 John Wiley & Sons, Ltd.

for the untransformed parameter is reasonably symmetric, so the estimated group difference is not significantly different from $\gamma = 1$. It seems reasonable then, given the limited information available, to proceed with the assumption of no group differences.

6. DISCUSSION

We have considered two different methods of trying to accommodate the within-patient dependence in multivariate survival analysis. Both approaches are sufficiently common in the literature to be considered standard (see for example Klein and Moeschberger [19]), but until now both the application and much of the theory has used partial likelihood in the Cox model. When there is double censoring, considerable adaptation is required. In particular there seems now to be no alternative to the estimation of the baseline survival functions, which brings both computational and inferential difficulties. These difficulties are further increased when we move from bivariate data, which has been the norm in the literature, to highly multivariate data ($p = 11$). Here we have taken a parametric approach to the modelling of the survival functions, but adapting it to a semi-parametric model in which the survivor functions are estimated non-parametrically would be very challenging. We have also had to assume that the treatment effect, and the frailties in the frailty model, are the same for each joint. Ideally one would want to check these assumptions, but given the sparsity of our data we have been unable to do so. The elbow and knee joints had one failure each during treatment, leading to confidence intervals of $(0.015, 1.03)$ and $(0.019, 1.09)$, respectively. It is clear from these that there is not enough data on individual joint failures to estimate and compare joint-specific treatment effects. There seems no alternative to the assumption that the effect of Accutane is homogeneous. We have noted that both joint failures affected the same patient. While it may be extreme to suppose that only this patient received no protection from Accutane, it is probably also extreme to suppose that Accutane protects every patient equally. Again we found no tractable alternative to the strong assumption made.

The most important conclusion is shared by both the marginal and frailty approaches; Accutane treatment leads to a significant reduction in the rate of involvement of previously unaffected joints. It is important to note however that the two approaches lead to different interpretations of the treatment effect, and to different marginal distributions for the joint involvements. The marginal model assumes that Accutane has the same multiplicative effect on each of the marginal hazard rates, and that the marginal distributions are Weibull. The frailty model assumes that Accutane has the same multiplicative effect on each of the conditional hazard rates of an individual given their frailty, and that the conditional distributions are Weibull. If we were to integrate out the frailty we would find that the marginal distributions were not Weibull, and that the effect of Accutane was not proportional on the marginal hazard. Since the Weibull model was originally chosen by examining the marginal distributions (Section 2), we might want to question the parametric forms used in the frailty model. The log-normal frailty distribution was chosen by examining the unrestricted frailty estimates, but other choices could have been made. The gamma distribution is a common choice, but its convenience is lost with double censoring. Hougaard [20] recommends a positive stable distribution since this preserves proportional hazards in the margins.

Both approaches as implemented here involve a large number of nuisance parameters, which given the small amount of data must raise concerns about the asymptotic results used. This

Copyright © 2002 John Wiley & Sons, Ltd.

is particularly true for the frailty model. With the frailties as fixed parameters there are in all 76 parameters and 671 observations. It is clearly preferable to regard the frailties as random effects, and the penalized likelihood approach does this, although ideally one would want to integrate the frailties out of the likelihood. This seems to be intractable with doubly-censored data. We have made some savings on the number of parameters by assuming that right–left pairs of joints have the same baseline hazards; that is, the same Weibull parameters. In Table II the κ parameters seem different for the hip and especially the knee joints, although these parameter values are not precise as evidenced by the large standard errors. Ideally we would like to test simultaneously the equality of all eight pairs of parameters, but the usual likelihood ratio test does not apply because it ignores within-patient correlation. A referee has suggested that differences between pairs might explain the lack of fit for hip in the treated group (Table II). However, the 29 pre-treatment failures comprise 14 left and 15 right hips. The lack of fit in this column is evidence that the treatment group appear to be more frail, as discussed in Section 5. That this frailty seems particularly pronounced for the hip joints is an interesting observation, but not one that we can explain.

