UNIVERSITY OF
Southampton
Southampton Statistical
Sciences Research Institute

Working Paper M11/09
Methodology

# A Covariance-Based Test For Shared Frailty In Multivariate Lifetime Data

Alan Kimber, Shah Jalal Sarker

## Abstract

We decompose the score statistic for testing for shared finite variance frailty in multivariate lifetime data into marginal and covariance-based terms. The null properties of the covariance-based statistic are derived in the context of parametric lifetime models. Its non-null properties are estimated using simulation and compared with those of the score test and two likelihood ratio tests when the underlying lifetime distribution is Weibull. Some examples are used to illustrate the covariance-based test. A case is made for using the covariance-based statistic as a simple diagnostic procedure for shared frailty in a parametric exploratory analysis of multivariate lifetime data.

# A Covariance-based Test for Shared Frailty in Multivariate Lifetime Data

Alan Kimber* and Shah Jalal Sarker**

* Southampton Statistical Sciences Research Institute, University of Southampton, UK

** Division of Health and Social Care Research, Guy's, King's and St Thomas' School of Medicine, King's College London, UK

## Abstract

We decompose the score statistic for testing for shared finite variance frailty in multivariate lifetime data into marginal and covariance-based terms. The null properties of the covariance-based statistic are derived in the context of parametric lifetime models. Its non-null properties are estimated using simulation and compared with those of the score test and two likelihood ratio tests when the underlying lifetime distribution is Weibull. Some examples are used to illustrate the covariance-based test. A case is made for using the covariance-based statistic as a simple diagnostic procedure for shared frailty in a parametric exploratory analysis of multivariate lifetime data.

Key words: Finite variance frailty; Gamma frailty; Multivariate lifetimes; Positive stable frailty; Proportional hazards; Score statistic; Survival analysis; Weibull distribution.

## 1. Introduction

Suppose we have an independent sample of *n* *p*-variate lifetimes (*p* > 1) where the *i*th observation is $(t_{i1}, t_{i2}, ..., t_{ip})^T$. A simple example of a situation in which such data would arise is a matched pair study (*p* = 2) in which component *j* (*j* = 1, 2) receives treatment *j* and the lifetime of each component is observed. Frailty modelling is a natural approach to modelling such data, see Duchateau and Janssen (2008) and Wienke (2011) for recent texts on this topic. In particular the shared frailty model is a natural extension to standard lifetime data modelling (see, for example, Collett, 2003). In this shared frailty model the conditional cumulative hazard of lifetime (*i*, *j*) is given by

$$H_{ij}(t_{ij} \mid z_j) = z_i H_{ij}(t_{ij}), \tag{1.1}$$

where $z_i$ is the frailty, a random effect that is shared by all lifetimes in observation *i*. Note in passing that if we replace $z_i$ by $z_{ij}$, where the $z_{ij}$ are mutually independent, then we obtain a marginal frailty model.

We shall refer to $H_{ij}(t_{ij})$ in (1.1) as the cumulative hazard of the underlying lifetime distribution for the random variable corresponding to $t_{ij}$. When $z_i \equiv 1$ then there is no frailty and we shall refer to this as the null model.

When $z_i$ varies with *i*, then the shared frailty model induces correlation between the *p* lifetime components and may also affect the marginal behaviour of the lifetime components. For example, if the underlying lifetime distribution is Weibull and the shared frailty is gamma distributed, then we obtain a multivariate Burr distribution for the *p*-variate lifetimes, which has positively correlated lifetime components and univariate Burr marginals (see, for example, Crowder, 1985).

A common approach is to adopt a Cox model for the underlying distribution with unspecified underlying cumulative hazard function:

$$H_{ij}(t_{ij}) = \exp(\beta_j^T x_{ij}) H_{0j}(t_{ij}), \tag{1.2}$$

where the $x_{ij}$ are fixed covariates and $H_{0j}(t_{ij})$ is an unspecified baseline hazard. Also, the distributional form, such as the gamma distribution, of the $z_i$ is commonly specified (see, for example, Siegmund *et al*, 1999). However, in some situations there may be background information on the form of the underlying distribution, perhaps via a weakest link argument, use of extreme value theory or from past experience. A Weibull distribution is a common choice, particularly in reliability and material strength applications. Moreover, the shared frailty is an unobservable random variable and so we may prefer not to be too prescriptive about the frailty distribution, at least at the exploratory stage of an analysis.

In this paper we consider testing for the presence of shared frailty using only simple quantities that are readily available from standard lifetime data analyses. We shall use models (1.1) with (1.2) but, unlike the Cox model approach, we shall specify the baseline hazard function up to a small number of unknown parameters in line with standard parametric lifetime data analysis. Taking $H_{0j}(t_{ij}) = t_{ij}^{\phi_j}$ would be appropriate for a Weibull-based model, for example. We shall assume that the frailty distribution has mean 1 and variance $1/\delta$ ($\delta > 0$). This is a weak assumption relative to assuming a specified distributional form but is an assumption nonetheless because, for example, it rules out distributions with infinite variance.

We shall also assume that the lifetime of component *j* of observation *i* may be censored at $c_{ij}$ and that $\delta_{ij}$ is an indicator taking the value 1 if the lifetime is observed and 0 if it is censored. Crowder and Kimber (1997) have obtained a score statistic to test for shared frailty but we will use the approach of Commenges and Andersen (1995) to decompose the score statistic into two parts, one of which is potentially useful in the shared frailty context. The score statistic in the notation is

$$\hat{U}_p = n^{-1}\sum_{i=1}^{n}\left\{\frac{s_{i\bullet}^2}{2} - \delta_{i\bullet}s_{i\bullet} - \frac{\delta_{i\bullet}(\delta_{i\bullet}-1)}{2}\right\}, \tag{1.3}$$

where $\delta_{i\bullet} = \sum_{j=1}^{p}\delta_{ij}$ , $s_{i\bullet} = \sum_{j=1}^{p}s_{ij}$ where $s_{ij} = \hat{H}_{ij}^{\delta_{ij}}(t_{ij})\hat{H}_{ij}^{1-\delta_{ij}}(c_{ij})$ and a hat indicates that any unknown parameters have been estimated by maximum likelihood under the null model.

The corresponding score statistic to test for marginal frailty in component *j* is, trivially adapting the result given Kimber (1996),

$$\hat{V}_j = n^{-1}\sum_{i=1}^{n}\left\{\frac{s_{ij}^2}{2} - \delta_{i\bullet}s_{ij}\right\}. \tag{1.4}$$

In section 2 we obtain the null properties of the new statistic. In section 3 we compare he power and robustness of the new statistic with the score test and also likelihood ratio methods. Some illustrative examples are given in section 4 and further discussion appears in section 5. Technical details are outlined in an appendix.

## 2. A covariance statistic and its null properties

Let

(2.1) $$\hat{T}_{j,k} = n^{-1}\sum_{i=1}^{n}(s_{ij}-\delta_{ij})(s_{ik}-\delta_{ik})$$

for *j, k* = 1, 2, … *p* (though we shall only use the cases where *j*<*k*).

Then, following Commenges and Andersen (1995), it is easy to show that

(2.2) $$\hat{U}_p = \sum_{j=1}^{p}\hat{V}_j + \sum_{j<k}\hat{T}_{j,k} \, .$$

This analysis of variance type decomposition of the score statistic for shared frailty into the sum of score statistics for marginal frailty and the sum of covariance type terms is a natural one because finite variance frailty affects the underlying survival distribution by acting on the marginal distributions and by inducing association between components.

The null properties of $\hat{U}_p$ and the $\hat{V}_j$ are available using the methods of Crowder and Kimber (1997) and Kimber (1996). So here we shall concentrate on the null properties of the $\hat{T}_{j,k}$. For the remainder of his section all the properties derived are for the null, no frailty case.

Consider initially the case of no censoring, so that for all *i* and *j* $\delta_{ij} = 0$. First, suppose for now that all the parameter values in the $H_{ij}$ are known. Then the probability integral transform indicates that in the null case the $s_{ij}$ are independent observations from the exponential distribution with mean 1. Thus, it is trivial to show that $\sqrt{n}\hat{T}_{j,k}$ has mean 0 and variance 1. Also, by the central limit theorem $\sqrt{n}\hat{T}_{j,k}$ is asymptotically normally distributed. Again, it is easy to show from first principles that the *p*(*p*+1)/2 terms in the decomposition (2.2) are uncorrelated.

Now, if we allow unknown parameters in the $H_{ij}$ and estimate them using maximum likelihood, we may use the approach of Pierce (1982) to find the asymptotic null distribution of $\sqrt{n}\hat{T}_{j,k}$ (as used by Crowder and Kimber (1997) and Kimber (1996) to find the asymptotic null distributions of $\hat{U}_p$ and $\hat{V}_j$ respectively). The key element here is that, unlike $\hat{U}_p$ and $\hat{V}_j$, there are no $s_{ij}^2$ terms, so that the "correction" term for the asymptotic variance of $\sqrt{n}\hat{T}_{j,k}$ that allows for parameter estimation is 0. See the Appendix for further discussion of this point.

Having obtained the null asymptotic variance of $\sqrt{n}\hat{T}_{j,k}$ and using the results of Crowder and Kimber (1997) and Kimber (1996), we see that the null asymptotic variance of $\hat{U}_p$ is simply the sum of the null asymptotic variances of the $p(p+1)/2$ terms in the decomposition, so that the terms in the decomposition (2.2) are asymptotically uncorrelated.

If we now allow for censoring to occur, then the above null properties still hold with some adjustment to the asymptotic variance of $\sqrt{n}\hat{T}_{j,k}$. This may now by estimated by

(2.3) $$V\hat{a}r(\sqrt{n}\hat{T}_{j,k}) = n^{-1}\sum_{i=1}^{n}(1-e^{-d_{ij}})(1-e^{-d_{ik}}),$$

where $d_{ij} = \hat{H}_{ij}(c_{ij})$ with any unknown parameters replaced by their null maximum likelihood estimates. In the simple case of type I censoring with no covariates, then $c_{ij} = c_j$ and $d_{ij} = d_j$ for all *i*, so that

(2.4) $$V\hat{a}r(\sqrt{n}\hat{T}_{j,k}) = (1-e^{-d_j})(1-e^{-d_k}).$$

Note that as expected the right hand side of (2.4) tends to 1 (the result for the uncensored case) as the censoring points grow large.