An advantage of the marginal approach is that it applies without our having to specify the type of dependence between joints. The frailty model assumes that these are conditionally independent given the frailty, which may not be a reasonable assumption; although the mechanism is not known, some clinicians believe that the disease progresses in a characteristic pattern [2, 21]. The frailty model approach is, however, of added interest since it provides (through the frailty parameter σ) information about the range of differences in the severity of the disease which might be useful to clinicians; estimates of individual frailties could also be helpful in predicting prognosis for individual patients.

ACKNOWLEDGEMENTS

We thank Jeannie Peeper, Dr Michael Zasloff and Dr Frederick S. Kaplan for providing the data, and two anonymous referees for helpful comments and suggestions.

REFERENCES

1. Kaplan FS (ed.). *Clinical Orthopaedics and Related Research*. Lippincott-Raven: Philadelphia, 1998.
2. Cohen RB, Hahn GV, Tabas JA, Peeper J, Levitz CL, Sando A, Sando N, Zasloff M, Kaplan FS. The natural history of heterotopic ossification in patients who have fibrodysplasia ossificans progressiva. *Journal of Bone and Joint Surgery* 1993; **75-A**:215–219.
3. Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal of the American Statistical Association* 1989; **84**:1065–1073.
4. Huber PJ. The behaviour of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1967; **1**:221–233.
5. Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of chronic disease incidence. *Biometrika* 1978; **65**:141–151.
6. Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 1979; **16**:439–454.
7. Oakes D. A model for association in bivariate survival data. *Journal of the Royal Statistical Society*, *Series B* 1982; **44**:414–422.
8. Clayton DG, Cuzick J. Multivariate generalisations of the proportional hazards model. *Journal of the Royal Statistical Society*, *Series A* 1985; **148**:82–117.
9. Huster JH, Brookmeyer R, Self SG. Modelling paired survival data with covariates. *Biometrics* 1989; **45**:145–156.
10. Finkelstein DM. A proportional hazards model for interval-censored failure time data. *Biometrics* 1986; **42**: 845–854.

Copyright © 2002 John Wiley & Sons, Ltd.

11. Alioum A, Commenges D. A proportional hazards model for arbitrarily censored and truncated data. *Biometrics* 1996; **52**:512–524.
12. Satten GA. Rank-based inference in the proportional hazards model for interval censored data. *Biometrika* 1996; **83**(2):355–370.
13. Goggins WB, Finkelstein DM, Zaslavsky AM. Applying the Cox proportional hazards model when the change time of a binary time-varying covariate is interval censored. *Biometrics* 1999; **55**:445–451.
14. Zasloff MA, Rocke DM, Crofford LJ, Gregory VH, Kaplan FS. Treatment of patients who have fibrodysplasia ossificans progressiva with isotretinoin. *Clinical Orthopaedics and Related Research* 1998; **346**:121–129.
15. Rocke DM, Zasloff M, Peeper J, Cohen RB, Kaplan FS. Age- and joint-specific risk of heterotopic ossification inpatients who have fibrodysplasia ossificans progressiva. *Clinical Orthopaedics and Related Research* 1994; **301**:243–248.
16. Turnbull BW. Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association* 1974; **69**:169–173.
17. McGilchrist CA, Aisbett CW. Regression with frailty in survival analysis. *Biometrics* 1991; **47**:461–466.
18. Kalbfleisch JD, Sprott DA. Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society*, *Series B* 1970; **32**:175–209.
19. Klein JP, Moeschberger ML. *Survival Analysis*: *Techniques for Censored and Truncated Data*. Springer-Verlag: New York, 1997; 405–418.
20. Hougaard P. Survival models for heterogeneous populations derived from stable distributions. *Biometrika* 1986; **73**:387–396.
21. Kaplan FS, Tabas JA, Zasloff MA. Fibrodysplasia ossificans progressiva: a clue from the fly? *Calcified Tissue International* 1990; **47**:117–125.

Copyright © 2002 John Wiley & Sons, Ltd.