Let

(2.5) $$\hat{T}_p = \sum_{j<k}\hat{T}_{j,k}.$$

Returning to the general censoring case, we have that $\sqrt{n}\hat{T}_p$ has a null distribution that is asymptotically normal with mean 0 and variance that may be estimated by

(2.6) $$V\hat{a}r(\sqrt{n}\hat{T}_p) = n^{-1}\sum_{j<k}\sum_{i=1}^{n}(1-e^{-d_{ij}})(1-e^{-d_{ik}}).$$

Thus, either $\hat{T}_p$ or the $\hat{T}_{j,k}$ may be used as diagnostic test statistics for detecting frailty to augment use of $\hat{U}_p$. Note that all these statistics depend the $s_{ij}$, which are simply Cox-Snell residuals (Cox and Snell, 1968). Such residuals are routinely available for standard survival distributions like the Weibull distribution. Large positive values of either $\hat{T}_p$ or the $\hat{T}_{j,k}$ are indicative of departures from the null hypothesis of no shared frailty.

Note that for these results to hold there must be separate parameters for each of the *p* components so that under the null, no frailty model we have essentially *p* independent data sets with no parameters in common. We shall discuss this further in section 5.

Note also that in (2.3), (2.4) and (2.6) the particular form of the underlying distribution enters only in calculating the $d_{ij}$. This is in contrast to the corresponding properties of $\hat{U}_p$ derived in Crowder and Kimber (1997) where different underlying distributions lead to different variances. For example, in the case in which the underlying distribution is Weibull with both parameters unknown the null variance of $\hat{U}_p$ when there is no censoring is

$$\frac{p(p+3)}{2} - p(1+\frac{6}{\pi^2}) \; .$$

The corresponding result for an exponential underlying distribution is

$$\frac{p(p+1)}{2} \, .$$

The reason is that the "correction term" in the variance to allow for parameter estimation (Pierce,1982) is non-zero for $\hat{U}_p$ and involves the expected information matrix of the null model, which depends on the particular underlying distribution being used. See the Appendix.

## 3. Power calculations

The most widely used standard parametric survival distribution is the Weibull and so we investigate our frailty tests for *p*-variate Weibull survival data, mostly with *p*=2 which corresponds to matched pairs. For each situation (*n*, *p*, frailty distribution, censoring regime) we used 2000 simulated samples of size *n* to estimate the percentage power, so that, using the binomial distribution, an upper bound (attained when the true power is 50%) for each percentage power estimate is 1.1% . The numerical results given here are reported for the case *n* = 50 and for a 5 per cent significance level. Other combinations give qualitatively similar results but are omitted for brevity. Since shared frailty can induce correlation and affect marginal behaviour, the results reported cover the cases: correlation and marginal effects (finite variance frailty), correlation only (infinite variance frailty) and marginal effects only (marginal frailty).

### 3.1 Comparison of the performances of score-based and likelihood ratio tests for frailty

A natural alternative to using score-based procedures is likelihood ratio testing for frailty. There are two issues here. First, likelihood ratio tests require the frailty model to be fitted explicitly, which is certainly more computationally challenging than fitting, say, a standard Weibull model followed by use of standard residuals. Secondly, a likelihood ratio test for frailty requires the frailty distribution to be specified (a gamma distributed frailty distribution is often used in practice). Since the frailty distribution is unobservable, the robustness of likelihood ratio tests in this context is not obvious.

We consider the bivariate case (*p*=2) where the underlying distribution is Weibull. We shall investigate the four test statistics $\hat{U}_2$, $\hat{T}_2$ $(=\hat{T}_{1,2})$, $L_G$ and $L_{PS}$, where the last two are likelihood ratio statistics assuming that the frailty distribution is gamma and positive stable respectively. Note that the positive stable distribution does not have finite variance and so lies outside the class of frailty distributions our score-based tests were set up to detect.

Tables 1 and 2 show respectively the power results for *n*=50 using a 5 per cent significance level with no censoring and no covariates for the situations where the frailty distribution is (a) gamma with mean 1 and variance $1/\delta$ $(\delta > 0)$ and (b) positive stable with characteristic exponent $\nu$ $(0.5 < \nu < 1)$. Note that $\delta = \infty$ and $\nu = 1$correspond to the null case of no frailty. Note also that since the Cox-Snell residuals are invariant to changes in the Weibull scale and shape parameters, an exponential underlying distribution with mean 1 was used in the simulation study, though in calculating all the statistics it was assumed that the density of each underlying Weibull distribution had unknown shape and scale parameters. Other values of *n* gave qualitatively similar results.

If we consider first the score-based tests, we see from Table 1 that in this "ideal" situation (i.e. shared frailty with finite variance), $\hat{U}_2$ is slightly more powerful than $\hat{T}_2$. However, in Table 2 we see that $\hat{T}_2$ is more robust than $\hat{U}_2$ in that it maintains its power much better when the frailty distribution has been misspecified. From these two tables it is clear that if the form of the frailty distribution is correctly specified, then the relevant likelihood ratio test is the most powerful of the four tests (though the score-based tests are still competitive in the gamma frailty case). However, if the form of the frailty distribution is incorrectly specified, then the likelihood ratio test does rather worse than $\hat{T}_2$. Thus, on robustness grounds, together with simplicity and numerical convenience, there is certainly a case for using $\hat{T}_2$.

### 3.2 The effect of censoring on the powers of the score-based methods

As before, Tables 3 and 4 show respectively the power results for *n*=50 using a 5 per cent significance level with no censoring and no covariates for the situations where the frailty distribution is (a) gamma with mean 1 and variance $1/\delta$ ($\delta > 0$) and (b) positive stable with characteristic exponent $\nu$ ($0.5 < \nu < 1$), but this time with three censoring regimes: (i) $c_1 = c_2 = 2.97$ (ii) $c_1 = c_2 = 1.80$ (iii) $c_1 = 1.20; c_2 = \infty$. Case (i) corresponds to approximately 5% censoring in each margin, case (ii) corresponds approximately to approximately 16.5% censoring in each margin, while case (iii) corresponds to approximately 30% censoring in one margin only.

The relative performances of $\hat{U}_2$ and $\hat{T}_2$ are qualitatively similar to those in the no censoring case: $\hat{U}_2$ is somewhat more powerful than $\hat{T}_2$ when the frailty has finite variance, but $\hat{T}_2$ is considerably more powerful when the frailty distribution is misspecified in the sense of having infinite variance. However, in the finite variance case the effect of censoring is to reduce the powers of both tests. In contrast, in the infinite variance case the effect of censoring is broadly to increase the powers relative to the no censoring case. An explanation is (Caroni and Kimber, 2004) that the effect of finite variance frailty tends to manifest itself as upper outliers, whereas the positive stable frailty distribution tends to produce lower outliers. Thus, for the gamma frailty distribution censoring tends to mask the effect of frailty, thereby making it more difficult to detect. However, for the positive stable frailty distribution, where frailty effects are concentrated in the lower tail and dissipated in the upper tail, censoring means that the lower tail effects are not watered down by relatively uninformative upper tail values.

### 3.3 Powers of the score-based tests when there is marginal but not shared frailty

Tests based on $\hat{U}_2$ and $\hat{T}_2$ have been designed to detect shared, finite variance frailty. In this section we consider the case in which there is only marginal frailty. We model this by simulating samples that comprise independent pairs of observations, each observation having a Burr distribution (Crowder,1985) obtained by combining an underlying Weibull survival time with a gamma frailty with mean 1 and variance $1/\delta$ ($\delta > 0$).

Tables 5 shows the power results for $\hat{U}_2$ and $\hat{T}_2$ for *n*=50 using a 5 per cent significance level with no covariates for the situation where there is marginal gamma frailty with mean 1 and variance $1/\delta$ $(\delta > 0)$ when there is (a) no censoring, (b) $c_1 = c_2 = 1.80$ (c) $c_1 = 1.20; c_2 = \infty$. Broadly speaking, when there is censoring in both margins the powers of $\hat{U}_2$ and $\hat{T}_2$ are both low. However, when one or more margin is not censored $\hat{U}_2$ has non-negligible power for moderate to strong marginal frailty, whereas $\hat{T}_2$ does not. Thus, $\hat{T}_2$ tends not to give a significant result when only marginal frailty is present but $\hat{U}_2$ tends to give a significant result when not all margins are subject to censoring.

### 3.4 Powers of tests based on the score-based tests when *p*=3

Tables 6 and 7 show respectively the power results for *n*=50 and *p*=3 using a 5 per cent significance level with no censoring and no covariates for the situations where the frailty distribution is (a) gamma with mean 1 and variance $1/\delta$ $(\delta > 0)$ and (b) positive stable with characteristic exponent $\nu$ $(0.5 < \nu < 1)$. The relative performances of $\hat{U}_2$ and $\hat{T}_2$ are qualitatively similar to those in the bivariate case: $\hat{U}_3$ is somewhat more powerful than $\hat{T}_3$ when the frailty has finite variance, but $\hat{T}_3$ is more powerful when the frailty distribution is misspecified in the sense of having infinite variance. Not surprisingly, the powers of the tests are higher than in the bivariate case with the same *n* and frailty distribution since there are 50% more items of data per sample when *p*=3 than when *p*=2.

## 4. Examples

### Example 1: infant nutrition

We first use a set of data on the ages of introduction of two types of potentially allergenic food in infant diets. Further details of the data may be found in Kimber (1996). The data comprise the ages in months at which fish and egg were first given to each of 55 infants in an infant nutrition study carried out in Madrid, Spain. There is no censoring present in the data and the null model assumed is that

$$H_{ij}(y_{ij}) = H_j(y_{ij}) = \exp(\alpha_j) y_{ij}^{\phi_j}, \tag{4.1}$$

with $i = 1, 2, \ldots, 55$ and $j = 1,2$. Thus, the null model is that the two ages of introduction are independent Weibull random variables, but it is felt that there may be a shared frailty effect, possibly corresponding to unmeasured psycho-social factors of the main carers for the infants.

Now, under the null model $\hat{U}_2$ is approximately Normal with mean 0 and variance $3(1-4/\pi^2)/n$; see Crowder and Kimber (1997) for details. The observed value of $\sqrt{n}\hat{U}_2 / \sqrt{3(1-4/\pi^2)}$ is 3.45, which is highly significant. Taken on its own, this result would seem to indicate the presence of shared frailty. However, $\sqrt{n}\hat{T}_2$= 1.41, which, using the results of section 2, may be referred to the standard normal distribution. Clearly, this result is not significant at the 5% level. Taking these two diagnostic tests together suggests that frailty may be present but that the evidence for shared frailty is relatively weak.

### Example 2: reaction times

Crowder and Kimber (1997) investigated the pre-test and post-test reaction times in seconds of $n$ = 9 rats in a study on the effect of lead levels. These data are a coherent subset of a larger data set, full details of which are given in Crowder (1989). A feature of the data is that reaction times that exceed 250 seconds have been censored. The data set is smaller than is ideal for detecting frailty but it does illustrate how the test may be applied in the presence of censoring. The null model is of the same form as (4.1) used in Example 1. The nine data pairs are (45, 214), (40, 218.5), (52.5, 250*), (57, 211), (40.5, 117.5), (26.5, 250*), (58.5, 179.5), (35.5, 193.5), (33, 141.5). With $c_1 = c_2 = 250$ and using Weibull maximum likelihood estimation for each marginal distribution, we obtain $d_1 = 2457.66$ and $d_2 = 1.53323$. Thus, using (2.4) we see that $\sqrt{n}\hat{T}_2$ has estimated null standard deviation 0.886. However, $\sqrt{n}\hat{T}_2 = -0.039$ and since large positive values of $\sqrt{n}\hat{T}_2$ are significant, it is clear that there is no evidence of frailty here. This is in line with the analysis based on $\hat{U}_2$ given in Crowder and Kimber (1997), which also gave a non-significant result. Thus, overall there is little evidence for shared frailty in the data.

**Example 3: braided cords**

Crowder and Kimber (1997) give trivariate strength data for 40 braided cords (so that the “survival time” is in fact strength in this example). The null model is as for (4.1) in Example 1 except that here *j*=1, 2, 3. Crowder and Kimber (1997) show that $\hat{U}_3$ is highly significant but we now carry out the test based on $\hat{T}_3$. Now, after fitting the null Weibull model to the three sets of strengths we obtain $\sqrt{n}\hat{T}_{1,2}$ =16.272, $\sqrt{n}\hat{T}_{1,3}$= 12.979 and $\sqrt{n}\hat{T}_{2,3}$= 11.498. Further, the null distribution of $\sqrt{\frac{n}{3}}\hat{T}_3$ is standard Normal since each of the three $\sqrt{n}\hat{T}_{j,k}$ terms has null asymptotic variance 1 and they are asymptotically uncorrelated. The observed value of $\sqrt{\frac{n}{3}}\hat{T}_3$ is $(16.272+12.979+11.498)/\sqrt{3}$ = 23.526, which is clearly highly significant. The two results together indicate that the shared frailty model is plausible here.

**Example 4: exercise times to angina**

Pickles and Crouchley (1994) studied the time to angina of patients after being given a dose of isosorbide dinitrite. Table 8 shows the data we shall use here. Here 21 patients were given a dose of the drug and 1 hour and 3 hours after receiving the drug they used exercise bikes until they felt angina or were too exhausted to continue, the latter situation corresponding to censoring.

The null model

$$(4.2) \qquad H_{ij}(y_{ij}) = \exp(\beta_{0j} + \beta_{1j}x_i)y_{ij}^{\phi_j}$$

was fitted with *i* = 1, 2, …, 21 and *j* = 1,2; here $x_i$ is the dose given to patient *i*. Table 9 gives the null maximum likelihood estimates for this model. Using these results to obtain the Cox-Snell residuals yields $\sqrt{n}\hat{T}_2 = 2.938$. Since there is censoring the null asymptotic variance of this statistic is, by inspection of (2.3), clearly less than 1. Thus, even without evaluating this variance explicitly, it is clear that the result is highly significant so that there appears to be shared frailty in the data.

## 5. Discussion

We shall discuss our results using the matched pair set up as an illustration.

Doing an initial check on whether there is shared frailty may have value in terms of modelling the lifetime data and may be useful in checking whether the matching has worked. For example, in a quality control procedure if one selects each pair from a different batch, a non-significant test for shared frailty may be indicative that the batch to batch variability is low. Conversely, an indication of high batch to batch variability may give a warning.

In the spirit of wanting a simple initial analysis that uses standard lifetime data methods both $\hat{U}_2$ and $\hat{T}_2$ are simple to calculate (but see below) and give slightly different information. If $\hat{U}_2$ and $\hat{T}_2$ are both significant, then there is evidence of shared frailty. If neither is significant, then the data give little evidence of shared frailty (though of course that could be because censoring has masked the presence of frailty). More interestingly, if $\hat{U}_2$ is significant but $\hat{T}_2$ is not, then there is an indication that marginal frailty may be more important than shared frailty. Likewise if $\hat{T}_2$ is significant but $\hat{U}_2$ is not, then this may suggest that there is shared frailty but possibly not with finite variance. Note that for the censoring regimes considered in this paper it is easy to obtain an estimate of the null asymptotic variance of $\hat{T}_2$ but it can be more tricky to do so for $\hat{U}_2$ (see the calculations performed in Crowder and Kimber (1997) and compare these with (2.3), (2.4) and (2.6)).

Models such as (4.1) with no explicit covariate information are clearly very simple but may still be useful. For example, if component *j* corresponds to treatment *j* then (4.1) allows for a treatment potentially to impact on all its within component parameters. Thus, there is flexibility here and at the exploratory analysis stage one is not forcing a particular treatment effect model on the data. On the other hand, models such as (4.2) still allow this flexibility of treatment effect but impose more structure on the effects of other prognostic variables. Because there must be no overlap in parameters between components, there is still some flexibility in modelling the effects of prognostic variables (i.e. we do not force the effect of a prognostic variable to be the same in each component).

We have concentrated on the Weibull underlying distribution because of its ubiquity in many branches of lifetime data analysis. The exponential and Rayleigh distributions remain popular choices in some branches of reliability and these can easily be incorporated in the models we have used by setting $\phi_j = 1$ and 2 respectively for all *j* in our Weibull models and the results of section 2 apply unchanged. Other lifetime distributions could also just be slotted in to the results of section 2 provided all maximum likelihood estimation required is regular (so, for example, the two parameter exponential distribution with unknown location parameter could not be used since estimation of the location parameter is not regular).

The simple structure of the null model means that the only constraint on the prognostic variables is that they allow all null parameters in a component to be estimable. However, we envisage statistics such as $\hat{T}_2$ as being most useful when the number of prognostic variables is small.

Another approach for paired data has been set out by Owen (2005) where by using ratios of variables in an accelerated failure time set up (see, for example, Collett, 2003) he enables shared frailty terms to cancel out, so that the frailty distribution essentially disappears from the analysis; see also Wang (2010). Since with a Weibull underlying distribution the proportional hazards and accelerated failure time frailty models are equivalent this is an alternative if frailty is simply a nuisance and of no interest in itself. Also, in order to obtain simple distributional results the ratio approach requires equal Weibull shape parameters across components, which may not be appropriate in all applications. For $p > 2$ such ratio methods would become increasingly complex.

## References

Caroni, C. and Kimber, A.C. (2004). Detection of frailty in Weibull lifetime data using outlier tests. *J Statistical Computation and Simulation*, 74, 15-23.

Collett, D. (2003). *Modelling Survival Data in Medical Research* (2nd edition). Chapman & Hall/CRC.

Commenges, D. and Andersen, P.K. (1995). Score test of homogeneity for survival data. *Lifetime Data Analysis*, 1, 145-156.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *J Roy Statist Soc*, **B**30, 248-275.

Crowder, M. J. (1985). A distributional model for repeated failure time measurements. *J Roy Statist Soc*, **B**47, 447-452.

Crowder, M. J. (1989). A multivariate distribution with Weibull connections. *J Roy Statist Soc*, **B**51, 93-107.

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

Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. Springer.

Kimber, A.C. (1996). A Weibull-based score test for heterogeneity. *Lifetime Data Analysis*, 2, 63-71.

Owen, W.J. (2005). A power analysis of tests for paired lifetime data. *Lifetime Data Analysis*, 11, 233-243.

Pickles, A. and Crouchley, R. (1995). A comparison of frailty models for multivariate survival data. *Statistics in Medicine*, 14, 1447-1461.

Pierce, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain type of statistics. *Ann Statist*, 10, 475-478.

Siegmund, K.D., Todorov, A.A. and Province, M.A. (1999). A frailty approach for modeling diseases with variable age of onset in families: the NHLBI family heart study. Statistics in Medicine, 18, 1517-1528.

Wang, Z. (2010). Tests for paired lifetime data with frailty models. *Communications in Statistics – Theory and Methods*, 39, 3122-3139.

Wienke, A. (2011). *Frailty Models in Survival Analysis*. Chapman & Hall/CRC.

## Appendix

We use Crowder and Kimber (1997) and Pierce (1982) to show that the null asymptotic variance of $\hat{T}_2$ is the same whether all parameters are known or estimated by maximum likelihood.

Let $v_n$ and $\hat{v}_n$ denote the null variances of $\hat{T}_2$ with and without regular parameter estimation respectively. Let $\lambda$ denote the parameter vector for the null model. Then putting Pierce's result in the present context

(A1) $$\hat{v}_n = v_n - B_n^T J_n B_n ,$$

where $J_n$ is the expected information matrix for the null model and $B_n$ is the null mean vector of $\partial\hat{T}_2 / \partial\lambda$ when there is no parameter estimation required. We now show that $B_n$ is the zero vector, which gives the required result.

To show this, take the contribution to $B_n$ of observation *i* and then drop the subscript *i* for simplicity and suppose without loss of generality that $\theta$ is a component 1 parameter. Then the contribution is $s_1 s_2 - s_1 - s_2 + 1$ if both components are observed, $s_1 d_2 - d_2$ if only component 2 is censored, $d_1 s_2 - d_1$ if only component 1 is censored, and $d_1 d_2$ if both components are censored. The derivatives of the contribution with respect to $\theta$ involve component 2 via a multiplicative factor $s_2 - 1$ when component 2 is observed and via a multiplicative factor $d_2$ when component 2 is censored. Since the variables in the null model are independent the double integral required to find the expectation is separable. Thus, the expectation of the derivative is proportional to

$$\int_0^{d_2} (s_2 - 1)\exp(-s_2)ds_2 + \int_{d_2}^{\infty} d_2 \exp(-s_2)ds_2 = 0 .$$

Hence the required result for $\hat{T}_2$. The result clearly generalises to $\hat{T}_p$ since the components of $\hat{T}_p$ are asymptotically uncorrelated under the null model and are of the same form as $\hat{T}_2$.

Note that the correction term in (A1) involves the information matrix, which clearly depends on the form of the underlying distribution. Thus, when $B_n$ is non-zero, as is the case for $\hat{U}_p$ (Crowder and Kimber, 1997) , then different underlying distributions will lead to different expressions for the variance. This is contrast to the variance of $\hat{T}_p$ where the variance result (2.6), for example, applies across underlying distributions.

Table 1 Estimated powers (%) of $\hat{U}_2$, $\hat{T}_2$ $(=\hat{T}_{1,2})$, $L_G$ and $L_{PS}$ when *p*=2, *n*=50, the underlying survival distribution is Weibull and the frailty distribution is gamma with mean 1 and variance $1/\delta$.

| $\delta$ | $\hat{U}_2$ | $\hat{T}_2$ | $L_G$ | $L_{PS}$ |
|---|---|---|---|---|
| 0.5 | 100.0 | 100.0 | 100.0 | 100.0 |
| 1 | 99.7 | 97.7 | 99.9 | 99.0 |
| 2 | 90.5 | 82.4 | 92.2 | 71.3 |
| 4 | 56.2 | 47.6 | 58.6 | 33.7 |
| 8 | 27.4 | 23.3 | 27.9 | 15.4 |
| 16 | 12.6 | 11.4 | 15.2 | 8.8 |

Table 2 Estimated powers (%) of $\hat{U}_2$, $\hat{T}_2$ $(=\hat{T}_{1,2})$, $L_G$ and $L_{PS}$ when *p*=2, *n*=50, the underlying survival distribution is Weibull and the frailty distribution is positive stable with characteristic exponent ν.

| $\nu$ | $\hat{U}_2$ | $\hat{T}_2$ | $L_G$ | $L_{PS}$ |
|---|---|---|---|---|
| 0.5 | 87.9 | 99.0 | 96.0 | 100.0 |
| 0.6 | 70.6 | 90.4 | 87.2 | 99.0 |
| 0.7 | 48.8 | 68.9 | 63.3 | 94.9 |
| 0.8 | 26.7 | 38.1 | 37.0 | 75.1 |
| 0.9 | 12.7 | 17.7 | 14.8 | 35.8 |
| 0.95 | 7.2 | 8.9 | 10.3 | 15.2 |

Table 3 Estimated powers (%) of $\hat{U}_2$ and $\hat{T}_2$ $(=\hat{T}_{1,2})$ when *p*=2, *n*=50, the underlying survival distribution is Weibull and the frailty distribution is gamma with mean 1 and variance $1/\delta$ under three censoring regimes: (i) $c_1 = c_2 = 2.97$ (ii) $c_1 = c_2 = 1.80$ (iii) $c_1 = 1.20; c_2 = \infty$.

| Case | (i) | | (ii) | | (iii) | |
|---|---|---|---|---|---|---|
| $\delta$ | $\hat{U}_2$ | $\hat{T}_2$ | $\hat{U}_2$ | $\hat{T}_2$ | $\hat{U}_2$ | $\hat{T}_2$ |
| 0.5 | 99.9 | 100.0 | 99.7 | 99.5 | 100.0 | 100.0 |
| 1 | 97.9 | 96.4 | 94.6 | 93.5 | 97.9 | 95.7 |
| 2 | 78.5 | 74.1 | 67.0 | 63.1 | 78.5 | 71.0 |
| 4 | 43.6 | 40.0 | 35.0 | 32.4 | 45.2 | 37.1 |
| 8 | 22.4 | 20.1 | 16.3 | 15.3 | 19.8 | 16.7 |
| 16 | 11.9 | 11.7 | 10.6 | 9.9 | 11.5 | 10.0 |

Table 4 Estimated powers (%) of $\hat{U}_2$ and $\hat{T}_2$ $(=\hat{T}_{1,2})$ when *p*=2, *n*=50, the underlying survival distribution is Weibull and the frailty distribution is positive stable with characteristic exponent $\nu$ under three censoring regimes: (i) $c_1 = c_2 = 2.97$ (ii) $c_1 = c_2 = 1.80$ (iii) $c_1 = 1.20; c_2 = \infty$.

| Case | (i) | | (ii) | | (iii) | |
|---|---|---|---|---|---|---|
| $\nu$ | $\hat{U}_2$ | $\hat{T}_2$ | $\hat{U}_2$ | $\hat{T}_2$ | $\hat{U}_2$ | $\hat{T}_2$ |
| 0.5 | 97.7 | 99.6 | 98.9 | 99.5 | 93.3 | 99.7 |
| 0.6 | 88.1 | 94.3 | 91.1 | 94.7 | 80.5 | 95.3 |
| 0.7 | 63.9 | 72.6 | 72.4 | 78.3 | 53.9 | 74.5 |
| 0.8 | 36.3 | 42.5 | 41.4 | 46.2 | 31.1 | 44.8 |
| 0.9 | 13.8 | 17.0 | 16.7 | 17.8 | 14.2 | 18.6 |
| 0.95 | 8.6 | 9.8 | 9.0 | 10.0 | 8.1 | 10.0 |

Table 5 Estimated powers (%) of $\hat{U}_2$ and $\hat{T}_2$ $(=\hat{T}_{1,2})$ when $p$=2, $n$=50, the underlying survival distribution is Weibull and there is marginal gamma frailty with mean 1 and variance $1/\delta$ under three censoring regimes: (a) no censoring (b) $c_1 = c_2 = 1.80$ (c) $c_1 = 1.20; c_2 = \infty$ .

| Case | (a) | | (b) | | (c) | |
|---|---|---|---|---|---|---|
| δ | $\hat{U}_2$ | $\hat{T}_2$ | $\hat{U}_2$ | $\hat{T}_2$ | $\hat{U}_2$ | $\hat{T}_2$ |
| 0.5 | 96.3 | 13.5 | 7.6 | 4.7 | 70.6 | 14.5 |
| 1 | 80.0 | 12.0 | 9.0 | 5.4 | 56.0 | 11.0 |
| 2 | 47.8 | 9.5 | 7.7 | 5.7 | 31.9 | 8.6 |
| 4 | 22.5 | 7.6 | 5.6 | 4.1 | 16.0 | 6.3 |
| 8 | 11.3 | 6.0 | 5.0 | 5.0 | 9.7 | 6.4 |
| 16 | 7.6 | 5.4 | 5.2 | 5.7 | 6.8 | 4.6 |

Table 6 Estimated powers (%) of $\hat{U}_3$ and $\hat{T}_3$ when $p$=3, $n$=50, the underlying survival distribution is Weibull and the frailty distribution is gamma with mean 1 and variance $1/\delta$.

| $\delta$ | $\hat{U}_3$ | $\hat{T}_3$ |
|---|---|---|
| 0.5 | 100.0 | 100.0 |
| 1 | 100.0 | 100.0 |
| 2 | 98.8 | 97.5 |
| 4 | 80.9 | 75.1 |
| 8 | 43.5 | 38.5 |
| 16 | 22.0 | 18.7 |

Table 7 Estimated powers (%) of $\hat{U}_3$ and $\hat{T}_3$ when $p$=3, $n$=50, the underlying survival distribution is Weibull and the frailty distribution is positive stable with characteristic exponent $\nu$.

| $\nu$ | $\hat{U}_3$ | $\hat{T}_3$ |
|---|---|---|
| 0.5 | 99.9 | 100.0 |
| 0.6 | 99.2 | 100.0 |
| 0.7 | 92.6 | 97.1 |
| 0.8 | 66.3 | 75.9 |
| 0.9 | 25.8 | 31.6 |
| 0.95 | 12.4 | 15.0 |

Table 8 Exercise time to angina in seconds of 21 patients 1 hour and 3 hours after each had received a dose of oral isosorbide dinitrite (* corresponds to a censored observation).

| 1 hour | 3 hours | Dose (mm/kg) |
|---|---|---|
| 445* | 393* | 0.58 |
| 232 | 258 | 0.24 |
| 121 | 110 | 0.38 |
| 504* | 519* | 0.41 |
| 110 | 123 | 0.37 |
| 230 | 264 | 0.24 |
| 540* | 370 | 0.49 |
| 733* | 492 | 0.2 |
| 250 | 150 | 0.38 |
| 651 | 624 | 0.51 |
| 565* | 504* | 0.51 |
| 306 | 206 | 0.34 |
| 248 | 298 | 0.37 |
| 580 | 613 | 0.32 |
| 264 | 210 | 0.37 |
| 145 | 172 | 0.53 |
| 403 | 290 | 0.44 |
| 432 | 291 | 0.31 |
| 743* | 566 | 0.24 |
| 559* | 557* | 0.27 |
| 327 | 280 | 0.24 |

Table 9 Null Weibull maximum likelihood estimates for model (4.2) for the data given in Table 8.

| Parameter | $\beta_{01}$ | $\beta_{11}$ | $\phi_1$ | $\beta_{02}$ | $\beta_{12}$ | $\phi_2$ |
|---|---|---|---|---|---|---|
| Estimate | -10.14 | -0.12 | 1.61 | -11.52 | -1.22 | 1.98 |